On the Acoustical and Perceptual Features of Vowel Nasality

by

Will Styler

B.A., University of Colorado, 2008

M.A., University of Colorado, 2008

A thesis submitted to the

Faculty of the Graduate School of the

University of Colorado in partial fulfillment

of the requirement for the degree of

Doctor of Philosophy

Department of Linguistics

2015

This thesis, entitled:

On the Acoustical and Perceptual Features of Vowel Nasality

written by Will Styler

has been approved for the Department of Linguistics

_____

Dr. Rebecca Scarborough, Committee Chair

_____

Dr. Kathy Arehart

_____

Dr. Mans Hulden

_____

Dr. Martha Palmer

_____

Dr. David Rood

_____

Dr. Wayne Ward

Date: _____

The final copy of this thesis has been examined by the signatories, and we find that both the content and the form meet acceptable presentation standards of scholarly work in the above mentioned discipline.

IRB Protocol #: 13-0668

Abstract

Styler, Will (Ph.D., Linguistics, Department of Linguistics)

On the Acoustical and Perceptual Features of Vowel Nasality

Thesis directed by Assistant Professor Rebecca A. Scarborough

Although much is known about the linguistic function of vowel nasality, either contrastive (as in French) or coarticulatory (as in English), less is known about its perception. This study uses careful examination of production patterns, along with data from both machine learning and human listeners to establish which acoustical features are useful (and used) for identifying vowel nasality.

A corpus of 4,778 oral and nasal or nasalized vowels in English and French was collected, and feature data for 29 potential perceptual features was extracted. A series of Linear Mixed-Effects Regressions showed 7 promising features with large oral-to-nasal feature differences, and highlighted some cross-linguistic differences in the relative importance of these features.

Two machine learning algorithms, Support Vector Machines and RandomForests, were trained on this data to identify features or feature groupings that were most effective at predicting nasality token-by-token in each language. The list of promising features was thus narrowed to four: A1-P0, Vowel Duration, Spectral Tilt, and Formant Frequency/Bandwidth.

These four features were manipulated in vowels in oral and nasal contexts in English, adding nasal features to oral vowels and reducing nasal features in nasalized vowels, in an attempt to influence oral/nasal classification. These stimuli were presented to native English listeners in a lexical choice task with phoneme masking, measuring oral/nasal classification accuracy and reaction time. Only modifications to vowel formant structure caused any perceptual change for listeners, resulting in increased reaction times, as well as increased oral/nasal confusion in the oral-to-nasal (feature addition) stimuli. Classification of already-nasal vowels was not affected by any modifications, suggesting a perceptual role for other acoustical characteristics alongside nasality-specific cues. A Support Vector Machine trained on the same stimuli showed a similar pattern of sensitivity to the experimental modifications.

Thus, based on both the machine learning and human perception results, formant structure, particularly F1 bandwidth, appears to be the primary cue to the perception of nasality in English. This close relationship of nasal- and oral-cavity derived acoustical cues leads to a strong perceptual role for both the oral and nasal aspects of nasal vowels.

Acknowledgements

First and foremost, thanks to my primary Advisor, Rebecca, for her constant willingness to help, hone my work, and eliminate entire unnecessary paragraphs from my writing, as well as for her vigilant efforts to keep me from getting too ambitious, too mired in code, or too engineer-ish. In addition, I am and will forever be grateful for her unflagging personal support and mentorship, through good and bad times, and for the many meetings where I left her office feeling less stressed and overwhelmed than when I walked in. I realize now that this was an incredible gift, and hope to some day pass it along to students of my own.

Thanks to my *other* advisor, Martha, who has helped Rebecca in supporting me, personally and financially, throughout graduate school. Martha has offered me incredible opportunities for training, employment, publication and experience, all seemingly without ever questioning what a phonetician was doing hanging around natural language processing projects. She has given me access to wonderful data, projects, and people, and has been a wonderful mentor, whose approach to complex problems and projects will likely always be reflected in my own.

A warm thank you to Lise Menn, my other, *other* advisor, who, despite a sincere attempt at retirement, has nevertheless continued to talk with me, mentor me, write kind things about me, and offer me tea. She also deserves credit as the teacher whose innocuous, mid-lecture comment about widespread across-speaker vowel formant variability, many years ago, set my obsession with complexity and variability in speech perception in motion.

Many students go through school never having a great advisor and mentor. I had three of them. I am a very lucky man.

Of course, I would like to thank the rest of my committee as well, for their willingness to meet with me, connect me with audiometry gear, read the occasional chapter, and to sit through 170 pages of writing on nasality.

Special thanks to Georgia Zellou for recording both Philadelphia-based speakers of French, to Story Kiser for assisting with data collection for Experiment 3, and to Luciana Marques for her help in recording and segmenting French files (as well as for her cheerful willingness to nerd out about nasality and share her insights).

Thanks to Pam Beddor, Andries Coetzee, Rob Hagiwara, Jesse Stewart, Paul Boersma, Susan Lin, Kate Phelps, Daniel Peterson, Shumin Wu, and all those other people who have provided me with useful tips, ideas, references, advice, motivation, or generalized phonetic moral support during this process.

I must also recognize the wonderful community in the CU Linguistics department: faculty, staff, students, and softball team. I will always fondly remember odd conversations in hallways, knowing looks mid-lecture, and watching the sunset from the outfield of the Mapleton softball fields, wearing the beschwaed jersey of the Illocutionary Force.

Of course, much love and gratitude to my family. To my parents, who have always supported my education and my passion for language, and who took my choice of fields rather well, all things considered. To my Godparents, who have always been there for me, and supported my education

and sanity with a steady stream of unintentionally extravagant dinners. To Cindy and Donna, who kept me laughing, and whose fur-children, past, present, and future, graciously lent their names to my English speakers.

Perhaps most of all, a loving thank you to my partner and fiancée, Jessica, for her constant support, her understanding when I arrived at the dinner table glazed over with a head full of nasality problems, and her willingness to both tear me away from writing my dissertation, and push me towards it, as she deemed healthy at the time. Nobody should ever be asked to share a partner with vowel perception, but somehow, she doesn't seem to mind, and for that, I am deeply grateful.

Finally, I'd like to thank vowels, for preventing the catastrophic nuclear implosion of syllables around the world, and for being both magical and fractally weird, thus ensuring that no matter what I accomplish in my career, I will *never* run out of questions to ask.

Contents

List of Tables

List of Figures

Note on Transcription Conventions and Terminology

All phonetic transcriptions, unless otherwise stated, will be made using the International Phonetic Alphabet (as described in International Phonetic Association (1999)). Characters will be used as canonically defined on the IPA chart (e.g. /y/ is a high front vowel, /j/ is a palatal glide).

In addition, this paper will discuss extensively the difference in coarticulatory nasality between CVC (consonant-vowel-consonant) words where there are no nasals (as in "bat"), and words which are also Consonant-Vowel-Consonant, but either or both of consonants are nasal (as in "ban", or "man"). As such, a slightly more specific set of symbols and meanings will be used in the description of phonological structures:

- C = any *non-nasal* consonant (e.g. /d/, /z/, /ʃ/)
- N = any nasal consonant (e.g. /n/, /m/, /ŋ/)
- V = any vowel *which is not phonemically nasal* (all English vowels, or French oral vowels like 'fait' (/fɛ/))
- Ṽ = any vowel *which is phonemically nasal* (as in the French 'fin' (/fɛ̃/))

Using this system, we can more easily distinguish syllables with environments for nasal coarticulation (CVN, NVC, NVN) from oral environments (CV, VC, or CVC), and we can discuss syllables containing contrastively nasal and oral vowels as classes (CVC vs. CṼC), rather than vowel-by-vowel.

Unless otherwise specified, in Tables and Figures, "en" refers to English and "fr" refers to French.

Finally, as this work focuses on phonetics and the production and perception of speech, "nasality", unless otherwise specified, refers to the phonetic, articulatory phenomenon in speech where sound is produced with the velopharyngeal port open, allowing nasal airflow, rather than phonologically contrastive nasality or [+NASAL] specification.

1 Introduction

Language is fundamentally about meaningful, functional, and perceptible contrasts. In order for one word to be different from the next, we must be able to reliably produce distinctions between sounds and patterns of sounds. More importantly, though, these contrasts can only work if listeners can, somehow, extract and notice them amidst the wildly variable acoustical signal that is speech.

Many of these contrasts change meaning in Language. Silence contrasts with speech. Voiced airflow contrasts with voiceless airflow. Vowels are meaningfully different from consonants, and two similar consonants made at different places in the mouth might carry the contrast between two words. These contrasts are phonologically relevant, consciously known to speakers, and crucial for understanding.

Some contrasts, although not salient enough to speakers to change meaning or to derail comprehension, are nevertheless useful and functional for speakers and listeners. The difference between two vowel pronunciations may not alter the word, but may indicate that the speaker "isn't from around here". A consonant moving forward in the mouth may serve to indicate a coming front vowel, and the addition of aspiration may serve to distance a voiceless stop from a voiced one. These contrasts, although not *consciously* attended to by speakers and listeners, are still of tremendous use as a secondary source of information, and the failure to adequately produce them can leave a speaker isolated and, well, marked.

Vowel nasality is a fascinating phenomenon in language. In French, as well as in Hindi, Lakhota, Navajo, Brazilian Portuguese, and many other languages (constituting nearly 30% of the WALS dataset (Haspelmath (2005))), vowel nasality is phonemic. This means that a word's meaning can change entirely depending whether the velum is raised or lowered during a given vowel's production. Thus, in French, *beau* [bo] can mean 'beautiful', and *bon* [bõ] can mean 'good'. In these languages, the production and perception of nasality is doing obvious linguistic work, and is of crucial importance to the communicative act.

However, in languages where nasality is not phonemically contrastive, vowel nasality is neither absent nor irrelevant, simply different. When speakers of any language produce nasal sounds in the vicinity of vowels, they *coarticulate*, that is, the nasal gesture overlaps the oral vowel gesture. Far from being simple articulatory 'slop', vowel nasality is still meaningful to listeners. Listeners use coarticulatory vowel nasality as a supplementary cue for oncoming nasals (Lahiri and Marslen-Wilson (1991), Beddor and Krakow (1999)), and there is strong evidence that nasality provides a cue for difficult words (Scarborough (2013)), for easing difficult contrasts in non-words (Scarborough (2012)), for other, completely unrelated consonant contrasts (Zellou (2012)). Even where nasality is not meaningful in the conventional sense, then, it is still attended to, and still useful.

Clearly then, the linguistic phenomenon of vowel nasality is doing meaningful work in Language and communication, no matter how a given language uses and classifies it. And of course, a phenomenon as useful and used as this one must be easily produced and perceived in order to be as common as it is. But we as linguists have a fundamental problem: although we know that

nasality is used by listeners in a variety of ways, *we do not know what listeners are actually listening for when they hear nasality*.

This is not to say we have no knowledge at all about the acoustics of nasality. There are a variety of measures and analyses which provide some information about the nasality of vowels, and over large datasets, these measures are accurate often enough to allow study of nasality based on acoustical data alone.

However, unlike human perception, at a token-by-token, functional level, none of these measures can reliably predict nasality based on acoustics alone. For example, if we take the most commonly used of these measures, Marilyn Chen's A1-P0 (discussed extensively in Section 2.3.2), and look for a phonetically-expected increase in coarticulatory nasality over a large dataset, we're likely to find a significant correlation between A1-P0 and nasality in the aggregate. But token-by-token, the measure does not provide sufficient precision nor recall to say meaningful things about the nasality of any particular word. Put differently, even when measuring vowels known to increase in nasality, current measures will indicate a drop in nasality almost as often as they indicate the expected rise. This is why, in the literature, all work using acoustical nasality measurement does so over large datasets, and without reference to particular tokens.

Further displaying the complexity of identifying nasality, in Patrice Beddor's 2009 *Language* paper, she explicitly states that when labeling the onset of nasality in words, "No single spectral criterion for identifying vowel nasalization onset could be applied to the tokens of all speakers." (Beddor (2009), Footnote 3). As such, she was forced to have a pair of human annotators mark boundaries considering several known features, and then adjudicate their work. A similar multi-feature human-centric approach was used in Delvaux et al. (2012). Any dataset will have atypical tokens, and any measure will have errors, but if this is the method preferred in modern linguistic research, we see that there is a massive gap between our ability to externally measure and classify nasal vowels and our internal ability to perceive them.

Vowel nasality, then, is a phenomenon which is complex beyond our present understanding. It quite clearly exists in human language as a relevant abstraction, and is explicitly identifiable by listeners, and any contrast which can serve a function for listeners must be present in the signal and able to be quantified by researchers. But at the present, we as linguists don't know which characteristics of the acoustical signal listeners are attending to, and we cannot simulate that percept without the benefit of an experienced human mind.

This raises a fundamental question which this dissertation will aim to address: How do we as humans distinguish oral vowels from nasal vowels?

To test this question, we will examine the acoustics of nasality in its two phonological forms: coarticulatory nasality (studied here in English), and contrastive nasality (here represented by French), and test the human perception of nasality in English. Through this process, we can hope to gain a clearer understanding of those acoustical factors which are useful, necessary, or sufficient for listeners to perceive a vowel as nasalized, as well as to understand the differences, if any, in the perception of nasality across these two languages.

Before going further into detail about the methods we will use to address this question, a word on nasality itself is required.

1.1 On Vowel Nasality

Vowel nasality, no matter its function or use, stems from a lowering of the velum during the vowel, which opens the velopharyngeal ("VP") port and allows air to flow out through the nose and nostrils (rather than just through the mouth). However, unlike with nasal stops, oral airflow is maintained as well throughout the nasal vowel. By doing this, additional nasal resonances and zeroes are added to the acoustical realization of the vowel, and listeners are, somehow, able to perceive a difference in sound (in phonemic cases, at least), and understand (consciously or unconsciously) that the VP port of the speaker is open, perhaps estimating the degree of the aperture, and adjusting their perceptions accordingly.

Nasal vowels provide a variety of information, both to listeners and to linguists. We might wonder how nasal a given vowel is relative to others from the same speaker (as a cue to contrastive nasality). We might want to know when the nasality starts (or ends) during the course of the vowel, and what that contour looks like (perhaps shedding light on the nasality of surrounding sounds). And, of course, linguists may want to compare the degree of nasality or the shape of these contours across several speakers, conditions, or even across different languages, as will be discussed in detail in this dissertation.

In order to address any of these questions, though, we must know what constitutes nasality *as it is used in language*. In the same way that speakers of a language "know" its grammar well enough for use, even without conscious awareness of the rules or intricacies, listeners are clearly quite capable of recognizing the acoustical features which indicate the presence, degree, and contour of nasality from which functional information can be gleaned. However, as with the use of our internal grammar, this internal process of nasality perception is opaque to listeners and observers, and can only be observed based on external behavior.

Because of this opacity, to understand the process of human nasality perception, we must analyze and measure the signal externally, using computational and signal analysis techniques which are transparent to researchers. Only once we can externally analyze and quantify the information present in the signal can we investigate its functions in language. Thus, in order to understand how nasality works for listeners, we must first examine how it can be analyzed by researchers and by machines.

1.2 Analyzing Nasality in the Laboratory

There are a variety of means to analyze vowel nasality which are available to researchers. These range from extraordinarily technical solutions, visualizing the velopharyngeal port itself using magnetic resonance imaging (MRI), to airflow measurement using transducers or flow-volume sensors, to simple impressionistic description based on listening. Each of these methods has its own distinct set of strengths and weaknesses, and many require specialized (and often expensive) equipment.

In linguistic research, the majority of studies are conducted either acoustically or using pneumotachography (nasal airflow measurement). Although nasal airflow measurement is excellent for gaining information about VP port activity (albeit still indirect information), it is also intimidating

for subjects, requiring a mask, nostril tubes, and other specialized equipment, and the researcher and equipment must be present in the room with the speaker in order for the data to be collected. Most importantly from our perspective, while airflow data is certainly clean and robust, it is not necessarily information that's available to the listener in a linguistic context, and does not necessarily provide direct information about the state of the velopharyngeal port (Shosted et al. (2012)).

Pneumotachography is excellent for production studies, where one is simply interested in the articulatory actions of speakers while producing that which is meant to be "nasal". It provides excellent time-resolution, and interpreting the data is easy, as it directly gives you the proportions of nasal vs. oral airflow. However, for any perceptual or listener-centric research, focused on what is available to and used by listeners, studying nasality using airflow data is far from ideal.

An alternative is the acoustical analysis of nasality. Simply put, one takes audio data, determines a property of that signal which will be taken to represent nasality, and quantifies that property to quantify nasality. This sort of analysis can be easily run on previously recorded data, no specialized equipment is required (beyond a computer and microphone), and most importantly, the analysis is based entirely on the speech signal which is made available to the listener during speech. For experiments dealing with perception, intelligibility, naturalness, or interactions within the speech signal, acoustical nasality measurement is the best mirror of real linguistic practice, and thus, is the best method available.

Its biggest downfall, however, is the simple fact that identifying and quantifying nasality as a component part of the greater speech signal is as complicated and fragile for the linguist as it is for the listener, but without the benefit of years of practice and millennia of evolution.

1.3 Dealing with noise in nasality measurement

Although it is easily categorically perceived and clearly available to listeners in coarticulatory contexts, finding acoustical correlate(s) of nasality is has proved to be far from a trivial task.

As previously discussed, a variety of acoustical correlates and measurements exist for analyzing nasality, and these suffice for studies using nasality as a measure or feature of interest over a large corpus of measurements or tokens. However, these methods do not provide sufficient by-token precision to say anything useful about the nasality of an individual vowel token, nor of small sub-classes. This limits the questions one can ask, and forces studies of nasality to perform an exceptionally large number of measurements in order to have meaningful data. However, this problem does not affect just the linguistic world.

In speech recognition, the lack of a clear method of classifying nasal vowels is a missed opportunity in languages which have contrastive nasality, in some cases forcing engineering workarounds (based on word frequency or context) rather than directly addressing the problem. Even for languages where nasality is not contrastive, a better understanding of vowel nasality would provide considerable additional information about surrounding low-amplitude nasal consonants.

In a different world altogether, speech pathology clinicians are often tasked with finding and treating hypernasality (variably caused by poor articulatory form or by anatomical malformations

such as Cleft Palate). Currently, the field of speech pathology uses nasal airflow measurement almost exclusively, with research in the field having found acoustical measures of nasality lacking sufficient "accessibility and sensitivity" for clinical use (Vogel et al. (2009), more detail in Buder (2005)). This forces clinics and clinicians to purchase and use expensive airflow measurement equipment, even for small practices, and as such, a more sensitive and accurate acoustical measure of nasality would be of considerable use to the speech pathology community as well.

So, we see that there is a large and problematic gap between our ability to externally measure and classify nasal vowels and our internal ability to perceive them, and that this is affecting more than just linguistic research. This leads us to our current state of affairs, where we have many ideas of what happens in nasal vowels, but little understanding of how, exactly, nasality is perceived.

1.4 Building a better understanding

This dissertation aims to narrow this gap, to gain a better understanding of how nasality is perceived in language, and to identify those features *actually used* in the perception of nasality, both to improve our understanding of human perception, and perhaps to improve our ability to measure nasality in laboratory settings.

To date, research on the acoustics of nasality has been quite rare, and, when conducted, usually focuses on individual measures of posited spectral properties for nasality (described in Section 2). Effectively, they have aimed to find a single point of data in production data which can be pointed as *the* indicator of Nasality in speech, across words, speakers, and contexts. Although this research has been useful and has established a number of useful measures, successful single-data-point measures are actually relatively unusual in the body of speech research.

For vowel articulation, tongue movement simultaneously affects F1, F2, and, to a lesser extent, F3. Lip rounding affects F3 primarily, but has a lowering effect on the other formants as well (due to the increased length of the cavity). Phonation type (breathy vs. creaky vs. modal) is most often measured using spectral tilt, but detailed analysis also reveals variation in periodicity, amplitude, formant structure, and more (Gordon and Ladefoged (2001)), all of which must be considered during analysis. In fact, most complex speech phenomena have multiple acoustical effects, and only as we've broadened our understanding of the underlying acoustical effects to include these secondary features have we been able to reliably detect these phenomena, and have we gained understanding as to how they may be perceived.

Coupling the oral and nasal cavity has a variety of acoustical effects, and thus, many potentially redundant cues to the perception and classification of nasalized vowels. As such, this dissertation will not focus on a search for "the one true Cue" which will alone model nasality in perception. Instead, we will test many cues *at every stage*, with the hope that, if humans are using multiple features, we will understand and capture the features and their weights.

In addition, we must consider the possibility that this perceptual process is different from language to language. Although it is intuitive that the extraction of similar information from the acoustical signal would be consistent across all humans, different goals may produce different approaches to nasality. For instance, it is certainly possible that speakers of contrastive nasality languages (like

French) are able to enhance certain acoustical properties of nasal vowels through practiced articulation, and that listeners will attend preferentially to those optimized cues. Similarly, speakers of languages where nasality is of less importance may aim to reduce the acoustical effects of this coarticulation, and listeners will thus attend to this particular realization instead.

Regardless of the nature or reasoning of any such cross-linguistic differences in perceptual cue utilization, the idea of such differences must be considered in order for any perceptual model to be complete. Evaluating the differences, if any, in nasal perception across these two languages is a primary area of interest for this study, and although French and English are compared in the present work, this dissertation will present a working paradigm for the evaluation of differences across other languages.

In order to address all of the issues raised so far, a more detailed understanding of the nature of nasality is required. To gain this understanding, we seek to examine a variety of acoustical effects posited for nasality, both in the literature and in the author's current work. Then, based on correlations, usefulness in machine classification, linguistic analysis, and eventually, human perception experiments, we will find those acoustical features which, alone or combined, provide the best information for the perception of nasality.

1.5 Proposed Course of Action

Identifying probable cues based not on *a priori* assumptions but on analysis of data is a complex task, and will require many steps, as well as four different experiments to do properly.

To begin, we will review what literature exists on nasal acoustics and create a list of acoustical features previously suggested in the literature, and combine them with a few additional measures and characteristics of nasality, derived from the author's own work over the last seven years.

Then, we will record a large corpus of words containing oral and nasalized vowels in English, and phonemically oral and nasal vowels in French. This corpus of sounds will then be annotated and automatically mined, collecting measurements for all of the features collected above. These measurements will then be used in a statistical study, examining the relationship of each feature to the phonologically (or phonetically) expected nasality in the vowels.

At this point, we must determine whether the features that we have isolated are simply correlates, or whether they are truly useful (and used) for the perception of nasality, whether by humans, or by machines.

To this end, the measurements generated for the above correlation study will fed into a machine learning based classification algorithm. This classifier, a complex statistical modeling program, will attempt to classify speech tokens by nasality in a series of experiments, examining features on their own, examining the importance of each feature in large models, and eventually, testing the features in sets, to isolate the most useful sub-groups for human testing. The output of each model will then be evaluated in terms of accuracy, that is, how often the computer labeled a nasal vowel as "nasal". Features (and feature sets) will be ranked by their utility for classification. These results will inform the human perception experiments in three key ways.

First, they allow the usefulness of the features to be tested in isolation, without the use of context or top-down processing available to humans. If these features do not, even combined, contain enough information for the computer to classify sounds accurately as oral or nasal at a better-than-chance rate, it will suggest that human nasality perception is based more on context and top-down processing than on the acoustics of individual tokens.

Second, we can assume that listeners would be less likely to rely on a potential feature which is poorly predictive of nasality, whether due to poor detectability, interference from other phenomena, or due to variation in production. In the same sense that we do not find the proper bus for our route by attempting to recognize the driver (as driver identity changes often and is difficult to see through a window), we would not want to identify nasal sounds using a feature which is infrequently present, variable, or troublesome to detect. These machine learning studies will allow direct measurement of the predictive power of each feature *in a token-by-token task*, rather than across the entire dataset.

Finally, we cannot test all 29 of our features with actual human listeners, due both to the experimental complexity and time required. Computer perceptual experiments are far less expensive and time-consuming to run. As such, we can use machine learning studies to narrow a large list of features to a testably small list of promising features. This is a far more rigorous approach to constraining the task than simply picking out those features that seem, *a priori,* like they might work best.

Through this machine learning study, we will narrow the list of features for evaluation down from 29 to 4 individual features. By considering the machine learning data, alongside the statistical data from the first experiment, we should be able to make hypotheses about the remaining features, which will inform the final step, a human perception experiment.

In this experiment, described more fully in Section 6, a series of stimuli will be tested in which each potential feature is added to an oral vowel (to perhaps produce a nasal percept) or removed from a nasal vowel (to perhaps produce an oral percept). These stimuli will be presented, alongside control stimuli and stimuli where *all* features have been modified, to English speakers in a lexical choice task with phoneme masking, and the resulting reaction time and responses will be evaluated.

With these data, we should be able to determine which of the features are necessary for a nasal vowel to be perceived as nasal, which are sufficient for an oral vowel to be perceived as nasal, and which appear to have no role at all in the perception of nasality.

At this point, we will run a parallel experiment, presenting the same stimuli tested in humans to our machine learning algorithms. This will allow us yet another perspective on cue utilization, and will allow us to make stronger claims about the correspondence, if any, between humans and machine perception.

At this point, we can discuss the feature or features used in the perception of nasality in English, and make hypotheses about the perception of nasality in French.

Finally, with all of this information in mind, we will discuss the acoustics, perception, measurement, and classification of nasality, and the consequences of these findings for the study of nasality in a linguistic context.

One note: although we will discuss the time course of nasality (the temporal pattern of degree of nasality throughout the vowel), and will use measurements at different time points as a factor in some analyses, the temporality of nasality will not be discussed as a perceptual cue in-and-of itself. It is known that nasality's time course varies both across languages and across coarticulatory contexts (discussed in detail in Section 2), and it is likely that this is a reinforcing source of information. However, in order to perceive (and gain information from) the temporal course of nasality, listeners must first be able to extract nasality from the signal, and this lower level task is what we are focused on here.

So, in summary, we hope to gain a better understanding of the acoustics of nasality, of the utility of different cues for machine classification, and of the perception of vowel nasality in English. This information will ideally prove useful for anybody interested in the measurement, classification, perception, or generalized study of vowel nasality in language.

2 Background: On Vowel Nasality in English and French

The goal of the present work is to analyze the acoustical and perceptual nature of nasality in French and English. But before such analysis can take place, some background on nasality must first be established, such that we can understand the phenomenon itself.

This chapter will first examine what we know about the nature of the phenomenon of nasality in English and in French, based on sources both in phonology and in speech production. We will then briefly discuss the possible sources of difference in the nature of nasality between English and French. Then, we will look more closely at our current understanding of the acoustics of nasality, with the goal of extracting useful and usable acoustical features from prior work for testing and evaluation in later experiments. Finally, some of our existing knowledge about the perception and automatic detection of nasality will be discussed and analyzed, both in terms of the questions already answered and in terms of the methodologies used in prior experiments.

2.1 Vowel Nasality in English

English vowel nasality is coarticulatory in nature. In the past, this has been understood to mean that nasality is not an inherent, phonological property of the vowel (+Nasal and -Nasal are not relevant), but instead, nasality arises during phonetic implementation, when a vowel is preceded or followed by a nasal consonant.

This is not to say that nasality is purely accidental or in free variation, nor that nasality is purely a matter of articulatory planning and ease. In fact, there are compelling arguments to be made that although nasality is not phonologically contrastive (as in French, or Hindi), coarticulatory nasality has become part of the phonology (see Lahiri and Marslen-Wilson (1991), Cohn (1990), Bybee (2003), among others). However, regardless of the level of its representation, it is clear that although vowel nasality is not phonemically meaningful in English, it is most certainly perceived and used by speakers.

In this section, we will discuss the articulatory nature of nasality, some causes of variation, and, perhaps most interestingly, the ways which nasality has been shown to be useful to native English speakers, even despite its coarticulatory nature.

2.1.1 The nature of vowel nasality in English

Vowel nasality in English is triggered by coarticulation, that is, by adjacency to one of English's three nasal consonants, /m/, /n/, or /ŋ/. This coarticulation occurs because producing a nasal consonant requires the velum to be lowered, opening the VP port and allowing airflow through the nose. Given that the velum and tongue can move independently, it is more anatomically efficient to decouple the two gestures, beginning (or ending) the velar gesture outside of the nasal consonant's boundaries.

This coarticulation is such that in a CVN word ('anticipatory coarticulation'), the velum begins lowering before the vowel has concluded, resulting in increasing nasality throughout the vowel.

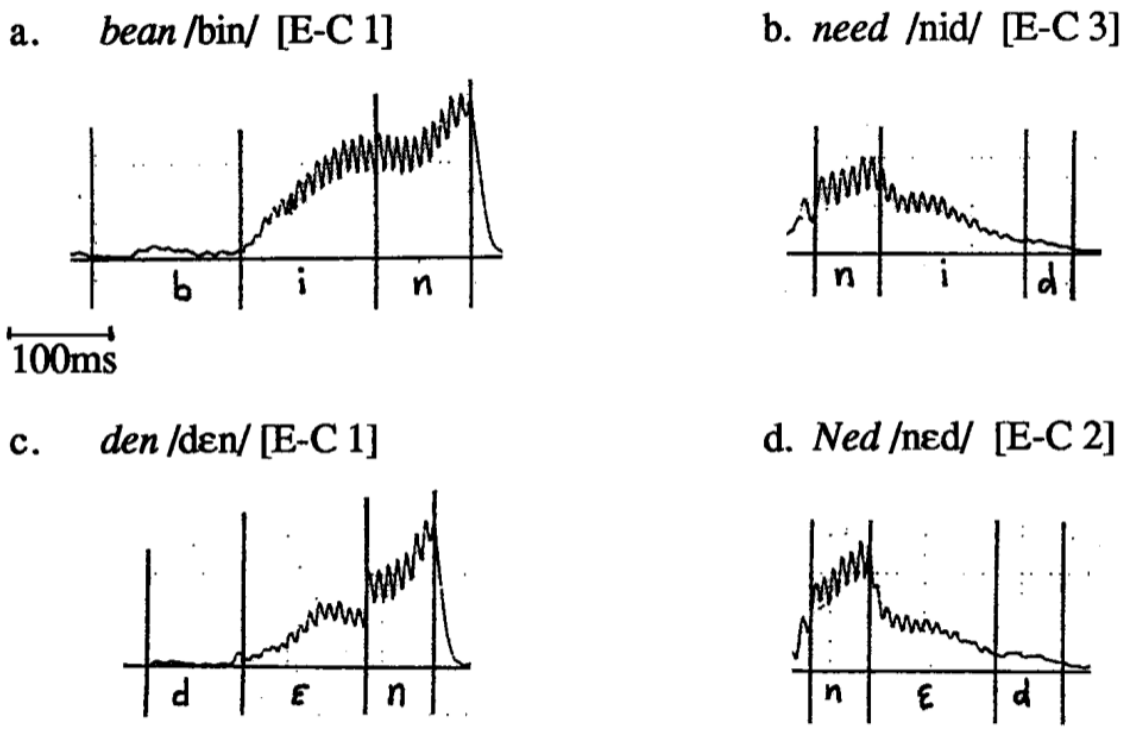


Figure 1: Airflow traces of CVN and NVC words (from Cohn 1990, Fig. 6, pp. 145)

In an NVC word, 'carryover' coarticulation occurs, with the velum raising during the course of the vowel, and in NVN words, the velum remains open throughout the vowel to conserve articulatory effort. This coarticulation, in any context involving a nasal consonant, allows nasal airflow during a vowel, creating a phonetically nasal vowel.

The exact nature of the velar gesture (and thus, the vowel's nasality) varies depending on the coarticulatory context. The nature of coarticulatory nasality in English was described extensively in Cohn (1990), where it was examined in detail using airflow measurement.

Cohn found that in anticipatory coarticulation cases (CVN), as one might expect, nasal airflow increases during the vowel (in the run-up to the nasal consonant). Similarly, in carryover (NVC) contexts, nasality decreases during the vowel (as distance from the nasal consonant increases). This overall pattern is shown in Figure 1.

It is worth noting, though, that the flow does not always increase or decrease linearly. Cohn observed some plateauing, especially in CVNC and in anticipatory cases where the vowel follows a voiceless stop or /h/ (pp. 145), and the airflow traces showed bumps and non-linearities even in other cases[1].

This variation is not only in time course, but in degree (or amplitude) of nasality as well. Cohn shows cases where the same word, in different repetitions, shows differing degrees of nasality. As

[1]Although airflow is not a direct proxy for the degree of VP port aperture (due to turbulence in the airflow, degree of oral closure, etc), it is quite robust for indicating the presence of aperture, as well as for examining the time course of nasality. Thus, these non-linearities should be assumed to actually exist in the speech signal rather than treated as measurement artifacts.

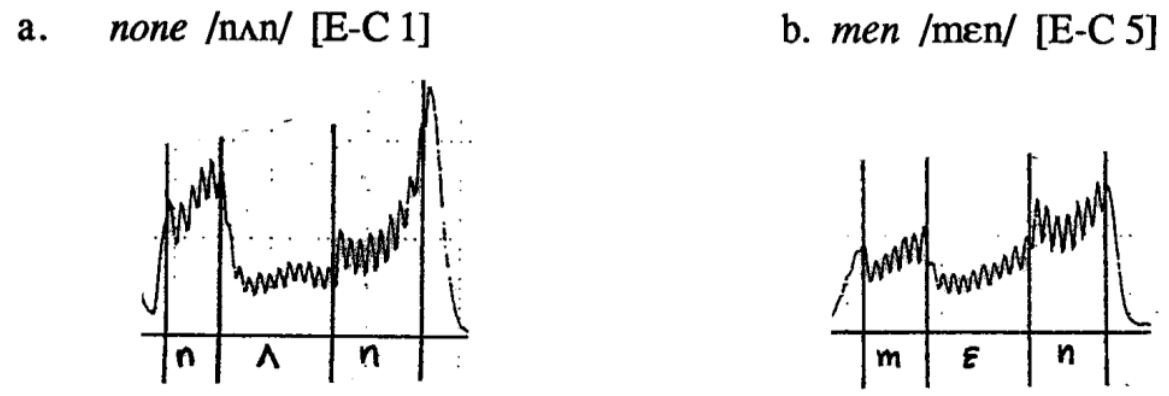


Figure 2: Airflow traces of NVN words (from Cohn 1990, Fig. 8, pp. 147)

Cohn puts it:

> "Often it appears to be the case that a fairly wide range of shapes, or interpolation functions are allowable. Thus, for any particular window, a range of acceptable paths are possible […]"

To generalize, in English CVN and NVC words, although there is a great deal of variation, we should assume that nasal airflow will generally be greatest adjacent to a nasal consonant, and least opposite the nasal, and that there is, as Cohn puts it, a tendency to straight interpolation along the most direct path. This finding is supported extensively in other acoustical studies (Scarborough (2013), Scarborough (2004), Beddor et al. (2013), Beddor (2009)), as well as in the acoustical measures of the present work.

NVN words show a different pattern. According to airflow measurement (again in Cohn (1990)), the nasality in NVN vowels is generally flat (although again, not strictly linear, and often demonstrating a sort of scooped pattern, lowest in the center of the vowel, see Figure 2. Cohn observes that the amount of nasal airflow is lower during the vowel than during the nasal consonants (which is to be expected, given that airflow is split among the oral and nasal cavities). But again, we can make the general claim that in NVN words, nasality will be strongly present and relatively steady throughout the vowel, and again, this is shown by acoustical studies elsewhere in the literature.

So, in terms of the velar gesture, both airflow and acoustical studies show that a nearby nasal consonant causes nasality in the part of the vowel closest to that nasal consonant, or, in NVN words, a plateau of nasality throughout the word. We also see significant variability in the exact time course (and degree) of the vowel's nasality from token to token, even of the same word, indicating that although the general trends hold, even in airflow data, the velar interpolation between oral and nasal is not necessarily linear.

Although we have so far focused on the lowering of the velum, the oral articulation (tongue positioning) of nasalized vowels can also differ from fully oral vowels. Carignan et al. (2011) find that nasalized /i/ vowels show increased height relative to oral /i/ vowels, although no such height difference appears for nasal vs. oral /a/ vowels. The authors of the study attribute these changes, which effectively negate some posited acoustical features of nasality, to a subtle

act of resistance among speakers, a small compensation to prevent nasality from growing more prominent (and thus phonologically important) than it already is. Similar results (although their cause is left unaddressed) are reported in Pruthi and Espy-Wilson (2004) as well.

This phenomenon of changes in oral articulation for nasalized vowels is not unique to English, as differing oral configurations for nasal vowels have been shown in Hindi as well (see Shosted et al. (2012)), albeit there it occurs for the purpose of maximizing contrast. We will need to ensure that acoustical markers of vowel height and position (namely F1 and F2) are included in the analysis for both languages.

So, in English, coarticulatory nasality is expressed by velar opening during the vowel, decreasing with distance from the nasal consonant. In NVN words, this appears similar to superimposing the coarticulatory traces of an NVC and CVN, leading to a vowel with flat or slightly "scooped" nasality. We also see that although the trend is towards direct interpolation, there is variability in both the time course and degree of nasality in English vowels, and finally, that even the oral articulations associated with the vowel often change with the introduction of nasality. Now, we'll discuss some of the other (known) sources of noise in the nasality signal.

2.1.2 Variation in vowel nasality in English

Although it is true that different phonological structures tend to result in different patterns of nasality, there are other sources of variation in nasality which are not easily attributable to pure articulatory efficiency or structure. One example, likely arising from anatomy, is the oft-stated phonetic idea that all things considered, low vowels are more nasal than high vowels in equivalent phonological situations. This has led to the assertion that typologically speaking, high nasal vowels are less common than low nasal vowels[2]. This has been shown in French (Delvaux et al. (2008a)), as well as in English and Hindi (Shosted et al. (2012)), and is generally hypothesized to be a form of compensation for the increased acoustical prominence of nasal/oral coupling in high vowels (Delvaux et al. (2008a)). Although this is a more straightforwardly anatomical and acoustical variation, other variations may suggest some degree of purposeful manipulation of nasality among English speakers.

It has been shown that speakers of English decrease the degree of nasality in words when asked to speak "clearly", or to somebody who is hard of hearing , and increase the degree of nasality when speaking casually (Scarborough and Zellou (2013)). This manipulation of degree of nasality must stem from the speaker's conscious awareness of the (real or imagined) conversational situation. Although the authors attribute this distinction to potential listener-directed accommodation, regardless of the source of this variation, it is clearly well beyond the simple coarticulatory variation which a "nasality is purely phonetic in English" approach, alongside variations caused by speech rate, articulation, and so forth, can predict.

As further evidence, other studies have emerged linking degree of nasality to lexical characteristics of words. One such characteristic involves the lexical "neighborhood", the number of words which differ by only one sound (so, "pat" would have "cat", "pit", "pad", and so forth, as neighbors), as

[2]This assertion is examined more closely in Hajek and Maeda (2000), where evidence is presented both for and against, and alongside possible causes of the effect.

well as its "density", the number of neighbors a word has. It has been shown that (alongside other hyperarticulatory features), speakers produce greater degrees of nasality for "hard" words (which have many frequent neighbors) than for "easy" words (prominent words with few neighbors) (Scarborough (2004), Scarborough (2013), Scarborough (2012)). Similar neighborhood density effects on nasality have been shown for nonwords, but the exact nature of the change in nasality based on neighborhood for nonwords, interestingly, appears to vary by speaker (Scarborough (2012)). The fact that speakers appear to be adjusting the coarticulatory velar gesture in response to lexical characteristics of words again appears to suggest lexical sensitivity in the process of planning the coarticulatory gesture, and again suggests a listener-directed change.

This, as we can see, provides further evidence that nasality in English, although coarticulatory in nature, varies with far more detail and control than can be attributed to "efficient phonetic implementation" alone. More importantly, these sorts of nuanced variations show that in English, the presence, degree and timing of vowel nasality encodes more than just "there are nasal consonants afoot", and may carry information or accommodation which is not just present in speech, but potentially useful.

2.1.3 The usefulness of vowel nasality to English listeners

The mere presence of vowel nasality is no proof that listeners will attend to it, especially when vowel nasality is of no phonological significance in a language. Here, we ask simply whether vowel nasality is relevant enough to English listeners to be attended to in speech perception.

The question at hand is not whether English-speaking listeners are *able* to perceive nasality in vowels in isolation or in experimental contexts, as this has been examined in a variety of studies, many of which will be described below. Instead, we can ask a more specific question - Is vowel nasality is *of use* to English speakers in the perception of language?

Unless we take the stance that listeners are steadfastly ignoring the coarticulatory vowel nasality in English or simply do not consider it during perception, we might expect that listeners would make use of nasality, consciously or unconsciously, as a cue to the presence of surrounding nasals, and this is indeed the case.

Lahiri and Marslen-Wilson (1991), as a means to proposing a psycholinguistic model of perception, examined the perception of nasality in English and Bengali. Although the theoretical psycholinguistic and phonological arguments about underspecification in the representation of vowel nasality are irrelevant to the present discussion, their method and results are supremely relevant. The authors presented English and Bengali listeners with "gated" stimuli, otherwise unmodified vowels where the sound is cut off before the natural ending of the word, and asked them to complete the word. In this particular case, the gating occurred part way through the vowel, before the coda of the word is audible. They tested a prediction that nasality at any point in the vowel in English would be attributed to a following nasal consonant, whereas they predicted that in Bengali, because there is ambiguity between coarticulatory and contrastive nasality, listeners would give equal numbers of CVN and CṼC responses.

Ultimately, their results were largely as hypothesized. Both Bengali and English speakers were

able to detect nasality in a gated vowel, and CVC vowels were categorized as nasal far less often than CVN and CṼC vowels. Interestingly, when presented with a ambiguous nasal vowel, Bengali speakers were far more likely to classify it as a contrastively nasal vowel, rather than attributing the nasality to coarticulation. Although this study uses natural nasal vowels, and makes no claim at all about the nature of the cues which indicate nasality, it does use an experimental design very similar to the one used in the present work. Perhaps most importantly, it shows that nasality on vowels, in both contrastive and coarticulatory contexts, is actually doing work, that is, listeners are using vowel nasality to support (or sometimes motivate) their judgement of upcoming or ambiguous sounds. Similar results are reported for English and Hindi in Ohala and Ohala (1995), further confirming the utility of this phenomenon for English speech perception.

Another study examining the perception of nasality among English speakers is Beddor and Krakow (1999). In this study, oral and nasal vowels were spliced into oral and nasal contexts (CVC and NVN) and placed in isolation. English and Thai speakers were then asked to evaluate the similarity of nasality of vowels in these contexts. Interestingly, the authors found that the context of evaluation mattered, and that all listeners were better at judging nasal similarity in non-nasal contexts. The authors attribute this difference to perceptual compensation, where listeners attribute the acoustical effects of vowel nasality to the nearby nasal, and they point out that the degree of compensation is affected by the native language of the listeners. This study is particularly relevant in that it shows both that English speakers are capable of evaluating nasality, and also that the context of evaluation of stimuli plays a strong role in the evaluation itself.

Beddor et al. (2013) confirms the results of Beddor and Krakow (1999) in an eye-tracking paradigm, showing that English speaking listeners start to fixate visually on a CVNC word (rather than a CVC word) as soon as they hear nasality in the vowel, rather than waiting for a following consonant. Although the paper does not directly concern itself with what parts of the signal constitute nasality, it does again show that English speakers attend to nasality, and that it is useful in making lexical decisions.

So, clearly, English-speaking listeners are using nasality to help with lexical decisions and word disambiguation. Because our English perception experiment methodology depends on listeners' attention to nasality for word completion, this is the use of nasality most important to the present work.

It is worth highlighting that English speakers appear to be attending to nasality even outside of phoneme masking or word completion tasks. In Scarborough et al. (2011), we manipulated nasality in natural tokens in such a way as to reduce or increase their nasality, and then presented them to listeners in a reaction-timed lexical decision experiment. For high neighborhood-density ("High ND") words (which are naturally produced as more nasal according to Scarborough (2013)), listeners preferred (responded more quickly to) the increased nasality tokens, compared to those similarly modified to have decreased nasality. This would seem to indicate that listeners are not only aware of speakers' tendency to use increased nasality in High ND words, but that they both perceive and expect it. Interestingly, this effect was not significant for Low ND words, and we offer several possible explanations for this asymmetry. Regardless, this study shows that listeners both attend to such subtle differences in nasality and store them, using them as a factor in lexical decisions.

All of these studies indicate that English speakers are not only attending to coarticulatory vowel nasality, but are taking full advantage of it during speech perception, using it as a cue not only to nearby nasal consonants, but as a property of the word itself.

To sum up our discussion, vowel nasality in English is not phonemic, but instead arises due to the presence of nasal segments adjacent to the vowel. The nature of the nasality, in terms of degree and time-course, varies depending on the phonological structure of the word, as well as based on the speech context, the lexical characteristics of the word, and speaker idiosyncrasies. Most importantly, though, despite nasality being "merely coarticulatory" in English, listeners do attend to vowel nasality, and use it as a cue not only to the presence and location of nearby nasal consonants, but also as a part of the representation of the word itself.

Thus, we find that vowel nasality in English is far from unimportant or uninteresting. Even in a "coarticulatory nasality" language, we see that nasality is not just present, but useful as a part of everyday speech perception, and most important of all, though, we see that English-speaking listeners, like speakers of any other language, have excellent reasons to both develop and hone their faculties for the perception of nasality in the acoustical signal of speech.

2.2 Vowel Nasality in French

Unlike in English, French vowel nasality can arise either from phonemic nasality in vowels, where the lowering of the velum itself indicates a change in word, or due to coarticulation, stemming from nearby nasal consonants as in English.

2.2.1 French coarticulatory nasality and the present work

Because French offers both contrastive and coarticulatory nasality, we must briefly discuss the role of both phenomena in the present work. Here, we are focusing on the search for differences in the perception or realization of nasality in languages where nasality is phonemic (as in French) versus languages in which vowel nasality is primarily coarticulatory (as in English). This cross-linguistic contrast allows us to study speakers with two different sets of goals for vowel nasality perception.

French listeners presumably *need* to be able to recover vowel nasality accurately and quickly as a part of the communicative act. If a French listener hears the sentence *il est bon* ('he is good'), in order to unambiguously interpret the sentence, he or she must presumably interpret the nasality of the final vowel to distinguish *bon* from the oral *beau* ('beautiful'). Thus, in French, nasality perception has far higher stakes than in English, where although nasality is useful (see Section 2.1.3), the interpretation of a word or sentence will rarely hinge on the nasality of a given vowel. This change in the consequences of proper perception is one likely source of any difference in perception among the two languages, and motivates our choice of using two languages, rather than one.

This focus on the contrast between coarticulatory and contrastive nasality in English and French leaves French coarticulatory (non-phonemic) nasality somewhat in the middle. In interpreting

a CVN word's vowels, French listeners are potentially contending with perceptual compensation and other coarticulation-related issues, while at the same time using their knowledge and understanding which arises from regularly interpreting phonemic nasality.

Effectively, French coarticulatory nasality is too influenced by contrastive nasality to be studied as coarticulation (in the context of the present work), and the segments are far too coarticulatory to shed light on contrastive nasality for our purposes. As such, although it is certainly interesting fodder for future studies, and the techniques and methods described here would allow effective study with only minor modifications, we will not be further examining French coarticulatory nasality in this paper.

2.2.2 Contrastive nasality in French

In French, we can draw a useful distinction between phonemic "oral vowels", those which are meant to be produced with little-to-no nasal airflow, and phonemic "nasal vowels", which are phonologically specified to include nasal airflow. The latter, referred to simply as "nasal vowels", are of significant interest in the present work.

Although a more precise (and theoretical) discussion of the phonological status of French nasal vowels can be found in the French language discussion in Cohn (1990), it will suffice for the present work to highlight the minimal pairs present (*beau* /bo/ 'beautiful' vs. *bon* /bõ/ 'good' or *bas* /ba/ 'low' vs. *ban* /bã/ 'ban'), and the fact that within the French language, they appear to act as single segments[3]. The precise phonological nature or underlying forms of French nasal vowels are not particularly relevant, as "nasal vowel" appears to be a distinct surface category in French, and perceptually speaking, the surface forms are of greatest importance.

As in many languages with nasal vowels, there is an asymmetry in French between the phonemic oral vowel system and the phonemic nasal vowel system. According to Fougeron and Smith (1999), phonologically speaking, Parisian French has an oral vowel system consisting of /i e ɛ a ɔ o u y ø œ ə/, but has only three nasal vowels, /ɛ̃, ɑ̃, ɔ̃/. According to Delvaux (2006), Quebecois French adds an /a ɑ/ contrast to the oral vowel space, and adds /œ̃/ as a fourth nasal vowel, while maintaining all other contrasts present in Parisian French. However, Fagyal et al. (2006) cite several sources showing that /ɛ̃/ and /œ̃/ are merging in Quebec, so alas, this contrast's days appear numbered.

Their identities and regional distributions aside, the nature of the nasality in these vowels has been well-studied.

Demolin et al. (2003) used Magnetic Resonance Imaging (MRI) to examine the velar gesture in French nasal vowels, providing a detailed look at the actual aperture of the velopharyngeal port during these productions. This provided information about the exact VP port aperture for each vowel, and about the differences in aperture for the different vowels. By comparing the various transversal slices, variation was shown in the shape of the VP port aperture across the four French

[3]There is a compelling argument to be made, based on borrowings, that nasal vowels are underlyingly VN clusters (as argued in Paradis and Prunet (2000)), but this hypothesis, however interesting, is not particularly relevant to the present work.

nasal vowels studied (/œ̃/ is present in the Belgian French dialect used), and /ɔ̃/ was shown to have the smallest opening diameter of all. Considerable inter-speaker variation was found as well.

Unfortunately, MRI techniques at the time of this 2003 study did not permit the capture of the time-course of the nasal gesture (the authors state that each set of slices required 13.8 seconds to capture). Because of this, no direct information about the velar gesture over time is available, and we must look to the next closest approach, nasal airflow.

Perhaps the most comprehensive aerodynamic study is Delvaux et al. (2008b). In phonemically nasalized vowels, as one would hope, there is nasality during the vowel, often with a plateau or central peak. However, there is a great deal of variability in both the degree and timing of this nasality (as shown nicely in Figure 3).

This finding of a frequent plateau or central bump of nasality in phonemically nasal vowels, coupled with variability in the time-course and degree, is mirrored in airflow studies from Cohn (1990).

Interestingly, Delvaux and Demolin show that phonemic nasal vowels can cause contextual nasalization *of the surrounding consonants*, particularly in a carryover setting (ṼC). Although this is quite interesting, it is unlikely to be a useful cue to vowel nasality, as according to the authors:

> "Note that none of the oral consonants that exhibit coarticulatory nasal airflow in the present study sounded nasalized to the experimenters, nor did the experimenters notice any significant effect of nasalization on their acoustic properties. Indeed, in CṼ items the temporal extent of anticipatory nasalization is limited to the last portion of the consonant, and in CṼC and CṼ.CV items, the second consonant is always a voiceless obstruent, so that even heavy carryover nasalization is unlikely to be detected." (Delvaux et al. (2008b), pp. 595)

Finally, it is also worth noting that, although they are outside the scope of the present work, nasal vowels are quite capable of exhibiting coarticulation with surrounding nasal consonants.

In addition to the change in velar configuration, there are differences in oral articulation between oral and nasal vowels. Delvaux et al. (2002b) examines the differences between the nasal vowels and their oral counterparts using both MRI and acoustical measurement. In this study, in addition to the expected differences in velar articulation, the authors find that /ɛ̃/ is more open and centralized than /ɛ/, that /ɑ̃/ is more rounded and back than /ɑ/, /ɔ̃/ is more rounded than /ɔ/ (as well as more back/closed for female speakers), and that /œ̃/ is slightly more open and back relative to /œ/. These findings were confirmed both on the basis of formant measures and in terms of actual tongue position (as measured by MRI). This finding, much like the per-vowel differences in coarticulatory nasality in English, points again to the potential utility of vowel formants and other, more typically oral phenomena as acoustical cues to the presence of nasal vowels. Similar work is performed in Carignan et al. (2015).

So, in summary, French nasal vowels show strong nasality, often with a plateau or even a peak in nasality towards the center of the vowel. However, as with English, the exact nature of the time course of nasality varies by context, by token and by speaker, meaning that even a "nasal vowel" is not evenly nor straightforwardly nasal. In addition, we see that nasal vowels often vary from

Figure 3: Time-Normalized Airflow traces of ṼN words (from Delvaux 2008, Fig. 6, pp. 594)

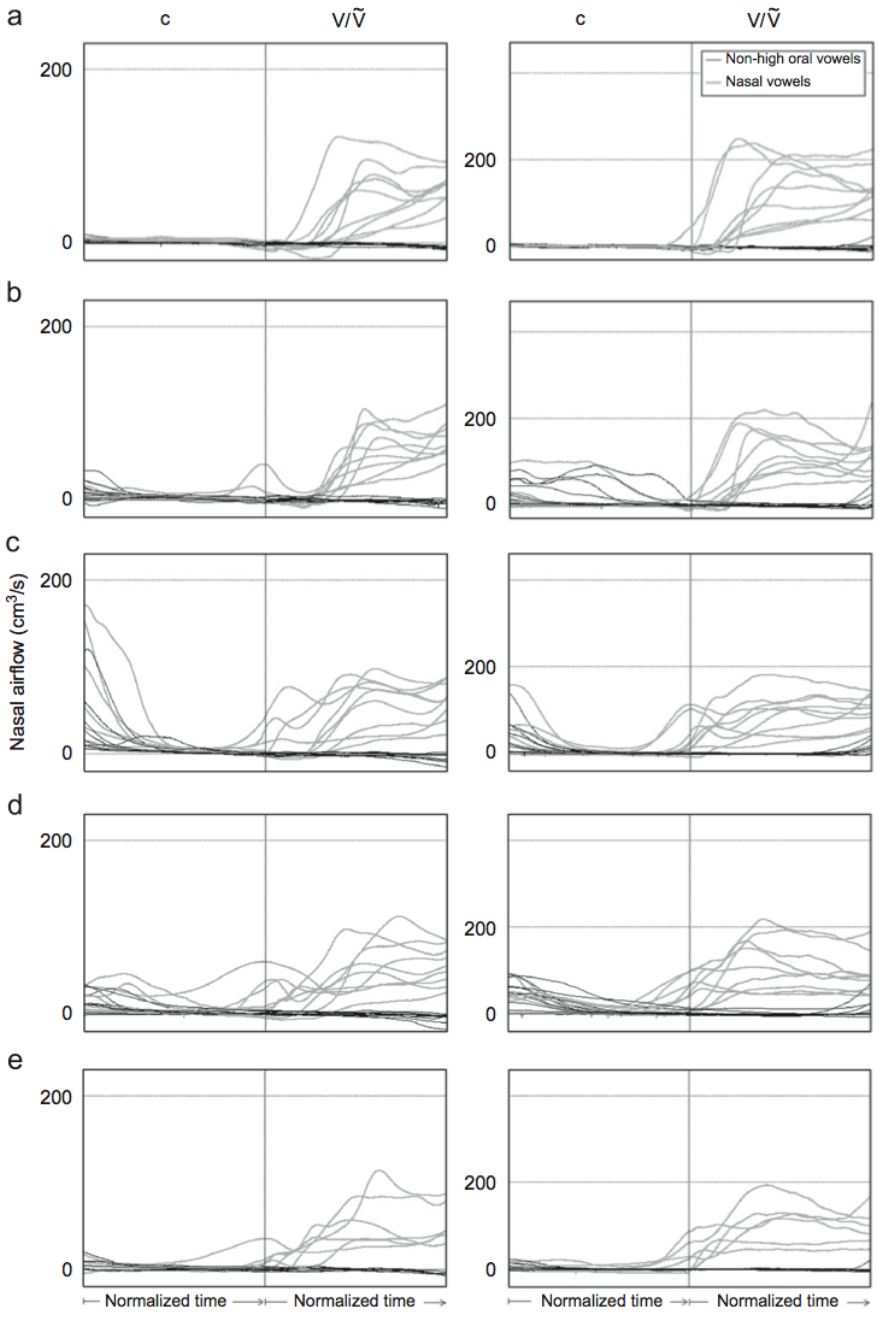


Fig. 6. Nasal airflow outlines for all the oral vowels recorded in CV and CṼitems for female speaker S3 on the left and male speaker S7 on the right: (a) voiceless stops, (b) voiced stops, (c) voiceless fricatives, (d) voiced fricatives, (e) liquids. Time is normalized separately for consonants and vowels.

their oral counterparts, differing in terms of oral articulation, and thus, in terms of the associated acoustical features.

Having established the canonical nature of these nasal vowels, we should now discuss some of the ways in which nasality has been shown to vary from word to word, as well as from speaker to speaker.

2.2.3 Variation in French vowel nasality

As with English, nasality is subject to conditioned variation.

There are variations in vowel nasality which are caused by changes in vowel height. In an effect first reported extensively in Basset et al. (2001) (which focused on coarticulatory nasality, and thus, is not extensively discussed), and then again confirmed in Delvaux et al. (2008b), high vowels in French appear to have far greater degrees of nasality. The origins of this phenomenon are discussed in depth in Delvaux 2008. Although this is an interesting phenomenon, it will not play a major role in the present work as French does not have any phonemically nasalized high vowels.

Delvaux et al. (2012) shows variation in the time course of nasality between Northern French (as spoken in Belgium) and Southern French (as spoken in France). These differences were solely in the temporal domain, showing variation in terms of vowel duration and in what the authors describe as “differing temporal organizations” of nasal vowels. The authors make no claims as to changes in the degree of nasality, nor do they claim any substantial difference in articulation outside of the temporal domain, but it does appear that dialect is a relevant factor in explaining vowel variation in French.

Finally, Roche et al. (1998) uses nasalance to examine the nasality of nearly 500 Canadian speakers of French and English. Nasalance here refers to the use of a plate pressed against the upper lip to collect the separate measurements of amplitude of the signal from the mouth and nose. In addition to some variations between speakers from Eastern and Western Canada, the primary findings are that females exhibited higher nasalance scores than males, and that there was a small increase in nasalance associated with age. Interestingly, these demographically-based nasalance differences, although small, manifested both in French and English, and thus represent true by-speaker variability.

Beyond the work described above, there does not appear to be much study of other sources of variability in French nasal vowels. Although coarticulatory nasality in French is extensively studied in terms of variability (Scarborough (2004), Basset et al. (2001), Oh (2008), among others), French phonemic nasality appears to have been largely ignored. But even based on the little research available, it is clear that French nasality is quite complex, and that considerable variability across speakers and dialects should be expected as this research proceeds.

2.3 The Acoustics of Nasality

Human speech perception is, at its most simple level, the process of turning acoustical information into mental abstractions. No matter the abstractness of these abstractions (phonemes, articulatory features, or even whole words), and putting aside supplementary sources of information (visual cues, context, probability of words or sentences), at some point, we must be able to process the incoming pressure waves and, from them, determine something about the speaker-intended abstraction.

However, speech perception is incredibly complex. Even for two people with no hearing or language disorders, sitting in a silent room, the cues which point to a given speech feature are not always transparent. Some are located in the time domain (geminate sounds, for instance). Some are marked largely through amplitude, such as the silence which (occasionally) indicates word boundaries, or the presence of voicing during a stop closure. Some are frequency-based, and rely on our innate ability to attend to the frequency of different sounds (for tone or fricative place, for instance), as well as the relative amplitude of sounds at different frequencies (vowel formants or spectral tilt). As if speech perception were not complex enough, the perception of many phenomena relies on multiple cues combined, for instance, diphthongs and glides are perceived by looking at changes in the frequency of different (amplitude-differentiated) vowel formants over the course of a vowel.

Nasality is a phenomenon which seems to be expressed using all of the feature types described above, and, as such, is complex beyond our present understanding. We have no shortage of acoustical features which are associated with the articulatory act of nasalization, but as discussed previously, we as linguists lack the ability to identify or quantify which characteristics of the acoustical signal represent (or add up to) the percept of “nasality” for listeners, and developing this understanding is the eventual goal of this dissertation.

In order to select the features tested in this process, we will examine the literature in order to extract and evaluate some of the previously-proposed features. These features will be grouped by their acoustical effects. First, we will discuss different types of spectral features, changes of the distribution of energy in the frequency domain. This will include discussion of the so called “nasal peaks” or poles in the signal, where energy is being introduced to the nasal signal, as well as nasal zeroes where an area of the spectrum is damped by the nasal coupling. In addition, we’ll discuss the effects that the nasal coupling can have on the resonances of the overall vocal tract (which leads to changes in vowel formants), and changes to the overall spectral shape of the vowel. Then, we’ll discuss amplitude features which with the amplitude of the signal itself (rather than of a component frequency) at different points in the word. Then, finally, we will look at temporal features, shifts in duration or changes over the course of the vowel or word related to nasality.

But let us begin with what are, perhaps, the most tantalizing of nasal features: the resonances directly associated with the nasal vocal tract, whose presence, one would hope, would simply and strongly indicate nasal coupling.

2.3.1 The Source and the Filter

Before we move too deeply into the specific consequences of nasality, we should back up and discuss the model of speech production most advantageous to this discussion. The "Source-Filter" theory of speech production simply breaks the act of speech into two parts.

The first of these is the "source", the creation of signal by the vibration of the vocal folds. At the level of the source, of voicing, we can modify the pitch of the voice (by changing the frequency of vocal fold vibration), we can change the amplitude of the voice (by increasing the air output and modulating vocal fold tension), and we can change the voicing type, choosing to use whispered, creaky, pressed, or breathy voicing.

The ultimate result, the acoustical "source", is a series of harmonics with a certain fundamental frequency, each one a multiple of the fundamental frequency, but slightly lower in amplitude than the last (falling off at a rate we refer to as "spectral tilt").

This signal, generated entirely at the larynx, must then travel through the remainder of the vocal tract, which has been configured by the speaker in a particular way to "filter" that signal, and thus create a certain speech sound. To create an /i/, for instance, the speaker would close the VP port, raise the tongue and move it forward in the mouth, advancing the tongue's root, and leave the lips un-rounded and open. The end result is a complex cavity, with particular acoustical properties. This cavity, for a generic American English speaking male, might produce a strong resonance centered at 250 Hz, raising the harmonics in the source signal in that region relative to the surrounding areas. This is what we refer to as a "vowel formant", a frequency band in which harmonics are raised. For /i/, we might expect other formants in the 2500 and 3000 Hz range as well.

However, even in acoustics, there's no such thing as a free lunch. In order for one region of the spectrum to grow louder, another area must grow quieter. Thus, we also have areas within the signal which are damped, or reduced in energy. These correspond to the frequencies between the formants. So, in addition to some harmonics increasing in amplitude between larynx and lips, some harmonics in the source signal are damped, and those spectral regions emerge from the mouth far quieter than at their initial generation.

Thus, we can consider the speech process to the unity of source and filter: We create a rather boring series of harmonics with the vocal folds, and then rearrange the various anatomical structures in the mouth to filter that signal into meaningful and perceptible speech.

This is crucial to understand for the present discussion, because although we will spend most of our time discussing changes to the various harmonics within the speech signal, nasality is a phenomenon which exists entirely within the "filter". The vocal folds do not (to the best of our current knowledge) function any differently when the velopharyngeal port is open, and the source harmonics at the laryngeal opening do not meaningfully differ in oral vs. nasal vowels. The only changes to the speech process stemming from nasality come from the addition of a new, complex resonator to the vocal tract.

Most importantly, though, we must keep in mind that acoustically speaking, when we produce a nasal (or oral) vowel, *there is only one filter* which reflects the configuration of both the oral tract

(tongue position, lip rounding, etc) as well as the nasal passages (the VP port's aperture and the prevailing mucosal weather in the sinuses, nasal passages, and various turbinates). It is not the case that nasality is doing something "on top of" or "after" the oral configuration, but instead, the articulatory opening of the velopharyngeal port serves only to add complexity to the already-complex supra-laryngeal vocal tract filter. Nasal resonances and anti-resonances are incorporated into "The Grand Filter" alongside the oral resonances and configuration, and at times, they cannot be de-convolved.

As a result, although we may talk about different "acoustical features of nasality" as individual phenomena for convenience, it is crucial to remember that these are not meaningfully separate or stand-alone phenomena. Instead of discussing different trees in a forest, we are naming and discussing ripples on the surface of a pond, each representing the local sum of a variety of disturbances in a complex system, and none straightforwardly traced to a single cause.

This is a part of what makes identifying nasality in the speech signal so complex: the acoustical effects of nasality are convolved with every other component of the "filter", and the same spectral real-estate is often shared among cues to a variety of phenomena. We will see that nasal cues (like any other cue) are often vulnerable to interference and covariance, and that finding a single cue which is unaffected by other changes to source and filter, although a pleasant fantasy, is unlikely.

The contributions of the nasal coupling to the overall filter, though, are not indescribable. The subtle effects are complex and myriad, but among the strongest effects are the addition of prominences and areas of reduced signal to the overall vocal signal. Huang et al. (2001) (pp. 286) provides an excellent summary:

> "In the production of the nasal consonants, the velum is lowered to open the nasal tract to the pharynx, whereas a complete closure is formed in the oral tract (/m/ at the lips, /n/ just back of the teeth and /ŋ/ just forward of the velum itself. This configuration is shown in Figure 6.12, which shows two branches, one of them completely closed. For nasals, the radiation occurs primarily at the nostrils. The set of resonances is determined by the shape and length of the three tubes. At certain frequencies, the wave reflected in the closure cancels the wave at the pharynx, preventing energy from appearing at nostrils. The result is that for nasal sounds, the vocal tract transfer function V(z) has anti-resonances (zeros) in addition to resonances. It has also been observed that nasal resonances have broader bandwidths than non-nasal voiced sounds, due to the greater viscous friction and thermal loss because of the large surface area of the nasal cavity."

Now that we understand the nature of filtration in the vocal tract, we can start by discussing the two principal phenomena associated with nasality, nasal resonances (also "nasal formants" or "poles"), and nasal anti-resonances (also referred to as "anti-formants" or "zeroes").

2.3.2 Nasal Resonances

When discussing or identifying resonances of any sort, we are concerned primarily with three characteristics. First, we want to know the location of the resonance within the frequency spectrum (e.g. "~250 Hz"). Second, we would like to know the bandwidth of the resonance (does it affect only a small part of the spectrum, or a much larger portion). Finally, we want to know the strength of the resonance, and how much we expect the signal in that area to be amplified (or damped). This strength can be discussed either in terms of absolute amplitude, in terms of relative amplitude (relative to surrounding frequency), or, as will be tested in the first experiment, the prominence of one peak relative to the surrounding harmonics.

We should keep in mind, though, that "peak" or "pole" does not necessarily imply that the frequency range in question will actually be raised in amplitude. For instance, it is not always the case that the harmonic(s) of a nasal pole will be raised in *absolute* amplitude in the nasal sound relative to the oral. A nasal formant may not have sufficient power to keep its harmonics from losing energy. Instead, it may be the case that those harmonics have dropped in amplitude, but have dropped less than the nearby frequencies[4].

With this in mind, we can now discuss some of the specific poles and zeroes which have been posited in the literature, and which we might expect to find in natural speech.

P0 (~250 Hz) Also occasionally referred to as "N1" or "the first nasal formant"), P0 is a low-frequency nasal pole, described most thoroughly as a correlate and measure for nasality in Chen (1997) and Chen (1995), although reference to it was made previously (c.f. Kingston and Macmillin (1995) and Mermelstein (1977), among others) and in a great deal of the subsequent literature on nasality (Stevens (1998), among others).

Chen (1997) describes P0 as occurring "between 250 and 450 Hz" (pp. 2360), usually on the first or second harmonic (H1 or H2), although speakers with exceptionally small vocal tracts may have a higher P0. Chen attributes this peak directly to the resonant properties of the sphenoid and maxillary sinuses, and describes the strength and bandwidth of this resonance as below:

> "[…] the pole- zero pair would introduce a 3.1 dB increase at 270 Hz if the bandwidths of the pole and zero are 120 Hz. If the bandwidths are 80 Hz, the pole-zero pair would introduce a 5.5-dB increase at 290 Hz." Chen (1997) (pp. 2362)

This resonance is, in the author's experience, quite visually prominent when present, as shown in 4, and it has an exceptionally strong effect on the corresponding harmonic. It is generally discussed relative to A1 (the amplitude of the first harmonic), and is measured not alone, but as "A1-P0" (the amplitude of the highest harmonic in F1 minus the amplitude of the nasal peak).

This peak is discussed extensively in Simpson (2012), which examines the interaction between nasality and breathy voice (which is conventionally measured using H1-H2, which are both good candidates for P0's amplification). Simpson's primary claim is that P0 interferes too much with the first and second harmonics for them to be reliably used for non-nasal measurements, which

[4] One can picture a line of 5 lit candles. After an hour of burning, all will be shorter than they started, but a candle which burns more slowly will appear taller than the others.

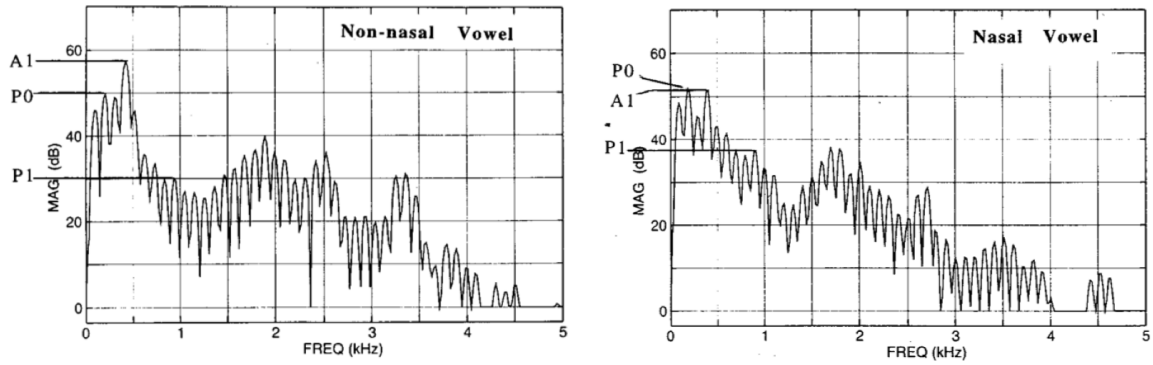


Figure 4: Spectra of Nasal and Non-nasal vowels showing A1, P0 and P1, and the changes in A1, P1 and P0 which occur when a vowel is nasalized (from Chen 1997, Fig. 2, pp. 2364)

leads to his claim, neatly summarized in the paper's title, that *The first and second harmonics should not be used to measure breathiness in male and female voices*.

However, in the same way that P0 can interfere with breathiness' effects on the first and second harmonics, F1 can (and does) interfere with P0. In high vowels, where F1 strays into the 250-450 Hz range, or when the speaker's fundamental frequency is greater than 250 Hz, P0 cannot be readily identified nor measured.
Chen does offer a (rather complex) equation which attempts to adjust the amplitude of P0 for the presence of nearby vowels using factors like the frequency and bandwidth of F1, as well as the frequency of P0. However, this is not a solution, especially in situations where P0 is completely subsumed within F1's prominence, as in many high vowels.

P1 (~950 Hz) Chen also discusses a second nasal resonance, which she refers to as P1, in the area around 950 Hz (pictured in 4). Chen (1995) describes P1 as having a frequency range from 790 to 1100 Hz, with an average of 950 Hz, and offers a bandwidth of "about 250 Hz". Chen offers no absolute intensity difference for P1, instead discussing the intensity of P1 relative to A1 (as "A1-P1").

Although A1-P1 is usually considered a good alternative for high vowels (where A1-P0 is not measurable), P1, like P0, can be affected (or overwhelmed completely) by nearby formants as well, and P1 is vulnerable to interference both from F1 and F2. Again, a correction function based on the properties of the formant is offered, but again, it is only helpful if the overlap is not substantial enough to prevent the proper choice of P1.

It is worth noting that both P1 and P0 are "found" for an individual speaker using relative and impressionistic criteria ("Look for the highest peak in [range] in a known-nasal vowel"). This is not unreasonable, given the cross-speaker differences in resonance that one might expect, but as such, it's less-than-ideal to work with these peaks automatically or on a vowel-by-vowel basis, especially for oral vowels (where there is no "peak" to find).

In the author's personal experience, the interference from two separate formants and variation in its location mean that this peak can be exceptionally difficult to find, even by hand, and that

consistent measurement of P1 can be quite difficult, more so than P0.

P2 (~1250 Hz) Finally, there exists a higher nasal pole, called P2, only discussed in a single source (Schwartz (1968)). Schwartz proposes that this peak is around 1250 Hz. No bandwidth nor amplitude estimates are offered, and again, no deterministic approach is offered for measurement, and the researcher is instructed to find the high peak in that rough area in nasal vowels.

Although it is not described in detail in any other literature on nasal acoustics, recent work from a colleague (Price and Stewart (2013)) indicates that it mirrors P0 and P1 in indicating the nasality of a vowel (and in fact, A1-P2 proved more robust than A1-P1 in their data).

Again, this peak is vulnerable to interference from formants (primarily F2 in this case), and although Chen's (fairly generic) adjustment function could be adapted here, no such function was proposed by Schwartz.

P0, P1 and P2 represent the sole nasal formants which have been proposed in the current literature. Although each pole has situations in which it can provide evidence for vowel nasality, interference from other parts of the filter renders each unreliable in certain contexts. So, we see that these poles provide useful evidence in favor of nasality, but do not themselves provide a nasal "smoking gun".

That said, in acoustics as in life, all poles must necessarily come with zeroes, and in many cases, it is the zero that gives pole meaning. So, rather than focusing solely on those frequencies amplified by the nasal coupling, we must now shift into those dark regions where energy is lost.

2.3.3 Nasal Zeroes

In addition to thinking about the new resonances of the nasal cavities, we must also consider the holistic effect of adding another resonator to the vocal tract. Due to this additional resonator, some resonances otherwise present in the oral cavity will disappear, and some frequency bands may be reduced in amplitude relative to oral speech. In addition, there is some sound which will flow into the nose and then be reflected back into the oral cavity. In certain frequencies, the reflected sound will be exactly out of phase with the original sound, and the peaks will cancel the troughs, and vice versa, thus completely eliminating some frequency bands from the signal (a phenomenon called "phase cancellation", forming the basis of most noise-canceling headsets in modern use).

Taken together, these damped frequencies and phase cancellations result in reductions of amplitude in certain regions of the frequency spectrum, a phenomenon which we refer to as "zeroes" or "antiformants". Given that these zeroes are just as indicative of nasal coupling as our previously discussed nasal peak, it is not unreasonable to think that humans may rely not only on finding nasal peaks, but on locating and measuring the strength of these nasal zeroes as well.

Zeroes associated with P0, P1 and P2 "Nasal poles" are, in many ways, more aptly described as pole-zero pairs. Both P0 and P1 are described in Chen (1997 and 1995) as having an accompanying

zero, with bandwidth roughly equivalent to the pole, usually above, but with some effect below. Given the laws of conservation of energy, a similar zero doubtless exists for P2, but is not described in the literature.

Identifying these pole-specific zeroes is less productive than one might expect, in part because the adjacent zero, practically speaking, serves mostly to identify the presence of a pole. As a result, these zeroes are largely ignored, with focus going to the pole. The present work does aim to capture their presence by examining what will be called the "prominence" of P0, P1, and P2, measuring the height of the harmonic representing the pole relative to the average height of the two adjacent harmonics.

These pole-specific zeroes are easily attributable to the same specific resonance phenomena which created the accompanying poles. However, not all zeroes are easily associated with a particular phenomenon. We will now discuss a few damping phenomena which are often present, but whose genesis is less easily explained.

Damping of F1's Amplitude ("A1") The most commonly discussed form of this damping in nasality is in the region of the first vowel formant, F1. This damping manifests as a sharp reduction of the amplitude of F1 (henceforth, the amplitude of F1 will be referred to as "A1") relative to the surrounding signal. A1 is relevant to the study of nasality in two ways.

First, the damping of A1 plays an integral part of the commonly used A1-P0, A1-P1, and A1-P2 relative measures discussed previously. These measures are able to show nasality more strongly by comparing a quantity expected to go up

The damping of the first formant is discussed extensively in the literature (as in Stevens (1998), Schwartz (1968), Delvaux et al. (2002b), Macmillin et al. (1999), and Pruthi and Espy-Wilson (2004), Delvaux (2009), among many others). It is worth pointing out, though, that the formant itself is not targeted. There is no particular phenomenon which seeks out and damps only the frequencies associated with the first formant (e.g.), but instead, they're a convenient-to-measure waypoint in a larger region which is damped due to phase cancellation, anti-resonance, or in some cases, the overlap of a pole-specific zero.

Given how often this feature is discussed in the literature on nasality, as well as its long history of association with the acoustics of nasality, one would expect a relative drop in F1's amplitude to play a strong role in the perception of nasality.

Damping of F2 and F3 The amplitudes of F2, and F3 (referred to as "A2", and "A3") relative to the surrounding signal have also been proposed as useful features for the description of nasality (Delvaux (2009), Schwartz (1968), among others). Again, we do not know the exact cause of these reductions in amplitude, but there is no shortage of plausible explanations, and these will certainly be analyzed as possible perceptual features.

2.3.4 Overall Reduction in Amplitude

So far, we've been discussing features of the distribution of power in the spectrum of the sound. But, the amplitude of the vowel on the whole is a relevant feature as well. Both Delvaux et al. (2002b) and Dickson (1962) mention that the amplitude of nasalized vowels is lower than that of oral vowels, and this makes intuitive sense, as the zeroes discussed above are generally wider and deeper than the poles, not to mention the considerable thermal loss and friction introduced by pushing air through the nasal cavity[5].

This amplitude drop is not necessarily consistent across the entire spectrum. Schwartz (1968) references an overall drop in amplitude, but states that "these energy losses have been studied and have been shown to affect the low frequencies more than the high" (pp. 133). This can result in a sort of spectral tilt due to nasality, as described in Atal (1985). This suggests that absolute amplitude (the measure of the volume of the overall signal), which is already complex due to differences in the amplitude baseline across recordings and microphones, may be less useful than the relative amplitudes of the high and low frequency sections of the vowel, and provides the possibility of a better approach to measurement.

As we have seen above, nasality varies throughout the word. As nasality causes an amplitude drop, Pruthi and Espy-Wilson (2004) mentions the idea of the variance of the amplitude across time in the vowel (variance in the "amplitude envelope") to be a useful parameter for detecting the presence of nasal consonants, and thus, for detecting nasal vowels.

Given these prior findings, as well as this acoustical fact, overall vowel amplitude, amplitude of different frequency bands, as well as amplitude relative to the per-speaker oral vowel average, are all potentially useful features for the perception of vowel nasality.

2.3.5 Changes to vowel formants

Formants have already been discussed in terms of the amplitude of their peaks, but there is evidence that nasality modifies other aspects of the vowel formants as well.

First, there is the possibility that formants are changing due to shifts in the oral articulation of nasal vowels. Some phonetic centralization of contrastive nasal vowels is far from unusual in language, and in many languages, phonologically allowable nasal vowels are a subset of the possible oral vowels in the language. But even when a nasal vowel is nominally in the same position (e.g. "High, Front, Unrounded"), studies report shifts in oral articulation relative to the oral vowel. In addition to the prominent study by Krakow, Beddor, Goldstein and Fowler (Krakow et al. (1988)), Delvaux et al. (2002b) mentions formants changing in frequency (indicating a change in tongue position) in nasal vowels in French, Shosted and Carignan mention a similar effect in Hindi (Shosted et al. (2012)), and both Pruthi and Espy-Wilson (2004) and Carignan et al. (2011) describe formant shifts in even in English, where nasality is not contrastive. Formant location and bandwidth shifts are also discussed as cross-linguistic facts in Kingston (2007).

[5]This strength of this amplitude drop seems unexpected at first, but is wholly understandable given the difference in amplitude between a sustained /u:::::/ and a sustained /m:::::/. Clearly the nasal cavities are not overwhelmingly efficient conductors of sound.

However, even in the (unlikely) case that oral and nasal vowels are articulated with the same tongue positions, the poles and zeroes discussed above could quite easily cause formant shifts. Because formants are, as mentioned previously, simply mountain ranges formed of many harmonic hills, the addition of an area of prominence near an existing formant will necessarily change its perceived width and central frequency. This, alone, could explain the sorts of formant shifts discussed in Kingston (2007), Delvaux and Huet (2006), and Delvaux (2009) (among others). These nearby poles can cause the broadening of the formants (specifically F1), which has been mentioned repeatedly as consequence of vowel nasalization (in Pruthi and Espy-Wilson (2004) and Arai (2006), among others).

As mentioned previously, one pleasant consequence of the single-signal nature of speech is that we need not distinguish between formant changes resulting from the nasal coupling and formant changes resulting from changes in oral articulations during nasal vowels. Although the two "types" of change may stem from two separate phenomena, ultimately, they both act in concert on the signal within the mouth and vocal tract, resulting in the net change in the formants audible to the listener. By simply working with this overall change in formants from oral to nasal, we are able to measure or manipulate changes in formants without making (or even needing to make) claims about the articulatory source (oral or nasal) of the change.

So, no matter their source in the signal, formant frequency shifts and bandwidth changes seem to correlate well, and should doubtless be considered as potential perceptual features for nasality.

2.3.6 Temporal Features

Finally, as we've mentioned previously, we cannot consider nasality to be a steady, binary phenomenon, especially in necessarily dynamic coarticulatory contexts. Beddor and Krakow (1999) manipulate the time course of nasal vowels to perceptual effect, and Stevens et al. (1987a) mention changes in perception due to modification of the duration of vowels, as well as to the duration of nasality within the vowel. This lends some credence to the idea that vowel duration, as well as duration of nasality within the vowel, will be a useful and relevant feature for nasality perception.

In addition, although there are no studies examining this (likely due to the considerable difficulty of modifying nasality in natural tokens), it is reasonable to suspect that the contour and time course of nasality over the course of the vowel provides useful information in disambiguating the acoustical effects of nasality from those of other, similarly-realized phonetic phenomena (such as voicing type).

Although manipulating the time course of nasality as a potential perceptual cue is necessarily out of the scope of this work, during machine evaluation of the data, several features will be evaluated across multiple timepoints, in an effort to capture the nasal status of the whole vowel, rather than at a single point. Similarly, in the human perceptual experiments, efforts will be made to ensure that the time course is preserved during modification of the vowels.

2.4 Can cues to vowel nasality differ across languages?

Given that this project will be considering (and comparing) both nasal production and perception across two very different languages, a bit of discussion of the possible sources of variation is warranted.

Cohn (1990) rather neatly showed that the nature of the airflow trace (in terms of timing, slopes, plateaus, and effect on surrounding consonants) varied substantially among English, French, and Sundanese. This establishes that, at one level, nasality is most certainly performed differently in different languages.

However, variations in the time course and degree of nasality are not particularly relevant in this work. The fundamental question being asked here is not "How do listeners recognize *French* nasality?", but instead, "How do listeners recognize nasality at all?". Certainly, the presence or absence of the expected temporal pattern of nasality will play a role in classification in a communicative context. However, the ability to make judgements about the time course, pattern, and degree of nasality presupposes the ability to perceive the presence or absence of nasality, and it is that more fundamental ability which we are interested in here.

So, rather than asking about differences in pattern or implementation, instead, we must ask if we expect the *fundamental nature* of the production and perception of nasality to change across languages?

A priori, one would not expect any differences in nasality at this level of functioning. We in the field of phonetics generally hold that speakers of different languages are not meaningfully different in physiology, and therefore, speakers of French are unlikely to have different nasal resonances and acoustical effects of oral-nasal coupling than speakers of English. The laws of physics and resonance apply identically in both populations, and thus, any differences must come down to articulation.

One source of variation could be the velar gesture itself. The velum ('soft palate') is raised and lowered by the action of the aptly named Levator Veli Palatini, whose contraction raises the velum and seals the velopharyngeal port. Demolin et al. (2003), previously discussed, make the following observation from MRI study:

> "The measurements of the opening of the velum for French nasal vowels, which are all rather low (between low and mid-low) do not suggest that there is a significant difference of activity of the levator palatini. This implies that there is only one position of velic lowering for French nasal vowels and that the velopharyngeal port is open passively as a result of the relaxation of the levator palatini as suggested by Bell-Berti." (Demolin et al. (2003), pp. 466)

This implies that there is an anatomical ceiling for nasality which is reached, and that the velar articulatory target for nasal vowels is simply "maximum VP port aperture". This would seem to imply that the performance of vowel nasality, at least in terms of degree, is the journey to and from a passive state, rather than an active performance of itself (which may be aimed at, say, optimization or differentiation of the perceptual cues to nasality). This seems to lessen the possibility of the velum as a nuanced and purposeful articulator, and to shift some of the burden

of additional performance to oral articulation.

This means that our search for differences in the performance of nasality must shift instead to the articulations going on within the oral cavity. Both Demolin et al. (2003) and Delvaux et al. (2008b) mention the possibility of changes to oral configuration leading to differences in nasal coupling and airflow. Perhaps more convincingly, Shosted et al. (2012) shows changes in oral articulation which serve to heighten the spectral characteristics of nasal vowels, and as mentioned earlier, Carignan et al. (2011) makes the claim that oral articulations are actually serving to counterbalance some of the acoustic characteristics of English nasal vowels. Demolin et al.'s findings of a largely passive velar articulation, coupled with Shosted and Carignan's (et al.) findings of enhancing differences caused by changes in oral articulation, can lead to a very different view of the idea of a "nasal vowel".

Perhaps it is the case, as suggested by Shosted et al. (2012), that a nasal vowel is not just an oral vowel with nasal aperture, but instead, is a paired combination of oral and velar gestures designed to produce a particular acoustical (or perhaps perceptual) effect which is holistically considered to be a "nasal vowel".

This is supported, at least in part, by the highly frequent asymmetry between oral and nasal vowel systems. According to an UPSID (UCLA Phonological Segment Inventory Database, Maddieson (1980)) search done in Kingston (2007):

> "Taking up size first, a little over half the languages with nasal vowels, 53 of 102, have fewer nasal than oral vowels in their inventories, and none have more nasal than oral vowels. The languages with fewer nasal than oral vowels have on average 2–3 fewer, and some as many as 6 fewer. […] In short, nasalization reduces height contrasts, and it does so most often by eliminating mid vowels."

This asymmetry of inventory implies that nasal vowels are, at least phonologically speaking, fundamentally different beasts than oral vowels (rather than being the result of the addition of a [+NASAL] feature to an existing oral vowel quality). If this is the case, there is no particular reason for nasal vowels' oral articulatory components to mirror oral vowel articulations. This idea, that nasal vowels may have distinctive oral components, when coupled with the fact that it isn't possible to deconvolve the oral and nasal gestures' contributions to the holistic acoustical notion of "nasality", we find that this is the best potential source of cross-linguistic differences in the nature of (and in the perception of) nasality itself.

French speakers, as mentioned, have a much greater need for nasality to be sent and received clearly, and therefore, they have a motivation to tailor their oral gestures in such a way as to enhance nasality. This possibility is supported by the finding that the quality of French nasal vowels is significantly different (c.f. Delvaux et al. (2002b)) than their oral counterparts. Although these differences *could* be unrelated to nasality, leading these changes to serve as secondary cues to nasality, but since communication tends towards optimization, and it's more likely that these differences serve to enhance the production or perception of nasality itself.

Thus, much like the purposeful, practiced and optimized gait of a professional sprinter will differ from that of the author attempting to catch a bus, it is possible that French speakers' holistic performance of nasality will differ fundamentally from that of speakers of a language where nasality is

not an end of itself. If this is the case, then, it stands to reason that French listeners would become attuned to that particular practiced performance of nasality, and thus, that French speakers may be listening for and perceiving nasality in a meaningfully different way than English speakers. So, as we proceed into our discussion of the acoustical features of nasality (in general), we must keep three ideas in mind.

First, we must remember that although the timing and degree of nasality will vary greatly across languages and tokens, the temporal nature of nasality is secondary information, available only after the initial identification of the presence of nasality in the signal. Thus, although time course and degree of nasality are most certainly interesting and useful in a per-language perception context, they are not particularly relevant to the present line of research.

Second, language aside, there are acoustical facts about the mouth and nose and their coupling which will not differ meaningfully across human speakers of any language. These will form the invariant, inescapable basis of nasality on which all languages must build, and from which all optimizations must begin.

Finally, we must keep in mind that there is more to the production of nasal vowels than the velar gesture, and that nasal vowels are not simply oral vowels produced with lowered velum. Speakers of languages who depend on nasality being reliably produced and perceived are likely to modify the oral components of the nasal vowel articulation in order to enhance both the ease of production and the perceptual salience of vowel nasality. Similarly, speakers of languages which do *not* use vowel nasality phonologically could easily modify the gesture to actively counteract the acoustical consequences of nasality.

So, although there are certainly invariant features of nasality which should remain the same among speakers of different languages, there is ample room for cross-linguistic variation in the implementation and perception of nasality.

With this understanding of what acoustical phenomena appear to reflect nasality in a laboratory setting, and how we might expect them to differ, we can now examine the more relevant question: What features are actually *used* by listeners to detect nasality in vowels?

2.5 The Perception of Vowel Nasality

Although there have been some papers which examine the perception of nasality, at this point, there appears to be little work which explicitly attempts to isolate cues to nasal perception, and none which aims to match those cues with extracted acoustical features of nasality. As such, the perceptual studies described here are all related to the question at hand and merit discussion, but none address it head on.

Unfortunately, those papers in the literature which have aimed to identify the acoustical correlates of nasality (discussed above) conducted no human perceptual study. Although identifying features is a strong step towards finding cues, there remains the possibility of strongly linked features which are not ultimately used by humans for the perceptual task of differentiating oral and nasal vowels. Thus, it is wise to consider perception explicitly, above and beyond studies of the acoustical features of nasalization.

2.5.1 The Perception of Vowel Nasality in English

There has been relatively little experimental work in nasality perception. Perhaps most relevant to the present work are Macmillin et al. (1999) and Kingston and Macmillin (1995). These studies, presented explicitly as two parts of the same work, examined the interaction between F1 and nasalization in human perception.

In the first study (Kingston and Macmillin (1995)), the authors tested their hypothesis that F1 and nasality are perceptually integral (that is, mapped to the same perceptual dimension) by synthesizing stimuli which varied in terms of positions of F1 and the nasal pole and zero, all using a Klatt synthesizer. By adjusting the relative locations of the poles and zeros and presenting them to listeners in a vowel nasality classification task, the authors found that lowering the frequency of F1 often caused listeners to report vowels as nasalized (and vice versa), and showed an integration of the two variables (in that changes in nasality affect the perceived height of F1, while changes in F1 affect the perceived nasality of the stimulus). This supports the idea that the perception of nasality is complex, and that although the pole-zero complex measurements currently in use (A1-P0) may be useful, they are also affected by factors other than nasality.

The second study, Macmillin et al. (1999), aims to find a reasoning for this interaction, again by generating varied stimuli using Klatt cascade synthesis and presenting them to listeners. In this case, stimuli varied not just on F1 frequency and nasal pole and zero, but also by time course of those three factors over the course of the word. Listeners were given the stimuli and ask to place them into previously-trained categories (one per stimulus preparation), and confidence measures were taken. These trials confirmed the integrality proposed in Kingston and Macmillin (1995), and showed that the integrality is greater in vowels followed by a nasal consonant (in an interesting parallel to Beddor and Krakow (1999)). The authors report that spectral center of gravity is a very prominent factor in this integrality (as it is easily affected by both F1 and the nasal pole).

A second experiment was performed using a finer set of distinctions in F1 and the location of nasal poles and zeroes, here classifying (depending on the task) vowels as either “High” or “Mid”, or “Oral” or “Nasal”. In this experiment, the authors found that listeners were most likely to judge nasalized vowels as “high”, and, as expected, “mid” stimuli were more likely to be judged as oral. The authors, interestingly, attribute this in part to the fact that in English, height is contrastive but nasality is not. Also of note is that no interaction is found in this experiment with consonantal context, unlike in the first experiment.

Although the methodology does leave several things to be desired (the task is not terribly linguistic and the stimuli are quite artificial), the most relevant takeaway from both of these studies is that nasality is closely related with F1, and that the perception of nasality affects judgement of the first formant, and vice versa.

One other pair of studies merits discussion here, even though they address neither English nor French.Hawkins and Stevens (1985b) conducted a study of the perception of nasality using speakers of Gujarati and Hindi (in which nasality is phonemically contrastive), Bengali (in which nasality is questionably phonemic), and English. Based on acoustical modeling, they predicted that nasal vowels can be synthesized by replacing F1 with a “Pole-Zero-Pole” complex (appearing, roughly, to refer to P0 and A1, with a zero between). In this experiment, a continuum of stimuli

were synthesized using a Klatt synthesizer, and then modified to exhibit the spectral properties being tested. Listeners were then asked to rate the nasality as "oral" or "nasal" for each point on the continuum. In a subsequent discussion of this experiment in Stevens et al. (1987b), the authors summarize their findings as

> "It seems as though the vowels that are perceived as nasal have the common property that the spectral peak corresponding to the formant is replaced by two peaks which are 200-400 Hz apart, and are the result of a Pole-Zero-Pole combination."

In the discussion of Hawkins and Stevens (1985b), the authors frame the finding slightly more specifically:

> "Another way of describing the stimuli [identified as nasal] is in terms of the degree of low-frequency prominence in the spectrum. […] As we introduce the pole-zero pair with an increased spacing, we are reducing the degree of prominence of the F1 peak in the spectrum, so that a single narrow spectral peak no longer dominates the low-frequency range. This reduced prominence is achieved by creating an additional spectral peak near F1 or by splitting or broadening the F1 peak"

Thus, the authors predict that an increase in F1's bandwidth or increased prominence of P0 will result in the perception of nasality.

Although this study only evaluates one potential cue to nasality, and does so in an isolated and unnatural listening condition, this is perhaps the closest analog to the present study, and provides the strongest prediction about the perception of nasality in the present literature.

2.5.2 The Perception of Vowel Nasality in French

There is comparatively little work on the perception of nasality within the French language, particularly from an acoustical point of view, and the majority of what work exists is itself written in French, by a small group of researchers.

The research which most closely maps to the present work is Véronique Delvaux's 2009 study *Perception du contraste de nasalité vocalique en français* (henceforth, Delvaux (2009)). This study, in many ways mirrors the goals of the present work. Delvaux begins her study describing the state of the art of nasal perception, and describes the situation as below[6]:

> "In summary, the acoustics of nasality is particularly complex, and the link between acoustics and perception is not transparent, except that listeners, even non-native, seem to treat nasal vowels as a natural class distinct from the oral vowels." (Delvaux (2009), pp. 25, trans. Will Styler)

Thus, we see that in 2009, even in the francophone world, the problem of the perception of nasality was far from solved, and Delvaux views the question as both rich and open for discussion. After discussing some of the potential acoustical features of nasality (which will be mentioned in detail

[6]This paper is written entirely in French. Any quotes or paraphrases are rough (albeit well-intentioned) translations by the present author.

later in Section 2.3), Delvaux discusses a series of experiments which tested three hypothetical features of nasality which could be perceptually useful:

1. Lowering of the amplitudes of F1 and F3
2. "The clarity or darkness of timbre" (shown by the frequency of F2)
3. The vowel's duration

All stimuli were generated using a Klatt synthesizer, creating stimuli at various points along a spectral prominence and F2 height continuum, and created long and short versions at each point. These were presented to listeners as part of four experiments. The first experiment was an identification task (using oral/nasal minimal pairs *in a coarticulatorily nasal context*). The second experiment mirrored the methodology described in Macmillin et al. (1999), and used a discrimination task to examine possible perceptual integration between F2's frequency and the amplitude variations in F1 and F3. The third and fourth experiments mirrored the first two, but used native English speakers unfamiliar with the French language. The four experiments delivered a number of interesting findings on the perception of nasality.

First, by analyzing the identification ("oral vs. nasal") experiments, it was found that all three factors proposed examined (F1/F3 amplitude, F2 frequency, and duration) contributed to the positive perception of vowel nasality. This held true for both English and French speakers. Unfortunately, they were not tested in such a way that the features' individual contributions can be gauged in the absence of the others, but according to Delvaux, the statistics do suggest that the F1/F3 amplitude variation "explained the largest amount of variation" (pp. 44). In fact, according to Delvaux:

> This confirms the results of earlier work which suggests that spectral prominence is related to the subjective impression of nasality, independent of the listener's native language. (pp. 44)

Analyzing the perceptual integration data, we see that indeed, this F2 variation is both covariant and perceptually integrated with the lowering of F1 and F3's amplitude. Delvaux goes on to discuss the (acoustically necessary) changes to F2's apparent frequency with the lowering of F1 and F3. Thus, Delvaux argues:

> "The covariation of F2 and [the F1/F3 prominence reduction] could be the necessary product of the lowering of the velum, and the result of a strategy developed by speakers – and fossilized in language – aimed to assure the robustness of the oral/nasal contrast, a strategy which uses the mechanisms which make up the general human auditory system." (pp. 53)

Although the use of synthesized stimuli, identification in a nasal coarticulation context, and a far smaller list of features set this study well apart from the present work, it does provide a valuable hint at some potential perceptual features of nasality, and serves as evidence that the present approach (isolate features and test in identification) can yield valuable information.

Finally, we should discuss Delvaux and Huet (2006), a prior study by Delvaux and Kathy Huet, which examined the perception of nasality in Belgian French. This study performed very similar identification experiments (finding nasal vowels in nasal context) to those described above, using

Belgian speakers, and identified F2's frequency as the primary cue to nasality among speakers. As the 2009 study can be viewed as a direct follow-up, providing additional detail to the findings of Delvaux and Huet (2006), in light of the detailed treatment of Delvaux (2009), this earlier study does not merit discussion beyond mention.

So, we see that although some work has been done to examine the perception of nasality in French, the exact nature of the perception of nasality is still very much indeterminate, and the more detailed examination proposed in the present work is poised to provide considerable new information for the field of French nasal perception.

2.5.3 Automated Detection of Nasality

Compared with the acoustics of many other speech phenomena, nasality detection is laughably understudied in the academic literature. At the time of writing, there are few modern papers dealing explicitly with the automatic measurement and detection of nasality in speech, most of which focus on detection of nasal consonants, the most modern and comprehensive of which is Pruthi and Espy-Wilson (2004).

This paper focused primarily, though, on the detection of nasal consonants in speech. In their work, they mention many of the features (and papers) listed above, but in the context of finding and characterizing the "nasal murmur" present for the production of nasal consonants. As such, the findings are not directly transferable to the present work.

That said, their overall accuracies for classifying nasal consonants across a large corpus of data using many of the features discussed here ranged from 87.8% (for intervocalic nasals) to a high of 95.8% (for post-vocalic nasals). This shows that classification in a related problem is possible, as well as giving a basis for comparison. In addition, it shows that the use of Support Vector Machine classifiers (SVMs) with cleverly chosen features is a worthwhile technique, supporting the approach and feasibility of the classification task proposed in the present work.

Of particular interest is a quote from the conclusion of the paper (Pruthi and Espy-Wilson (2004), pp.238):

> *One area which needs work is development of APs [acoustical parameters] to capture nasalization in vowels. We believe this can potentially give a big improvement in the detection of nasals especially because at times nasals might be articulated only as vowel nasalization.*

This again shows that at the time of writing, vowel nasality detection is an unsettled issue, and is of potential use outside of the testing the linguistic questions discussed here.

We should also consider that in the speech recognition world, a great deal of work is being done using Hidden Markov Models, cepstral coefficients, and other forms of analysis which do not rely on analysis in the frequency/amplitude/time domain as humans appear to. Using such methods, the search for discrete "acoustical features" is simply not relevant. Instead, large amounts of annotated data are poured through complex statistical models, creating highly-trained black boxes which identify and classify speech according to parameters which make little sense outside of the models. Although these approaches are unquestionably effective, they require little understanding of the nature of linguistic features, and provide little insight in return.

Finally, we must highlight the sad fact that many advances in automatic speech recognition are likely locked away in the "proprietary information" vaults of large speech technology organizations, and such companies have a vested interest in not sharing information about their methods and techniques (for fear of "aiding the competition" through the free dissemination of knowledge). We must assume, of course, that any speech technology company working with French, Hindi, or other languages which use nasality contrastively have developed methods for classifying nasal vowels (or have chosen an approach as described above which does not rely on features at all), but, unfortunately, such methods are opaque, proprietary, and potentially legally unavailable for discussion or use.

Nasality is a complex phenomenon, which varies significantly from language to language, as well as from environment to environment. Its realization is complex to characterize, as it results from the attachment of a second set of resonators to an already-complex acoustical system, and there is no shortage of interactions between nasality and other factors in speech. There are many potential features which have been discussed as relating to nasality in measurement, but there is no strong consensus as to the "best" feature(s) used for measurement, and even the most-often used have strong disadvantages.

Clearly, there is no magical solution to nasality. In the current literature, there is no one feature which captures nasality perfectly (or even nearly so). There is no one approach which works for all vowels, or in all contexts. In fact, there appear to be nearly as many solutions to the measurement or identification of nasality as there are papers examining the phenomenon. And yet, despite our great difficulty in the linguistic community, human listeners appear to have little trouble hearing, identifying, and even using the most subtle aspects of nasality.

Thus, in order to bring our understanding closer to our perceptual abilities, we must proceed with the measurement and evaluation of the features already known, features hitherto undescribed, and with combinations yet untested, and then evaluate our findings as potential bases for perception, both human, and computer.

But in order to do any of this, we must collect the natural language data upon which all subsequent analyses will be based.

3 Data Collection and Feature Extraction

Although many corpora exist for both French and English, there are a few key issues that prevented the off-the-shelf use of corpus data in this project, for correlations, for machine learning, and as a basis for the stimuli used in the human perception experiments described later.

First, the recording quality of many of the large spoken corpora is simply lower than this task demands. For instance, many corpora have been downsampled (or captured based on phone conversations which are themselves downsampled). This has the effect of removing the highest frequencies from the recording, complicating the evaluation of features which might depend on these highest frequencies, or on the distribution of power throughout the word. In addition, many corpora were recorded across different settings, with different amounts of noise, different microphones, and different methods, adding significant complexity to the precise extraction of these acoustical features from words.

Complicating matters further, those available corpora or datasets which were reliably recorded are not necessarily well suited to this use. Nasalized vowels, especially in English, are viewed by many as a complication and hindrance to the collection of non-nasal data. As such, collected corpora will often have asymmetries of collection which prevent their use for the present work (for instance, only a few NVN words relative to many CVC words, or complete avoidance of nasal onsets or codas).

As a result of these quality and distribution issues, corpora of spoken data will need to be collected as a part of this project.

3.1 Recording method

All recordings took place in the University of Colorado Phonetics Lab, inside our sound-attenuated booth. For all speakers (across both languages), the stimulus list was displayed as a slideshow on a flat-screen monitor in the booth, with words displayed individually, advanced by mouse-click. All data was collected according to the methods specified for data recording in the CU IRB Protocol 13-0668.

English Recordings were made using an Earthworks M30 microphone, recording at a 44,100 Hz sampling rate with 16-bit amplitude resolution using an Apogee Mini-Me Firewire Analog-to-Digital conversion box. French Recordings were made using an Shure Head-mounted microphone, recorded with the same ADC and settings.

Due to a peculiarity of the Hellems Building's HVAC system resulting in a ~3 Hz hum which penetrates our sound booth, all session recordings were filtered to remove the bottom 60 Hz of the signal (10Hz roll-off) using Praat. Upon completion, session recordings were saved in .wav format.

At the time of recording, speakers were each assigned pseudonyms to maintain confidentiality. In all subsequent discussion, speakers will be referred to only using those pseudonyms. The pseudonyms were assigned by dice roll from a list of assorted names.

3.2 The English Dataset

For English, a series of groups of minimal pairs in citation form were solicited for a variety of English vowels, balanced across the four coarticulatory structures principally examined in this work (CVC, CVN, NVC, NVN).

For each vowel, four tetrads are elicited, designed to limit the differences to nasality alone and to provide variety for later use in perception and production experiments. The first set contains the vowel in CVC, CVN, NVC, NVN format with only alveolar consonants. The second set uses only bilabials. The third uses bilabial onsets and alveolar codas, and the fourth tetrad uses alveolar onsets and bilablial codas.

All of the vowel-adjacent consonants are voiced[7], as English vowels are longer before voiced stops (Ladefoged and Johnson (2011)) and this length difference could complicate the study of duration as a cue. Similarly, the homorganic nature of the consonants (all sharing a place of articulation) means that consonantal place cues will not differ across pairs, removing them from consideration in the lexical choice task.

Velar consonants are not present because of the severely limited phonotactics of the English velar nasal (/ŋ/) which prevented the construction of an adequate paradigm. Similarly, due to its restricted distribution, /ʊ/ was not recorded, and ɔ/ɑ were not distinguished due to their frequent merging in the dialect region in which the study took place.

Where there exists a non-phonologically-conditioned hole in the distribution of nasals in English, nonsense words ('non-words') are recorded instead. Attempts have been made to limit the number of non-words by introducing a limited number of two-syllable words. In all cases, the nasalized vowel is in the same syllable as both consonants (e.g. "nimble" syllabifies as /nɪm.bl̩/, "monster" as /mɑn.stɚ/). Some non-words, however, were required to fill out the paradigm.

Because the lexical choice task which is used to test English perception of nasality depends on linguistic function (recovering a word's identity in a case of ambiguity), non-words will not be used to produce perceptual stimuli. However, Scarborough (2012) showed little difference in degree and nature of nasality between non-words and real words, and as such, non-words can still be safely used for the feature extraction and machine classification tasks.

The final dataset, shown in Table 1, was recorded twice per speaker, presented in a single randomized ordering across all speakers in the carrier sentence "The word is X". Speakers saw each word in a carrier sentence, presented in a Powerpoint-style slideshow, and were instructed to read the sentence aloud.

This wordlist then generates 160 tokens per read-through, with 2 read-throughs per speaker, yielding (in perfect circumstances) 320 tokens from each speaker.

For English, participants were recruited using the Linguistics Department's subject pool. Only self-identified native English speakers were recorded.

[7] The inclusion of the voiceless-final word "nit" is an error noticed only after analysis was conducted. However, this word was not used in the human perception tasks, and thus, the final finding should be unaffected.

Table 1: The English Word List

	CVC	**CVN**	**NVC**	**NVN**
/i/	deed	dean	need	neen
/i/	beeb	beam	meep	meme
/i/	bead	bean	mead	mean
/i/	deeb	deem	neeb	neem
/ɪ/	did	din	nit	ninny
/ɪ/	bib	bim	mib	mim
/ɪ/	bid	bin	mid	mint
/ɪ/	dib	dim	nib	nimble
/ej/	dade	deign	neighed	inane
/ej/	babe	bame	maybe	maim
/ej/	bayed	bane	maid	main
/ej/	dabe	dame	nape	name
/ɛ/	dead	den	ned	nen
/ɛ/	beb	bem	meb	memories
/ɛ/	bed	bent	meds	men
/ɛ/	deb	dems	neb	nem
/æ/	dad	dan	nad	nancy
/æ/	babble	bam	mab	ma'am
/æ/	bad	ban	mad	man
/æ/	dab	dam	nab	namble
/ɑ/	dodd	dawn	nod	non
/ɑ/	bob	bomb	mob	mom
/ɑ/	bod	bonfire	mod	monster
/ɑ/	dobson	dom	knob	nom
/aj/	died	dine	snide	nine
/aj/	imbibe	bime	mibe	mime
/aj/	bide	bind	mide	mine
/aj/	dibe	dime	nibe	nime
/ʌ/	dud	dunce	nudge	none
/ʌ/	bubba	bum	mub	mum
/ʌ/	bud	bun	mud	month
/ʌ/	dub	dumb	nub	numb
/oʊ/	doughed	don't	node	known
/oʊ/	bobe	bome	mobe	moam
/oʊ/	bowed	bone	mode	moan
/oʊ/	dobe	dome	noble	gnome
/u/	dude	dune	nude	noon
/u/	boob	boom	moob	moom
/u/	booed	boon	mood	moon
/u/	doobie	doom	noob	noom

3.2.1 Post-processing the English dataset

In order to use the dataset for feature extraction as well as for modification during stimulus creation, all words must be stored separately and annotated with Praat TextGrid files, showing the location of the vowel in the word.

A downsampled version of each session file (sampled at 11,025 Hz) was passed through the Penn Phonetics Lab Forced Aligner (Yuan and Liberman (2008)), which completed initial alignment against the script given to the speakers during recording. Although this alignment was quite good out-of-the-box, the author hand-verified the word and vowel boundaries for the target words for each speaker, paying special attention to vowel-nasal boundaries, providing more precise boundaries for measurement and generation of the phoneme-masked stimuli.

In situations where the speaker produced an unexpected pronunciation which corresponded with another target (e.g. /doʊd/ for "dodd" or /dajn/ for "deign"), the word was labeled as pronounced (here, "dode" and "dine", respectively). Where the pronunciation was other-than-expected (e.g. /boʊbi/ for "bobe"), the word was not labeled and excluded from later analysis. In this way, although not all tokens will be available for all speakers, measurements among vowels will be consistently in the proper category.

Some tokens were rendered unusable for other reasons. In some cases, the speaker coughed or bumped the table mid-word, and did not repeat the sentence as instructed. In others, the speaker trailed off mid-word, leading to a fractured pronunciation. Finally, some speakers simply missed a token (perhaps advancing the slideshow twice. In all cases, these tokens were removed from the session file, but in most cases, because all words were recorded twice, at least one valid exemplar of each word was recorded.

Once the labeling and time-alignment for each word was hand-confirmed by the author as accurate, and unusable tokens were removed, the full-signal (44,100 Hz sampling rate) session file was chopped into individual words, with vowel labels preserved.

3.2.2 About the English data

All participants were undergraduates at the University of Colorado in Boulder, recruited from the subject pool in large undergraduate linguistics classes. Unfortunately, the strongly female gender-bias of these classes is reflected in the subject pool, and as such, only one male speaker ("Max") was recorded (despite active efforts to recruit others). In Table 2, along with the age of the participant at the time of recording (in February 2014) and the state(s) in which they were raised. Two participants ("Annie" and "Greta") declined to give demographic information, although they appeared to be in a similar age group with other participants, and spoke with American-sounding speech.

Due to a recording script error, for the first seven speakers (Daisy, Ellie, Greta, Hazel, Max, Molly, and Olivia), the word "ban" was recorded four times, and the word "bam" was not recorded. For these speakers, all 4 tokens of "ban" were kept and analyzed. For later participants, the error was fixed and both "ban" and "bam" were recorded twice as planned.

Table 2: The Recorded English speakers

	Psuedonym	Age	Home State
1	Olivia	20	Tennessee/Colorado, USA
2	Daisy	19	Colorado, USA
3	Hazel	18	Colorado, USA
4	Molly	19	Colorado, USA
5	Max	21	New York/Illinois, USA
6	Greta	-	-
7	Ellie	18	Colorado, USA
8	Sue	19	New Jersey, USA
9	Isabella	18	Hawaii, USA
10	Amelia	19	South Carolina/Colorado, USA
11	Emily	20	Wisconsin, USA
12	Annie	-	-

It is also worth highlighting that 3 of the participants (Olivia, Daisy and Emily) displayed exceptional amounts of creaky voicing ("vocal fry") in their recordings, exacerbated by the sentence-final position of the target word within the carrier sentence. Olivia, particularly, produced practically no modal voicing in her recordings, and in many cases, the vowel itself was produced as a large gap between consonants with 3-6 aperiodic glottal pulses. Although their speech was still recorded and annotated, this creak will almost certainly cause measurement and modification difficulties, and these speakers will be analyzed separately, where prudent.

Ultimately, 12 speakers were recorded, resulting in a total of 3,823 tokens for analysis.

3.3 The French Dataset

The French dataset for this project need not be so lexically constrained as the English data, as the comparisons will be across vowel type, rather than across phonological structure, and because the final perception task involves identification of vowels in isolation, rather than in ambiguous phonological context. This flexibility allows improved control for the data, and negates the need for non-words.

The French data recorded for this project is comprised of 30 CVC/CṼC minimal pairs, differing only in the phonological nasality of the vowel. To find these pairs, a Python script was used to find all word pairs in the Lexique corpus (New et al. (2001)) which met the below requirements:

- Monosyllabic Nouns
- Identical consonantal context surrounding the vowel
- All instances of nasal vowels (/ɛ̃, ɑ̃, ɔ̃/) are replaced with oral vowels (/ɛ, ɑ, ɔ/)
- No glides (/j/ and /w/)

275 CVC/CṼC pairs met this criterion. In order to avoid picking highly infrequent or unusual words, and to avoid frequency-specific variations in nasality, the final pairs were chosen from this

list in such a way to maximize the log lexical frequency of the words chosen, while minimizing the difference in lexical frequency between members of a pair. In addition, words which were present in Lexique but not present in the Larousse online dictionary (Larousse (2014)) were excluded. 10 words per nasal/oral vowel pair were chosen, making a total of 60 minimal-pair words. As with the English dataset, all words were recorded twice, both to increase the number of tokens, as well as to provide flexibility in generating stimuli during later steps. This resulted in 120 words per speaker being recorded.

The finalized dataset is shown below in Table 3:

As a practical expediency, this data was collected in coordination with Dr. Rebecca Scarborough. Six speakers were recorded by Luciana Marques, a phonetician in the CU Phonetics Lab, and two were recorded by Georgia Zellou (at University of Pennsylvania). The French words for this project were recorded alongside a set of lexical-neighborhood-controlled words for a separate project, and were presented randomly in a single block with the additional words. All words were recorded in the carrier sentence « Dites [word] s'il vous plaît ».

The decision to include two speakers from outside of Colorado stems from the relative sparseness (and tight-knit nature) of the Francophone community in Boulder, CO. In addition to getting two more speakers, by recording two people entirely divorced from the French community in Colorado and using their data to create stimuli for the perceptual experiment, any French speaker can be freely recruited in Colorado with vanishingly low chance of the listener being familiar with the speech of that particular speaker, including, potentially, speakers who participated in the production study.

3.3.1 Post-processing the French dataset

As with English, in order to analyze the French data effectively, the labels must be time-aligned using Praat TextGrid files.

Because the Penn Phonetics Lab Forced Aligner does not work in French, the author was able to obtain an early release of the SPLAligner French Forced Alignment tool from Peter Milne (described in Milne (2014b) and in an unpublished manuscript Milne (2014a)). This is a Python reimplementation of the Penn Forced Aligner, using Markov Models trained on French data.

As with English, downsampled version of each session file (sampled at 11,025 Hz) was passed through SPLAligner, which completed initial alignment against the script given to the speakers during recording. Luciana Marques then verified the word and vowel boundaries for all recorded words for each speaker, paying special attention to vowel-nasal boundaries, providing more precise boundaries for measurement and generation of the phoneme-masked stimuli. The criteria used for segmenting and confirming the alignment for the French dataset were discussed, and the same criteria were used as were used in English.

Table 3: The French Word List

	Nasal Word	IPA	Gloss	Oral Pair	IPA	Gloss
ɛ̃	train	tʀɛ̃	train	trait	tʀɛ	trait
ɛ̃	lin	lɛ̃	linen	lait	lɛ	milk
ɛ̃	rein	ʀɛ̃	kidney	rai	ʀɛ	ray (of light)
ɛ̃	crin	kʀɛ̃	hair	craie	kʀɛ	chalk
ɛ̃	prince	pʀɛ̃s	prince	presse	pʀɛs	press
ɛ̃	plein	plɛ̃	full	plaie	plɛ	wound
ɛ̃	inde	ɛ̃d	India	aide	ɛd	help
ɛ̃	gains	gɛ̃	profits	gay	gɛ	gay man
ɛ̃	fin	fɛ̃	end	fait	fɛ	fact
ɛ̃	bain	bɛ̃	bath	baie	bɛ	bay
ɑ̃	pan	pɑ̃	section	pas	pɑ	pace
ɑ̃	temps	tɑ̃	weather	tas	tɑ	heap
ɑ̃	planque	plɑ̃k	hideout	plaque	plɑk	sheet
ɑ̃	plan	plɑ̃	plan	plat	plɑ	dish
ɑ̃	gland	glɑ̃	acorn	glas	glɑ	knell
ɑ̃	gant	gɑ̃	glove	gars	gɑ	guy
ɑ̃	chance	ʃɑ̃s	luck	chasse	ʃɑs	chase
ɑ̃	camp	kɑ̃	camp	cas	kɑ	case
ɑ̃	grande	gʀɑ̃d	big one	grade	gʀɑd	rank
ɑ̃	ban	bɑ̃	cheer	bas	bɑ	quietly
ɔ̃	ronce	ʀɔ̃s	bramble branch	rosse	ʀɔs	nag
ɔ̃	rhombe	ʀɔ̃b	a musical instrument	robe	ʀɔb	dress
ɔ̃	ponte	pɔ̃t	clutch	pote	pɔt	buddy
ɔ̃	pompe	pɔ̃p	pump	pop	pɔp	pop music
ɔ̃	onde	ɔ̃d	wave	ode	ɔd	ode
ɔ̃	once	ɔ̃s	ounce	os	ɔs	bone
ɔ̃	honte	ɔ̃t	shame	hot	ɔt	hot
ɔ̃	conque	kɔ̃k	conch shell	coq	kɔk	rooster
ɔ̃	comte	kɔ̃t	account	cote	kɔt	quotation
ɔ̃	bombes	bɔ̃b	bombs	bob	bɔb	sun-hat

3.3.2 About the French data

All participants were French speakers of French, recruited by word-of-mouth and poster. Three identified as Male, five as Female. Six of the participants were living in Boulder, Colorado at the time of recording, and two were living in Philadelphia, PA. Although French was the mother-tongue of all speakers recorded, all were at least proficient in the English language. In Table 2, the gender and age of each participant at the time of recording (in May 2014), their city of origin, and where they were recorded (Colorado or Pennsylvania) are listed.

Table 4: The Recorded French speakers

	Pseudonym	**Gender**	**Age**	**Origin**	**Rec. Location**
1	CB	F	23	Eaubonne, France	CO
2	CJ	F	26	Tours, France	CO
3	ST	F	40	Angers, France	CO
4	SV	F	40	Paris, France	CO
5	YH	M	28	Rennes, France	CO
6	JT	M	40	Angers, France	CO
7	AZ	M	42	Compiegne, France	PA
8	DD	F	45	Paris, France	PA

Ultimately, 8 speakers were recorded, resulting in a total of 955 tokens.

However, before we can begin examining this data, we must first discuss the selection of features which will be used, and how these features must be extracted from the data collected.

3.4 Defining a Feature Set

Now, we must define the features which will be evaluated in the current work, and give a more precise description of how to find and measure them. This set of features is *far* larger than the subset of features which will be evaluated using human listeners, but represents a starting point for the gradual narrowing process which will take place in the following chapters.

It's also worth noting that while we've included all features which seem, *a priori*, to have at least plausible connections with nasality, it is not presented as "complete", but rather, sufficient to cover the most probable candidates. Although additional useful features may yet be lurking in the darkest acoustical depths of the signal, those that have escaped scrutiny in 40+ years of speech research are unlikely to show the sort of strong correlation required of a useful perceptual cue. Ultimately, it is simply not prudent to test otherwise implausible features with hopes of stumbling onto the Loch Ness Monster of nasal cues.

In addition, although relative and absolute amplitude of spectral features are used extensively, for the purposes of this experiment, measures of overall amplitude are not being considered. Although there is some evidence (c.f. Delvaux et al. (2002b)) that nasal vowels are quieter than oral vowels, none of the recordings were done with calibrated microphones, and speakers were

free to move around, allowing uncontrollable between-token variation (as is also the case in day-to-day perceptual contexts). This is not to say that nasality has no effect on overall amplitude, nor that it cannot be a perceptual cue, but that we cannot reliably measure or modify it within the current experimental approach.

Finally, there is no indication, either in past research or in early pilots with this data, that F_0 has anything to do with nasality. Given that F_0 is a source feature and nasality is a filter feature, no acoustical interaction is expected, and there is no evidence that speakers use pitch as a secondary cue for nasality. So, both *a priori* and on the basis of pilot experimentation, we can exclude features which measure nothing but F_0 (or multiples thereof), such as Freq_H1, Freq_H2, and so forth.

3.4.1 Features to be tested

The below features in Table 5 will be tested and modified through the remainder of this experiment.

Some of these features have been suggested, either directly or by analogy, in the literature on nasality, as described in Section 2. For these features, the paper which most clearly suggested it is listed as the "Provenance".

Some of the features are not (to the best of the author's knowledge) directly attested in the literature on nasality, but instead, are based on extensions of existing approaches, on the author's observations over his years of research, and on other *a priori* principles. In some cases, the features are fairly straightforward repackagings of other approaches (The ZeroDiff family, for instance, is heavily based on Chen's A1-P0). In other cases (such as spectral center-of-gravity, P0Prominence, or A3-P0), the features are untested with regards to nasality in the literature, but seem to merit some evaluation.

One feature, A3-P0, was actually introduced considerably later in the project, only once the author began to notice the strong spectral tilt effects of nasality in French. Although it's clearly derivative of the A1-_ features discussed in Chen (1997) and Schwartz (1968), and will often correspond to the A3-H1 measure occasionally used for spectral tilt in the voice quality literature, the specific contrast with the nasal peak and application of this tilt measure to nasality is, to the best of the author's knowledge, novel, and shows considerable promise as a feature of nasality.

Each feature in Table 5 is given a short name, for use throughout the remainder of the paper, a technical description describing its measurement, and either its provenance in the literature or, for novel features, its reasoning. For convenience, sets of relative features which are identical except for the compared value (like P0Prominence and P1Prominence) are described only once.

3.5 On Feature Extraction

As mentioned previously in Section 3, all words in our two corpora (French and English) have been annotated with Praat TextGrid files, which show the extent of the word on one tier, as well as hand-corrected boundaries of the nasal vowel on another. This will provide the necessary

Table 5: Features of Nasality for evaluation

Name	**Description**	**Provenance**
Duration	Duration of the Vowel	Stevens et al. (1987a)
Amp_F1	Amplitude of the First Formant (”A1”)	Chen (1997)
Freq_F1	Frequency of the First Formant	Delvaux et al. (2002b)
Width_F1	Bandwidth of the First Formant	Hawkins and Stevens (1985b)
Amp_F2	Amplitude of the Second Formant (”A2”)	Delvaux et al. (2002b)
Freq_F2	Frequency thereof the Second Formant	Delvaux et al. (2002b)
Width_F2	Bandwidth of the Second Formant	Pruthi and Espy-Wilson (2004)
Amp_F3	Amplitude of the Third Formant (”A3”)	Delvaux et al. (2002b)
Freq_F3	Frequency of the Third Formant	Delvaux et al. (2002b)
Width_F3	Bandwidth of the Third Formant	Pruthi and Espy-Wilson (2004)
Amp_P0	Amplitude of the higher of H1 or H2	Chen (1997)
Amp_P1	Amplitude of the highest peak near 950 Hz	Chen (1995)
Amp_P2	Amplitude of the highest peak near 1250 Hz	Schwartz (1968)
A1P0_HighPeak	Amp_F1 - Amp_P0	Chen (1997)
A1P1	Amp_F1 - Amp_P1	Chen (1995)
A1P0_Compensated	A1P0_HighPeak using Chen’s correction function	Chen (1997)
A1P1_Compensated	A1P1 using Chen’s correction function	Chen (1995)
A1P2	Amp_F1 - Amp_P2	Analogy from Chen (1997)
A3P0	Amp_F3 - Amp_P0	Captures spectral tilt concretely
H1MinusH2	The height of H1 minus the height of H2	Simpson (2012)
(P0/P1)Prominence	Difference between P0 or P1 and the average of the two adjacent harmonics	Shows the local prominence of P0 and P1.
LowZeroDiff(P0/A1)	Difference of amplitude between P0 or A1 and the frequencies between P0 and P1	Detects a wider zero pattern than F1 alone
MidZeroDiff(P1/A1)	Difference of amplitude between P1 or A1 and the frequencies between P1 and P2	Detects a nasal zero higher in frequency than F1
Ratio_F1F2	Freq_F1/Freq_F2	Captures normalized formant interactions
Ratio_F2F3	Freq_F2/Freq_F3	Captures normalized formant interactions
SpectralCOG	Spectral Center of Gravity of the vowel	Captures damping of higher frequencies, alluded to in Macmillin et al. (1999)

temporal information as input to the measurement script, which will then be used to actually extract the features.

Measurements are taken at two points per vowel, at the 1/3 and 2/3 duration points (avoiding the rather fraught vowel boundaries altogether), Information about each individual feature for each vowel is here extracted using the Praat (Boersma and Weenink (2012)) Phonetics software suite along with the CU Nasality Automeasure script. All points which were "flagged" as problematic or likely erroneous were excluded from the final dataset.

All feature information was then extracted to a tab-delineated text file, one feature per column, and then expanded by lookup to include information on the vowel, phonological structure, surrounding consonant place of articulation, as well as additional per-word information ("Is this a real word? Does it contain a diphthong?"). This file (and formatting variants thereof) will serve as the dataset for Experiments 1-3.

3.5.1 About the feature extraction script

This script is the end result of around six years of work in automation and error checking by the author, itself based on a nasality measurement script originally developed by Dr. Scarborough and Sarah Johnstone.

This script takes as input a series of sound files and textgrid annotation files, then opens each file and takes a series of measurements at a fixed number of timepoints in each vowel. These measures range from vowel duration and amplitude to a variety of spectral measures, and the output of this script (or previous versions of it) has already been used in several published papers and posters, at the University of Colorado and elsewhere.

Without discussing the process used by the script in too great of detail, the script works as follows:

1. Ask the user for a number of inputs, specifying the number of timepoints at which a measurement should be taken, the spectral regions to search for certain features, as well as the location and labeling schema for the files.
2. Isolate the nasal vowel from the remainder of the word.
3. Calculate the absolute time during the vowel corresponding to each one of the N timepoints for measurement, as well as vowel duration.
4. Generate a formant object (using the Linear Predictive Coding (LPC) built into Praat) for the sound and extract the formant measures (frequency and bandwidth) at each timepoint.
5. Extract one glottal pulse at the time of each timepoint.
6. Iterate that glottal pulse to create a large enough window to be able to do frequency analysis only on that part of the signal which is at the timepoint under observation[8].

[8]This "extract-and-iterate" process is optional, but quite useful. In addition to ensuring that each timepoint's measurement is not influenced by other parts of the vowel, it also hardens the measurements against aperiodicities in the nearby signal and provides a cleaner FFT for feature extraction. The script has been tested extensively with and without this process, and although the amount of nasality shown is similar with and without, the error tolerance of the script

7. Use the iterated file to create a Fast Fourier Transform (FFT), which shows variations in the signal's power by frequency.
8. Detect the spectral peaks associated with H1, H2, F1, F2, by finding the highest points in an expected range based on previously generated values.
9. Measure the amplitude of each peak, along with any other spectral features (Spectral COG, for instance)
10. Perform certain sanity checks to ensure that the script has properly detected the frequencies of the various formants, harmonics, and has not extracted an otherwise faulty pulse. If an error is detected, first attempt to fix it automatically (usually by changing a setting and re-running the analysis), or, if no fix is possible, flag that measurement for human review.
11. Move on to the next timepoint or word.

Before the start of this project, the script captured (or provided information which can be processed to capture) 24 of the features proposed in Section 3.4, and it was adapted to capture the remaining features with relative ease. The result is a largely automated feature extraction script, which takes TextGrid annotated sound files as input, and outputs a single, tab-delineated text file with one series of measurements (Freq_F1, Amp_F1, Width_F1, etc.) for each of N timepoints in each measured vowel in each word.

3.5.2 Notes on the extraction of some specific features

Extracting most of the features is fairly straightforward, and will proceed in the manner described in the "Description" column of Table 5. However, a few merit further discussion.

Formant Amplitude Measures (e.g. Amp_F1, Amp_F2, Amp_F3) are determined by first finding the frequency of each formant by LPC, then creating a search space which is defined as [FormantFreq $\pm$ F_0]. Within this search space, the peak greatest in amplitude is found, and the amplitude and frequency of that peak is used as the formant's amplitude and frequency.

Amplitude and Frequency for Individual Harmonics (e.g. H1, H2, H3, all used in composite measures) are found in a similar way. First, the harmonic's expected position is calculated using F_0 as obtained from Praat's pitch tracker, using the formula [ExpectedHarmonicFreq = HarmonicNumber * F_0]. Then, as with formants, a search space is created [ExpectedHarmonicFreq $\pm$ F_0], and the amplitude of the highest peak within the search space is the harmonic's amplitude.

A1-P0 is calculated by finding Amp_F1 (as above) for A1, then finding H1 and H2 (as above). The harmonic with greatest amplitude among H1 and H2 is considered to be "P0". This is in line with Chen (1997), and is the approach taken in all papers examined which specified this detail.

Note that this means that in some cases, the harmonic designated as P0 may vary unpredictably within an individual speaker, especially for oral vowels. This is not problematic, as P0 is defined as a marked gain in a low harmonic in nasal vowels. In oral vowels, there is no nasal resonance, and thus, no true P0, and H1 and H2 will show normal variation, against which the specific nasal variation must be compared. Compensation is applied per the formula in Chen (1997)

is significantly improved by its presence.

A1-P1 is calculated by finding Amp_F1 (as above), then finding the highest amplitude harmonic in the range 850-1050 Hz. That harmonic is considered to be P1.

A1-P2 is calculated by finding Amp_F1 (as above), then finding the highest amplitude harmonic in the range 1150-1350 Hz. That harmonic is considered to be P2.

P0- and P1Prominence is calculated by measuring the amplitude of the harmonics adjacent to P0 or P1 on either side, and then subtracting the average of their amplitudes from that of P0 or P1. So, if P1 is H9 for a given timepoint:

```
P1Prominence = [H9Amplitude] - ([H8Amplitude] + [H10Amplitude] / 2)
```

The same formula is used for P0Prominence if P0 $==$ H2. If P0 $==$ H1 (and thus, there is no lower harmonic for comparison):

```
P0Prominence = [Amp_H1] - [Amp_H2]
```

Spectral Center of Gravity is calculated across the entire bandwidth of the sound (22,050 Hz), in order to give preferential damping and absorption of higher frequencies the best chance possible of affecting the measure.

3.5.3 Automated Feature Extraction: Advantages and disadvantages

All feature extraction for this project was done automatically. This approach has practical advantages and disadvantages, and these merit some discussion here.

The chief advantage of automatic processing of data is speed. For the A1-P0 measurement alone (which is one of 30+ features needing measurement at each timepoint in order to calculate all 29 tested features), completely manual measurement of two timepoints in a word takes around 1 minute per word, including processing the file in preparation and logging the results. Computer assisted A1-P0 measurement (where the computer does the processing and attempts a measurement and then presents it to the human for verification) takes 2-4 seconds per-measurement-per-timepoint, and around 15 seconds per measure when a timepoint requires verification (around 7-10% of the time).

Assuming 30 original measurements must be made for each of two timepoints, this works out to 229,380 individual measurements for the 3,823 tokens in the English corpus. Figuring conservatively, manual measurement of this data would take around 1,900 hours of constant measurement, which could be reduced to a mere 360 hours of measurement through computer assistance. Fully automatic measurement of all 29 features (and those measures needed to derive them) for two timepoints per word *for every word the entire English corpus*, using the above script, takes approximately 12 minutes. Clearly, then, with two large corpora to examine, automatic measurement is the only feasible approach given the time frame of this project.

The primary concern with automatic measurement is the fact that human researcher, unlike a computer, should intuitively understand that F_0 of a listener's voice is unlikely to be 1500 Hz, and that an FFT spectrum which shows no peaks will not yield reliable peak measurements. Because of this, when no human is involved with each measurement, careful programming and sanity

checking is needed to ensure that bad data is not analyzed as good. This issue is addressed by the present approach in three ways.

First, in the measurement script itself, a series of sanity checks are performed to detect common failure modes. For instance, the captured value for F_0 is checked against user-specified high and low values, and, if F_0 is above 300Hz, the measurement is attempted again with different settings. To avoid analyzing the peak-less (or sawtooth-like) spectral slices which often result from aperiodic (usually creaky) speech, the script performs another type of check, making sure that there exists a valley between the first and second harmonics, and that that valley is greater than 4 dB. If such a valley is not found (or, if any other failure mode is detected which cannot be fixed), the measurement is "flagged" as being a failed measure, and the data is branded with an error code which is available at the time of analysis. These "flagged" measurements, constituting between 7 and 10% of the data, can then be excluded from the analysis. It is worth noting that this script has been used extensively in other experiments, and its current state and error-control measures are the result of 6 years of iteration, testing, and revision.

Second, the data, throughout the correlation and machine learning stages, will be extensively reviewed. Unusual or unexpected patterns will be examined, and if needed, the script will be edited to detect and flag these errors as well. Given the relative speed of running, re-running all the data with better code or parameters which result in improved data is not particularly difficult.

Finally, the sheer volume of data collected allows significant amounts of measurement noise to be absorbed and removed. Even in the extreme case that 10% of the data is flagged, and 5% of the remaining data is erroneous in an undetectable way, in the English dataset alone, we are left with nearly 3,270 accurately measured tokens (6,540 timepoints), a respectable number by any reckoning.

So, although automated feature extraction isn't ideal and has some drawbacks relative to hand-measurement, for the purposes of the present work, there is no other feasible choice. More importantly, active steps have been taken to minimize the weaknesses of this approach, so, despite this necessity, these data should provide a solid basis for further analysis.

Now that data has been collected, annotated, measured, and coerced into the proper formats for analysis, all that remains are the analyses themselves.

4 Statistical Analysis of the Acoustical Features

We must now focus on finding and evaluating relationships, both in terms of statistical relationships and predictive power, between these nasal features and and phonologically expected nasality in actual speech data. This will be accomplished first through in-depth statistical analysis of the relationship of these features with nasality, then, in the next chapter, a series of related machine-learning studies aimed to understand the role of each feature in the nasal classification process.

In this process, we must keep in mind the goals and scope of this study. Although we are clearly interested in the production of nasality, the primary goal here is not to produce an exhaustive account of the acoustical consequences of nasality. Although we are evaluating 29 different possible acoustical features here, there are an infinite number of features and combinations that *could* be considered. In addition, other factors (such as speaker gender, age, and language) are not present in the spread we might desire for a purpose-built and comprehensive production study. Thus, although this process will yield information about the acoustical properties of nasality, it should not be considered a definitive production study.

Instead, our goal is to provide meaningful data for consideration as we attempt to find the *features which are most likely to be perceptually useful in each language*, and prepare for the subsequent listener studies. Thus, rather than focusing on *any* significant correlations, our criteria for evaluating these features need to be grounded in perception, and we must focus on those features which not just vary with nasality, but vary in consistent, perceptible, and *useful* ways. We will also want to consider these features in their greater context and in juxtaposition to each other, rather than treating them as wholly independent entities. So, given groups of interrelated features (such as A1P0, Amp_F1, and Amp_P0), the best representative feature will be chosen and discussed, rather than attempting to evaluate each component independently. There will likely be features which are significantly correlated with nasality and vary perceptibly from oral to nasal vowels, but which, due either to redundancy or other factors, do not meet the threshold for testing.

In this evaluation process, we will reduce our present 29 features to a set of 5 particularly promising features, each of which demonstrates strong statistical link with nasality, a likely-to-be-perceptible degree of change, and strong utility for classification.

These features, alongside the other 24, will then be evaluated through Machine Learning and then honed to a final set in Section 5. At this point, we will also discuss the cross-linguistic differences in these features, and any other issues relevant to the eventual goal of preparing a perceptual experiment to test their relative utility.

Before we proceed with the analysis and evaluation, we must discuss the two evaluative methods used and their relative strengths and weaknesses.

4.1 Statistical Processing vs. Machine Learning

Although both statistical studies and machine learning involve heavy statistical processing of extracted data, they require different considerations, produce different output, and are meant to

accomplish two very different things.

Our statistical evaluation will examine the whole of the data set(s), and will use a series of linear mixed-effects models to examine the correlations between each feature and the presence of level of nasality across all of the items, while controlling for random factors like speaker, word, and repetition. The end goal of this analysis is to have a list of features, their correlations with nasality (showing the strength of the relationship), and details of the effect of nasality on the feature (Does the measured value go up or down? By how much?). These statistical links can then be compared, examined, and used to better understand the empirical reality of the link between nasality and each feature.

This more conventional statistical analysis is useful in two ways. First, it's useful for replication. Most existing work on "the features of nasality" finds these by examining oral and nasal data, and finding correlations between expected nasality and the acoustics of the data. These statistical data are useful, then, as they provide more straightforward points of comparison to prior work.

Second, it's useful for capturing nuance in the data. Machine learning is a practical process, working vowel by vowel and designed to focus on the *best* criteria, and to accomplish the task. These statistical analyses are able to look at the entire dataset in the aggregate, and to find those patterns which are too small to emerge token-by-token, and perhaps not strongly useful for classification, but which still have a measurable link to nasality.

The machine learning studies in Section 5 will also look at the data in an effort to find statistical patterns, but from a different perspective. Here, these patterns will be used for the purpose of *classification*, that is, feature-based prediction of nasality *in each individual vowel token*. This classification can be conducted using the entire feature set, subgroups of those features, or even individual features, and in each case, we are primarily concerned with *classification accuracy*, that is, "How effectively does this feature (or set) allow us to predict whether a token is oral or nasal?". The end result will not be simple correlations, but something more akin to perceptual utility, an understanding of which features and groups are *predictively useful* in classifying a given specific vowel as nasal or oral.

So, statistical studies and machine learning effectively provide two different perspectives on the same questions, and their different strengths combined allow a more nuanced and useful analysis than either experiment might alone.

4.2 Methods: Statistical Analysis (Experiment 1)

In this experiment, we will examine the statistical relationship of each individual feature described above with the expected relative nasality of each vowel, as determined by the structure of the word. For each analysis, we will use both speaker and vowel as random effects, and data from each language will be considered separately.

Using speaker and vowel as random effects will both harden against (and test) the idea that the acoustical correlates of nasality differ across different speakers' productions, and comparison between the English and French results should show whether our English and French speakers are performing nasality differently in our different languages.

The goal of this initial experiment is to gain some understanding of the link between each of these various features and nasality, to evaluate the possibility of cross-speaker and cross-language variation, and to use our understanding to remove from further consideration features which are not meaningfully associated with nasality as it occurs in language.

4.2.1 Data Collection

The process of collecting and measuring features in the French and English data collection is described extensively in Section 3. The resulting output is then read into the R Statistics Suite (R Core Team (2013)), which is used for statistical analysis, deterministic generation of the included tables (using the `stargazer` package, Hlavac (2014)), and then eventual machine learning experiments.

4.2.2 The Structure of the Comparison

The wordlist recorded was purposefully designed to allow straightforward comparisons of nasal and oral vowels. It consists entirely of pairs of words which differ (at a phonological level) only with respect to nasality. The contrast between "bed" and "men" is the [±NASAL] status of the surrounding consonants (and thus, of the vowel). Similarly, in French, the sole expected difference between 'fin' (/fɛ̃/) and 'fait' (/fɛ/) is the phonological nasality of the vowel.

Because these minimal pairs differ only by nasality, when holding the speaker constant, we can safely expect that acoustical differences in the vowel across these pairs (when borne out over large amounts of data) stem from nasality. Thus, the analysis is simple: Observe the 36 features being considered at various points in each pair, then compare the difference in the feature's measured value between oral and nasal vowels. This comparison, henceforth "ΔFeature", is:

ΔFeature = [Value of Feature | "Oral"] - [Value of Feature | "Nasal"]

Our null hypothesis states that each feature does not differ meaningfully between oral and nasal vowels (ΔFeature ≈ 0), and we can reject the null hypothesis if we can show that ΔFeature differs statistically from 0 (as any correlation, positive or negative, suggests that the feature is linked in some way to nasality).

In order to conduct this comparison, though, we must define "oral" and "nasal", and this definition must differ for our two languages.

In our French data, "oral" and "nasal" correspond to phonological categories. In French, vowels like that in 'fin' (/fɛ̃/) are "nasal", and those as in 'fait' (/fɛ/) are "oral". The link between articulatory nasality and phonological nasality has been shown repeatedly in both airflow and acoustics (see Section 2.2.2). It is true that the degree of nasality in nasal vowels will vary, both word-to-word and moment-to-moment, and there will always be aberrant articulations, but it is a very safe bet that nasal vowels, at any given point, will exhibit greater articulatory nasality than oral vowels. Thus, we can safely make the assumption that comparing any given point in a phonologically oral vowel to any given point in a phonologically nasal vowel *is* comparing an oral articulation to a nasal one.

So, in French, for each feature, we will compare the measured values between the CVC and the CṼC, yielding three ΔFeature measures per oral/nasal comparison (one per timepoint), and all French data will be coded for "Nasality", where "0" means "oral vowel" and "1" means "nasal vowel".

In English, vowel nasality stems from the surrounding nasal consonants, and thus, our "oral" category will be the CVC words. Here, both surrounding consonants are oral, and thus, we have no reason to expect any nasality at all (putting aside speaker errors, pathologies, or other forms of articulatory creativity).

The "nasal" category is a bit more complex. Clearly, any vowel with surrounding nasal context will have *some* nasal influence. As shown by airflow (c.f. Cohn (1990)), there is an some degree of nasal airflow throughout vowels in CVN, NVC and NVN contexts (whereas there is none in CVC words) . The most consistent nasality at all points in the vowel will come from NVN vowels, where there is no articulatory reason to raise the velum at any point during the vowel. In NVN words, all points should be nasalized, and the degree of nasality should be most constant. In CVN and NVC contexts, the degree of nasality at any given point in the vowel is much more variable (albeit predictably so). However, in most cases, any given point in a CVN/NVC vowel should be more nasal than the equivalent point in an oral (CVC) vowel.

The most conservative position, then, would be to only use vowels in NVN words as the "Nasal" vowels, and to compare them to the CVC "oral" vowels. However, because we still expect *some* difference in nasality between CVC and CVN/NVC vowels at most points (even if it's not as strong at all points), there is little harm in classifying CVN, NVC *and* NVN vowels as "nasal", especially given that doing so will triple the number of comparisons possible.

So, in English, we will compare the measured values for each feature between oral (CVC) vowels, coded "0" for Nasality, and nasal (CVN, NVC, NVN) vowels, coded with a "1" for nasality, at each timepoint.

Because we are now making the same comparisons in English and French ("oral" vs. "nasal"), the data can be arranged as a series of oral/nasal feature comparisons in which differences in each feature can be tested for significance *en masse*, and cross-linguistic comparison is relatively straightforward.

Using this coded data, our goals here are to find ΔFeature in our above-defined "oral" vs. "nasal" vowels, to characterize that difference, and to examine the interactions or additional factors relevant in determining the final difference.

To accomplish these goals in complex speech data, though, will require a slightly more nuanced approach to analysis.

4.2.3 Using Linear Mixed Effects Modeling

Although "how does this feature change when nasality changes" is very much the domain of classic linear regression, it is prudent to also consider the other sources of variability in these features in this data.

For example, the vowel quality, timepoint, and repetition are all likely to affect the values of each feature. These are what are conventionally referred to as "fixed effects", as they are controlled by the experiment, and the range of options (within the experimental domain) are all present and measured. Put differently, these fixed effects are fully represented in the data, and by evaluating both repetitions, the entire range of influence of that effect can be seen[9].

These can be compared to our two "random effects", word and speaker. There exist (rather large) groups of both speakers and words in English and French which were *not* recorded for this experiment, and thus, the full range of neither is represented. These two sources of variability will introduce noise that cannot be built into a paradigm of possible effects, and practically speaking, is random.

Speaker has an additional complication. Speakers may have not just a different *degree* of nasality for a given word, but also a different *amount of change* in nasality from oral to nasal. This means that the variable speaker should have not just random intercepts (representing that speakers will have different degrees of nasality), but random slopes as well (representing that speakers will show different changes in degree of nasality from oral to nasal).

We will establish the significance (or lack thereof) of ΔFeature for each feature using a Linear Mixed Effects Regression (henceforth, "LMER"), as implemented in the `lme4` package in R (Bates et al. (2014)).

To conduct each analysis as outlined above, for each language and feature, an LMER was run, taking into account each of the fixed and random effects discussed above. So, for example, to evaluate the relationship between the amplitude of A1 ('Amp_F1') and nasality in English, the below R code would be run (noting the difference in syntax between random and fixed effects, as well as the by-speaker random slope for nasality):

```
Amp_F1.lmer = lmer(Amp_F1 ~ nasality + repetition + vowel + Timepoint + (1+
    nasality|speaker) + (1|Word), data = eng)
```

This outputs coefficients for all fixed effects (one each for nasality, repetition, and timepoint, and one per vowel, expressed as "difference from /a/"), as well as the variance absorbed by word and speaker. In addition, for each fixed effect, we are given a "t-value".

According to Baayen (2008):

> A simple way of assessing significance at the 5% significance level is to check whether the absolute value of the t-statistic exceeds 2.

Thus, we will be using $|t| > 2$ as our test of significance for each fixed effect, and any feature for which the $|t|$associated with nasality is < 2 will be considered to have a non-significant correlation with nasality.

[9]This explanation, as well as the use of LMER here, owes a great deal to the straightforward and pleasant LMER and R tutorials published by Bodo Winter (Winter (2013))

4.3 Experiment 1: Criteria for Feature Evaluation

We will consider features' usefulness and statistical link to nasality in terms of four statistical criteria.

First, the feature must show a strong and statistically significant correlation with our binary approach to nasality, in either or both languages. If a feature has no meaningful correlation with nasality, needless to say, it is unlikely to have any meaningful perceptual link with nasality. This will be shown by examining the | t | values for the correlation with nasality all three datasets.

Second, there must be a meaningful difference in this feature between oral and nasal vowels. Put differently, even if the correlation is significant in some or all of the datasets, a change small enough to be lost in the inherent noise of speech will be of little use. To do this, we will compare the coefficient for nasality (ΔFeature) to the standard deviation of the feature *in oral vowels*, under the assumption that a more perceptible feature might change by a larger amount relative to the oral standard deviation (the baseline "noise").

Then, for a feature to be considered an *inherent acoustical property* of nasality, the languages must agree on the direction of the change. Although English and French speakers could change their production in subtle ways to differently accentuate certain acoustical features, it seems highly unlikely that speakers of two different languages, when making the same gesture, would experience *directly opposite* acoustical effects[10].

Finally, each feature chosen should be independent of the others chosen. For maximal robustness, we would want each feature to focus on and measure a different aspect of the acoustics of vowel nasality (such that all of our perceptual ability is not dependent on a particular peak or spectral region). Because we have included sets of interdependent measures which all capture the same phenomena (Amp_A1, Amp_P0, A1P0_Highpeak, and A1P0_Compensated, for example), we should eliminate from further consideration those features in these sets which show the same pattern, but do so with smaller coefficients, or with increased noise relative to others. Put differently, our final feature set should reflect a variety of acoustical properties of nasality, and use only the most effective feature for each.

So, we will consider a feature to be a good candidate for further evaluation and discussion only if:

1. The feature has a significant correlation with nasality.
2. There is a ΔFeature between oral and nasal vowels which is greater than the perceptual limen and is meaningful given the normal variance within the category.
3. The feature is linked with nasality similarly in both languages, even if to different degrees, or is plausibly implemented separately from the nasal gesture.
4. The feature is the best feature for capturing a unique and independent aspect of the acoustics of nasality

[10]There are some features which do not arise directly from the nasal gesture (e.g. duration, F_0, or formant changes reflecting changes in tongue positioning) which could be *associated with*, rather than *caused by* nasality. These secondary phenomena may improve the salience of the nasal contrast (c.f. Carignan (2014); Carignan et al. (2011); Shosted et al. (2012)), or may simply co-occur, but regardless, they most certainly can be implemented in language-specific ways.

4.4 Experiment 1: Statistical Study Results

Our principal goals in conducting this experiment are to evaluate the links between each feature and nasality, to narrow the field of features by eliminating those irrelevant to nasality, and to evaluate the variability caused by different speakers, vowels, timepoints and repetitions.

To this end, we will examine the output of our linear mixed-effects regressions, examining the relevant statistics needed to fully understand the trends in the data.

Running 29 independent LMER analyses generates a considerable amount of raw data, far more than should be presented to an innocent reader. Rather than including a 30 page appendix of R readouts, or worse still, including each analysis inline, we will examine aggregated data for all features in terms of their link with nasality in our dataset.

We will then examine the role of speaker and vowel for some of the more promising features, and then dive into the groups of related features to identify the best candidates for future consideration, which will be evaluated as a group in our machine learning analysis, and then discussed in detail, as well as in cross-linguistic contrast, in Section 5.8.4.

Below, in Tables 6 through 11, are the aggregate results of LMER testing for each feature, for each language and across both. For the sake of clarity, we will show in separate tables (7 and 11) those features which do not meet our $|t| > 2$ threshold for significance.

Each row of the below tables gives data for one feature, listing a variety of figures, each explained below:

- **Feature**: The shortname of the feature, as listed in Table 5.
- **Nas.Coef** ("Nasality Coefficient"): The ΔFeature which the model attributes to nasality. Put more practically, the expected change in the feature when an oral vowel is nasalized.
- **Nas.t** ("Nasality t-value"): The t value produced by the model for the correlation with nasality. Again, when $|t| > 2$, the effect is considered to be significant.
- **OralMean**: The mean value of the feature in vowels labeled "oral"
- **NasalMean**: The mean value of the feature in vowels labeled "nasal".
- **CoefvsSD**: The nasality coefficient divided by the standard deviation for the feature *in oral vowels only*, to gauge the strength of the change relative to expected noise.

For ease of display and interpretation, all features where $|t| < 2$ have been placed in a separate "non-significant" table, and the will be discussed separately for English and French. Both tables have been sorted by CoefvsSD.

4.4.1 English Statistical Results

The results of the English statistical study are presented in Tables 6 and 7. 19 of our 29 features reached the threshold for significance in the English dataset. Several expected patterns from the literature are reassuringly present, and rather strongly so: Amp_F1, A1-P0, A1-P1, and A1-P2 all

fall with nasality, F1's bandwidth grows as the formant falls in frequency, and H1MinusH2 rises, as expected based on Simpson (2012).

Table 6: Correlation with Nasality in English (Significant Features)

Features	Nas.Coef	Nas.T	OralMean	NasalMean	CoefvsSD
A1P0_Compensated	-4.114	-7.584	5.099	1.133	0.570
A1P0_HighPeak	-4.065	-7.471	2.932	-0.982	0.562
Width_F1	96.091	6.579	171.093	262.532	0.479
LowZeroDiffA1	-1.917	-7.961	-20.762	-22.643	0.469
Amp_F1	-2.533	-7.186	46.445	43.962	0.376
LowZeroDiffP0	2.155	5.239	-23.695	-21.660	0.362
A3P0	-3.580	-6.052	-14.965	-18.179	0.346
A1P1_Compensated	-3.550	-5.793	19.398	16.137	0.327
A1P2	-3.141	-4.307	17.050	14.422	0.266
P0Prominence	1.868	4.195	10.290	12.083	0.263
A1P1	-3.180	-4.807	15.653	12.801	0.253
Amp_P0	1.536	3.850	43.513	44.944	0.249
Duration	-17.553	-2.280	217.984	203.653	0.229
Amp_F3	-2.059	-4.042	28.547	26.765	0.211
MidZeroDiffP1	2.903	3.942	-33.866	-31.224	0.186
Width_F3	98.559	3.526	498.014	589.278	0.139
Freq_F1	32.249	3.129	601.869	629.036	0.137
Ratio_F1F2	0.024	2.521	0.377	0.394	0.124
SpectralCOG	-40.797	-2.351	758.323	717.497	0.119

Table 7: Correlation with Nasality in English (Non-Significant)

Features	Nas.Coef	Nas.T	OralMean	NasalMean	CoefvsSD
H1MinusH2	1.131	1.618	3.217	4.341	0.103
P1Prominence	0.781	1.831	1.458	2.043	0.096
Ratio_F2F3	-0.012	-1.215	0.641	0.634	0.077
Freq_F2	-25.176	-0.876	1,770.088	1,762.837	0.052
Amp_P1	0.651	0.935	30.792	31.161	0.046
Amp_P2	0.591	0.839	29.395	29.540	0.043
Freq_F3	8.170	0.417	2,756.662	2,775.772	0.026
Amp_F2	-0.280	-0.567	35.768	35.259	0.025
MidZeroDiffA1	-0.217	-0.779	-18.212	-18.423	0.025
Width_F2	14.763	0.414	421.553	442.663	0.019

We also see that a few long-shot features did not reach significance. The frequency of F2 and F3 (as well as their ratio) were not significantly correlated with nasality, and amplitude alone wasn't correlated with nasality for P1 and P2 (meaning that these peaks do not reliably change in any absolute sense).

Duration proves to be correlated with nasality, with nasalized words shorter overall. This is not

particularly surprising, given duration's tendency to act as an enhancing cue for other speech features.

Some novel features performed well as well. P0 Prominence shows a healthy ΔFeature, although P1Prominence fails to reach significance. Finally, the unattested LowZeroDiff(A1/P0) features show a strong effect, although this is not as shocking, given that these measure the same zero used by the A1-PX family. In addition, A3-P0 emerges as a strong feature for nasality, besting the attested A1-P1 family in every way.

The CoefvsSD data provides some interesting perspective. Meant to show the salience of the ΔFeature relative to normal oral variation, this column compares the size of the coefficient to the size of one standard deviation of the feature in oral vowels. Although none of these features show an oral-nasal change greater than one standard deviation for the oral vowel, we can see that, for instance, a 95+ Hz bandwidth change is far more unusual (in terms of the variability found in oral vowels) for F1 than for F3, and that A1-P0-related measures are particularly strong relative to oral vowel-to-vowel variation. These relative changes will be of considerable importance when choosing the most promising features.

It is also worth mentioning that the same overall patterns held when CVCs were compared *only* to NVNs (vs CVN/NVC/NVN above), as shown in Tables 8 and 9.

Table 8: Correlation with Nasality in English CVC/NVNs (Significant Features)

Features	Nas.Coef	Nas.T	OralMean	NasalMean	CoefvsSD
A1P0_Compensated	-5.425	-7.803	5.099	-0.162	0.752
A1P0_HighPeak	-5.370	-7.689	2.932	-2.271	0.743
Width_F1	128.889	7.609	171.093	296.747	0.642
LowZeroDiffA1	-2.561	-6.985	-20.762	-23.329	0.627
A3P0	-5.335	-7.012	-14.965	-19.629	0.516
Amp_F1	-3.438	-7.421	46.445	42.904	0.510
LowZeroDiffP0	2.824	5.378	-23.695	-21.058	0.474
Duration	-36.022	-3.654	217.984	182.211	0.469
P0Prominence	2.982	5.926	10.290	13.176	0.419
A1P1_Compensated	-4.300	-5.156	19.398	15.420	0.396
Amp_F3	-3.400	-5.018	28.547	25.546	0.349
A1P2	-3.821	-4.061	17.050	14.039	0.324
Amp_P0	1.937	3.888	43.513	45.176	0.314
A1P1	-3.835	-4.217	15.653	12.207	0.306
SpectralCOG	-76.783	-3.134	758.323	684.064	0.223
MidZeroDiffP1	3.473	3.321	-33.866	-30.729	0.223
Width_F3	131.739	3.686	498.014	615.589	0.185
P1Prominence	1.502	2.562	1.458	2.826	0.185
Ratio_F1F2	0.031	2.395	0.377	0.397	0.162
Freq_F1	36.371	2.339	601.869	632.548	0.154

Here, we can see that all features found significant above are still significant in a CVC/NVN comparison, although the coefficients are universally higher (representing the greater and more

Table 9: Correlation with Nasality in English CVC/NVNs (Non-Significant)

Features	Nas.Coef	Nas.T	OralMean	NasalMean	CoefvsSD
Ratio_F2F3	-0.020	-1.429	0.641	0.629	0.134
Freq_F2	-42.262	—-	1,770.088	1,762.474	0.088
H1MinusH2	0.671	0.679	3.217	3.651	0.061
Amp_F2	-0.603	-0.971	35.768	34.595	0.053
Freq_F3	16.387	0.619	2,756.662	2,795.524	0.052
MidZeroDiffA1	-0.323	-0.804	-18.212	-18.522	0.037
Amp_P1	0.396	0.410	30.792	30.697	0.028
Amp_P2	0.366	0.379	29.395	28.865	0.027
Width_F2	-0.219	-0.005	421.553	438.916	0.0003

consistent degree of nasality in NVNs).

4.4.2 French Statistical Results

The results of the French statistical study are presented in Tables 10 and 11, using the same columns and data format. In French, 19 of our 29 features reached the threshold for significance.

Once again, the results do mirror some expectations, with A1-P0 showing strong and salient ΔFeature between V and Ṽ words, alongside the damping and widening of F1 and the gain in H1MinusH2 predicted in Simpson (2012). Duration also proves to be a very strong feature of nasality in French, with nasal vowels around 30ms longer than oral vowels (compared to English, where nasals were around 17ms shorter).

Interestingly, although they both showed reasonable changes in English, neither A1-P1 (in all forms) nor A1-P2 is significant in French. The cross-linguistic differences do not end there, and in fact, the only feature which did not reach significance in either English or French is the (rather improbable) Freq_F3. These differences will provide ample fodder for discussion later in Section 4.8.

Also rather surprising is the marginal significance (t = -1.971) of the drop in F2's frequency reported in Delvaux et al. (2002a). 180 Hz is a substantial formant frequency drop, and without using speaker as a random-sloped factor, the drop is present and strongly significant (at roughly the same magnitude as in models including random slopes for speaker, see Table 15). No other formant frequency effects are found here, which is understandable given the random effect for vowel (discussed more thoroughly in 4.5.6).

One particularly striking set of results involves amplitude measurements: The amplitude of P1, P2, F1, F2 and F3 all drop strongly in nasal vowels, and with high CoefvsSD, indicating a strong divergence from the oral norm. The amount of drop appears to be closely linked to frequency, with F3 dropping most, then F2, then P2, and so forth, indicating that this is not a universal drop in amplitude ("nasal vowels are quieter"), but instead, represents spectral tilt ("harmonics in nasal

Table 10: Correlation with Nasality in French (Significant Features)

Features	Nas.Coef	Nas.T	OralMean	NasalMean	CoefvsSD
A3P0	-10.288	-11.663	-12.537	-22.816	1.161
A1P0_HighPeak	-5.665	-6.713	6.500	0.719	1.105
A1P0_Compensated	-5.633	-6.471	8.681	2.932	1.100
Amp_F3	-11.291	-8.344	26.579	15.154	1.013
LowZeroDiffP0	4.253	4.975	-26.962	-22.641	0.868
Duration	29.994	3.493	107.315	137.206	0.802
Amp_F2	-7.475	-4.416	35.843	28.181	0.745
Amp_F1	-6.673	-7.166	45.616	38.690	0.666
Width_F3	359.829	4.604	400.098	760.728	0.653
Ratio_F2F3	-0.072	-3.070	0.594	0.523	0.556
SpectralCOG	-143.244	-7.189	779.721	636.496	0.545
Amp_P2	-6.638	-3.164	28.237	21.486	0.540
H1MinusH2	3.137	3.009	-2.962	0.170	0.457
Width_F1	157.591	3.602	205.784	365.041	0.433
P0Prominence	2.676	2.490	2.997	5.645	0.425
LowZeroDiffA1	-1.436	-2.614	-20.461	-21.922	0.404
Width_F2	229.980	2.497	374.930	605.643	0.377
Amp_P1	-4.643	-2.699	33.322	28.478	0.354
MidZeroDiffA1	-1.218	-2.263	-18.046	-19.270	0.082

Table 11: Correlation with Nasality in French (Non-Significant)

Features	Nas.Coef	Nas.T	OralMean	NasalMean	CoefvsSD
Freq_F2	-180.646	-1.971	1,588.074	1,410.198	0.441
Ratio_F1F2	0.040	1.559	0.431	0.470	0.254
A1P1_Compensated	-1.827	-1.235	15.657	13.764	0.245
A1P1	-2.025	-1.277	12.294	10.211	0.239
Freq_F3	53.984	0.670	2,676.358	2,729.481	0.152
Freq_F1	-24.140	-0.841	637.208	613.365	0.149
P1Prominence	0.845	1.931	0.378	1.206	0.127
Amp_P0	-1.001	-1.023	39.116	37.970	0.106
MidZeroDiffP1	0.788	0.468	-30.339	-29.481	0.046
A1P2	-0.040	-0.018	17.379	17.203	0.004

vowels fall off more quickly"). This is confirmed by a sharp drop in spectral center of gravity in nasal vowels, indicating that more of the energy for nasals is found in the lower frequencies.

This spectral tilt, in turn, led to a surprising victory for a dark-horse feature, the newly-proposed "A3P0". This feature, the difference between Amp_F3 and Amp_P0, is far and away the best feature for nasality in French, besting both of its components (Amp_F3 and Amp_P0), as well as the A1-P0 family. It will be discussed further in Section 4.5.3.

Finally, it's worth noting that although CoefvsSD values are generally higher for French (indicating a greater difference in these features between oral and nasal vowels than in English), A1-P0 showed the greatest change relative to oral variation in both.

Clearly, some interesting features have emerged, several of which merit close consideration as potential perceptual cues.

4.5 Discussion: Selecting the most promising features

Although we have thus far focused on describing the acoustics of nasality in French and English, the practical goal of the present experiment is to select from our initial 29 a select few which seem most likely to be perceptual cues.

Because all but one of our features showed a significant correlation with nasality in either English or French, this will necessarily involve leaving correlated and somewhat promising features behind. Given the strength of the entire feature set, taking a reductive approach ("How can we eliminate all but five of these features?") will be very difficult, as some features must be disqualified solely because they aren't *as promising* as others. As such, rather than "throwing out" features based on performance, we'll focus on choosing the very best, and move forward with those.

4.5.1 The A1, P0, and A1-P0 Family

Clearly, the family of features which measure P0, A1, and their relative amplitudes have had an excellent showing in both languages. Below are coefficients and CoefvsSD ratios for each of the significantly correlated features affiliated with A1 and P0. – indicates values for features failing to reach significance:

Features	en.Nas.Coef	en.CoefvsSD	fr.Nas.Coef	fr.CoefvsSD
A1P0_Compensated	-4.114	0.570	-5.633	1.100
A1P0_HighPeak	-4.065	0.562	-5.665	1.105
Amp_F1	-2.533	0.376	-6.673	0.666
Amp_P0	1.536	0.249	—	—
LowZeroDiffA1	-1.917	0.469	-1.436	0.404
LowZeroDiffP0	2.155	0.362	4.253	0.868
MidZeroDiffA1	—	—	-1.218	0.082

In both languages, the strongest ΔFeature vs oral variability is shown by the A1-P0 measures (with and without Chen's compensation formulae). We see there is some variability in the ranking of the second-place measurements (English shows a stronger ΔFeature of A1 vs. the P0-to-F1 zero, whereas French shows a stronger rise of P0's prominence above that zero), but far and away, the relative measures appear to show the strongest and most consistent ΔFeature values, and showed significances in both languages. Thus, they show the most promise, at this point, as perceptual features.

Amp_F1 and Amp_P0 are particularly interesting in this data. Although A1 (Amp_F1) dropped substantially in both languages, the degree was quite different, and P0's change failed to reach significance in French. Thus, although both features are quite clearly linked to nasality, and both will both inform the manipulation methodology for A1-P0 in each language, individually, they are not particularly robust, and don't merit individual testing.

Thus, as it is the strongest feature in this array of related features, and as it is easier to manipulate than A1P0_Compensated, it appears that **A1P0 Highpeak** will move to the next round of perceptual testing.

4.5.2 Prominence features

Although absolute amplitude of P0 did not convincingly and consistently reflect nasality, prominence proves a more robust cue. As before, – indicates values for features failing to reach significance:

Features	en.Nas.Coef	en.CoefvsSD	fr.Nas.Coef	fr.CoefvsSD
P0Prominence	1.868	0.263	2.676	0.425
P1Prominence	—	—	—	—

We can see here that in both languages, the relative prominence of P0 is significantly correlated with nasality, and shows a reasonable change relative to noise in oral vowels. Neither language shows a significant prominence change for P1, perhaps due to the complications of P1 discussed below in Section 4.5.7.

Given that both languages agree on the direction of the change, both languages appear to be using P0 (at least relative to A1), and that the ΔFeatures are rather strong and relatively less variable across vowels, **P0Prominence** is another promising feature which merits perceptual evaluation.

4.5.3 Spectral Tilt

Spectral Tilt was measured in two different ways in this study, directly, using SpectralCOG, and indirectly, by taking the amplitude of features distributed throughout the spectrum. Both are shown below, sorted by coefficient in French, for reasons to be discussed later:

Features	en.Nas.Coef	en.CoefvsSD	fr.Nas.Coef	fr.CoefvsSD
SpectralCOG	-40.797	0.119	-143.244	0.545
Amp_F3	-2.059	0.211	-11.291	1.013
A3P0	-3.580	0.346	-10.288	1.161
Amp_F2	—	—	-7.475	0.745
Amp_F1	-2.533	0.376	-6.673	0.666
Amp_P2	—	—	-6.638	0.540
Amp_P1	—	—	-4.643	0.354
Amp_P0	1.536	0.249	—	—

The Spectral Center of Gravity was measured for each vowel at each timepoint, as to capture the overall distribution of energy in the spectrum. For both languages, the spectral center of gravity is lowered for nasal vowels. This indicates that nasalized vowels tend to have more power in the lower frequencies than in the higher frequencies. For English, the change in COG was small relative to the oral variability, but in French, the change was stronger.

We can also examine the amplitude of various harmonics in the spectrum, shown above. In French, the amount of amplitude drop from oral to nasal is laid out *exactly* in order of increasing frequency, that is, F1 drops less than F2, which drops less than F3. The same ordering is visible in the "Nasal peaks", where P1 (~950Hz) drops less than P2 (~1250Hz), and ΔP0 fails to reach significance. This, coupled with SpectralCOG, seems to indicate that in French, nasality is associated with a sharp increase in spectral tilt.

This pattern is not as straightforwardly shown in English, where, although the SpectralCOG finding indicates an increase in spectral tilt and F3 shows some drop in nasal vowels, the majority of amplitude changes are not significant. Although spectral tilt is suggested by these data, it doesn't appear to be as clear as in French.

Due to the strong presence of spectral tilt coupled with relatively poor measurements (in terms of correlation and ΔFeature), the author tested several other possible measures for spectral tilt in nasality, and in doing so, implemented A3-P0. Designed to capture spectral tilt in a relative way (such that it could be interpreted vowel-by-vowel) and to incorporate the prominence of the nasal peak, A3-P0 appears to be an *excellent* feature for spectral tilt, and for nasality in general.

We also see that A3-P0 is strongly correlated with nasality in both languages (even given the lesser tilt in English), with the same direction of movement. The ΔFeatures are comparable to the A1P0 family in both languages and the coefficients are quite reasonable relative to the oral standard deviation (in fact, it has the highest CoefvsSD found in French, with a oral-nasal change greater than 1 Oral StDev). Most importantly, it captures spectral tilt in a more nasality-specific way than simple SpectralCOG.

Given both the strength of its French result and the parallels in the English data, **A3P0** is clearly the most promising nasal feature for French, and nearly equals the A1-P0 measure in English, and as such, it will most certainly be tested.

4.5.4 Formant Bandwidth

Formant bandwidth, that is, the amount of the spectrum which is amplified by each of the vowel formant resonances, shows strong nasal influence in both languages:

Features	en.Nas.Coef	en.CoefvsSD	fr.Nas.Coef	fr.CoefvsSD
Width_F1	96.091	0.479	157.591	0.433
Width_F2	—	—	229.980	0.377
Width_F3	98.559	0.139	359.829	0.653

We see that in both languages, F1 and F3 gain significant width in nasal contexts, and that in French, bandwidth increases with formant frequency, showing a nearly linear increase in bandwidth through F1, F2, and F3.

We must be cautious here: This increasing bandwidth with formant frequency is very likely caused by the powerful spectral tilt in French, as a lowered formant peak will appear to have wider "shoulders" on an LPC, and thus, a far wider bandwidth. So, although some of this bandwidth increase likely does come from nasality, in French nasal vowels (with their admirable spectral tilt), it's likely that much of the seeming strength of the bandwidth change can be attributed to tilt instead, and that there is a fixed bandwidth change, on top of which tilt is acting.

That said, the presence of these bandwidth changes in both languages, coupled with the strong prediction in the literature (particularly Hawkins and Stevens (1985b)) all point to their utility for classifying and perceiving nasality. Thus, **formant bandwidth** will certainly be evaluated as a potential perceptual cue to nasality.

4.5.5 Duration

Duration is significant in both languages:

Features	en.Nas.Coef	en.CoefvsSD	fr.Nas.Coef	fr.CoefvsSD
Duration	-17.553	0.229	29.994	0.802

Because duration is articulatorily independent of nasality (no aspect of the articulation of nasality should *cause* duration changes, or vice versa), the presence of strong, directly opposing patterns in English and French is not particularly troublesome. In all likelihood, duration serves as a complimentary cue for nasality in both languages (as it does for the English Tense/Lax distinction), with the two speech communities having settled on different degrees and directions of duration change over time.

Given its known perceptual utility as a secondary cue to other contrasts in languages around the world, it is clear that **Duration** should be investigated as a perceptual cue.

4.5.6 A note on Formant Frequency Effects

Formant frequency is another feature which need not vary with nasality (as the tongue can move independently of the velum), but merits discussion given its repeated mention in the literature. Below are the formant frequency changes in French and English:

Features	en.Nas.Coef	en.CoefvsSD	fr.Nas.Coef	fr.CoefvsSD
Freq_F1	32.249	0.137	—	—
Freq_F2	—	—	—	—
Freq_F3	—	—	—	—

The relative lack of formant-related changes is not terribly surprising, given our model. There are two possible sources for formant changes associated with nasality. First, there are nasality-driven formant changes, where the articulation of a nasal vowel *necessarily* changes the formants, whether by moving energy in the spectrum, or by other complex acoustical effects. There could also be changes in oral articulations associated with nasality, where the tongue is moved to enhance (or perhaps diminish) the contrast of the nasal (as has been discussed extensively in the literature, see Section 2.3.5, specifically, Shosted et al. (2012)), Carignan et al. (2011)).

Both sources of vowel variation would be expected to have differing effects on different vowels. Oral articulation changes would by necessity look different when applied to different base formant patterns, and similarly, if nasality itself is shifting the formants about, these shifts may vary depending on the initial state of the formants. Given that we have vowel as a fixed effect in this analysis, as well as random slopes for speakers, we would anticipate and hope that formant effects would be absorbed, so to speak, by the vowel and speaker variation in the model. This argument is bolstered by the ~180 Hz drop (as predicted by Delvaux) which is hovering just below significance in the present model, and which is fully significant without random slopes.

Thus, the lack of strong formant-related ΔFeature (outside of the small shift for English) in the present model should *not* be interpreted to mean that nasality does not affect formants. Instead, we should consider this to be evidence that the formant effects of nasality are *vowel- and speaker-specific*, and that vowels must be considered in isolation when evaluating such effects (as is already the case for most other formant-related measures).

This would seem to encourage a more nuanced view of formant variation in nasality (along the lines proposed by Shosted, Carignan, and others in the papers above), wherein nasality does not inherently affect vowel formant patterns, but where speakers can (and do) modify the formants in nasal contexts and conditions. This allows secondary oral articulations free reign for compensation, contrast enhancement, or vowel quality shifts, and indicates that quality differences associated with nasal vowels are likely strategic, rather than deterministic.

So, because formant frequency is already the principal cue for vowel perception, and because vowel-specific effects are strongly suggested, we cannot simply ignore formant frequency, but any formant frequency changes will need to be implemented differently for each vowel, and likely present a cue to the overall **vowel quality**, of which nasality is a part, rather than a cue to nasality specifically.

4.5.7 A note on P1

Finally, a note on P1-related features, which, despite the literature's optimism, will *not* be further examined. Here are the English and French data for P1-related features:

Features	en.Nas.Coef	en.CoefvsSD	fr.Nas.Coef	fr.CoefvsSD
A1P1	-3.180	0.253	—	—
A1P1_Compensated	-3.550	0.327	—	—
Amp_P1	—	—	-4.643	0.354
MidZeroDiffP1	2.903	0.186	—	—
P1Prominence	—	—	—	—

For English, P1 does show some usefulness. The A1-P1 measures move in the right direction, on roughly the same magnitude as A1-P0. Also, although P1's amplitude and prominence do not change significantly, it does rise relative to the F1-to-P1 space (MidZeroDiffP1).

In French, P1 performs poorly. Neither A1-P1 measure is significant, P1's amplitude drops precipitously (as would any harmonic at that height, given the spectral tilt), and prominence is again of no significance.

We should step back and discuss the function of A1-P1. In the literature, A1-P1 is used almost exclusively for high vowels, where A1-P0 is generally held to be less effective (because of the influence of F1). If we assume that P1-related measures are simply poorly suited to low and mid vowels, this height distinction could be causing some of P1's poor performance, as French has no high phonemically nasal vowels, and even in English, the majority of our vowels (7/10) are low or mid vowels. To this end, let us examine the P1-related features in our English dataset, focusing on the difference between high and non-high vowels:

Features	HighV.Nas.Coef	HighV.Nas.T	NonHighV.Nas.Coef	NonHighV.Nas.T
A1P1	-7.044	-6.157	—	—
A1P1_Compensated	-6.729	-6.120	-2.227	-2.927
Amp_P1	4.804	4.668	—	—
MidZeroDiffP1	6.760	5.906	—	—
P1Prominence	2.368	2.480	—	—

We can see here that vowel height *strongly* affects the utility and significance of these features, and that in non-high vowels, only A1-P1 Compensated (which includes Chen's vowel formant correction formula) reaches significance for low vowels (another strong performance for Chen's compensation algorithm). Although the current stance in the literature is that A1-P1 is *mostly* useful for high vowels, these findings go one step further, seeming to indicate that A1-P1 is, in fact, useful *only* for high vowels.

Given the lack of phonemically nasal high vowels in French, in light of these findings, we can safely say that A1-P1 (and related features) seem to play *no role* in French perception of contrastive nasality. But there exists the possibility of cue trading in English, where A1-P1 features are used

for high vowels, and A1-P0 features are used for non-high vowels. To this end, let's examine the height effects on A1-P0:

Features	HighV.Nas.Coef	HighV.Nas.T	NonHighV.Nas.Coef	NonHighV.Nas.T
A1P0_Compensated	-3.676	-6.377	-4.358	-6.316
A1P0_HighPeak	-3.534	-6.027	-4.347	-6.306
Amp_F1	-2.243	-3.864	-2.698	-6.502
Amp_P0	1.308	3.666	1.642	3.529
LowZeroDiffA1	-3.004	-6.902	-1.505	-5.110
LowZeroDiffP0	—	—	2.834	5.819
MidZeroDiffA1	—	—	—	—

We can see here that, although the coefficients are higher (representing a greater ΔFeature) in low vowels[11], our A1-P0 related features show significant links to nasality in high vowels and in low vowels [12].

Compounding these issues is the simple fact that P1 is particularly difficult to capture. Although we *appear* to be finding P1 effectively for English, in French, where it is so heavily affected by spectral tilt (and may well function differently regardless), finding P1 is not straightforward[13]. This difficulty could have resulted in measurement noise hurting P1-related features' changes of reaching significance, or masking a stronger ΔFeature.

Although it's not unreasonable to consider the idea that P1-related features are *also* used for nasality perception in high vowels, based on these data, A1-P0 is more consistently useful, both across languages and in different vowel categories. These data do not serve to "disprove" A1-P1, nor to suggest it is without perceptual merit. Instead, they suggest that P1-related features play a niche role as possible supplemental cues to nasality in high vowels, but that it's not a generally applicable feature (or measurement) for vowel nasality.

Given the phonological specificity, poor performance, and the findings above, P1-related features do not merit further investigation here.

4.6 A Preliminary Feature Set

Given all of the factors discussed above, based on these data, we can set out our test set, those features showing the most consistent and meaningful oral to nasal changes in English and French. In no particular order, we should consider:

- A1P0_HighPeak
- P0Prominence

[11]The change in ΔFeature in high vs. low vowels could also be attributed to the seemingly inherent increased nasality of low vowels (c.f Delvaux et al. (2008a), Shosted et al. (2012),Delvaux et al. (2008a)

[12]It should be noted that we have excluded as "unmeasurable" those tokens where A1's harmonic *is* P0. Thus, there are some cases where A1-P0 would be unmeasurable.

[13]This mirrors personal experience, where P1 has proven elusive even to a trained eye, in both French and Lakota data

- A3P0
- Duration
- Formant Frequency (Freq_F1, Freq_F2, and Freq_F3)
- Formant Bandwidth (Width_F1, Width_F2, and Width_F3)

This feature set will be evaluated using machine learning in Section 5, and will help lead to the final feature set, which we will establish and discuss in Section 5.8.4.

4.7 Discussion: Evaluating Sources of Variability in our Features

Before moving on to machine learning, it is prudent to look more closely at how these features are functioning in the variety of conditions and contexts in which they were recorded. To do this, we'll evaluate four different potential sources of noise or variability included in our model: Timepoint, Repetition, Speaker, and Vowel.

For the sake of brevity and efficiency, we will examine only those promising features discussed in Section 4.6, above.

4.7.1 Testing the value of Timepoint and Repetition

It's worth first touching on two of our fixed effects: Timepoint and Repetition. These effects are designed to capture any change in the features stemming from position within the word and the number of times the speaker had previously repeated the words.

The immediate question is "Do Timepoint and Repetition matter at all for the production of the words?". In order to determine whether these parameters are worthwhile and these differences cause any variation, we'll use the "Likelihood Ratio Comparison" approach discussed in Baayen (2008) (pp. 253).

In this approach, we compare the goodness of fit (as represented by log likelihood) between two models, with and without the parameters in question. Put simply, we will conduct a chi-squared test (via R's ANOVA function) testing the hypothesis that the two models' goodness of fit differs. This will yield a p value, representing the probability that the two models *do not* differ in terms of goodness of fit.

Evaluating the role of timepoint for A1P0_Highpeak in English, this process looks like:

```
Amp_F1.lmer = lmer(A1P0_Highpeak ~ nasality + repetition + vowel + Timepoint +
    (1+nasality|speaker) + (1|Word), data = eng)
Amp_F1.lmer.noTP = lmer(A1P0_Highpeak ~ nasality + repetition + vowel + (1+
    nasality|speaker) + (1|Word), data = eng)
anova(Amp_F1.lmer, Amp_F1.lmer.noTP)
```

This produces, among other things, the relevant p-values. These likelihood comparison tests were conducted for both repetition and timepoint, and the resulting values are given below for each of our features, in Table 12.

Table 12: Likelihood ratio comparison p-values for models with and without Timepoint and Repetition parameters

Features	en.Repetition.p	fr.Repetition.p	en.Timepoint.p	fr.Timepoint.p
Freq_F1	—	0.002	0.00000	0.011
Freq_F2	0.043	—	0.003	0.00002
A1P0_HighPeak	—	0.00000	—	0
Duration	0.002	—	—	—
Width_F1	—	—	0.0003	0.0001
Width_F2	0.016	—	0.007	—
Freq_F3	—	—	—	—
Width_F3	—	—	—	—
A3P0	0.0001	0.001	0.0001	0

(p > 0.05 shown as —)

We can see here that for our chosen features, including repetition and timepoint significantly change the model's goodness of fit in only half of the cases (and indeed, in the full feature set, significant goodness changes are less common still). This suggests that although there may be a subtle improvement to some models, they are not profoundly useful nor consistent.

This can be further confirmed by examining the model output, looking at the t values associated with each fixed effect. Tables 13 and 14 show the t-value and coefficients for both English and French associated with Repetition ("rep") and Timepoint ("tp"), organized by feature, and only showing those features where the effect was significant.

Table 13: Features with a significant effect of Repetition in English or French

Features	en.Rep.Coef	en.Rep.T	fr.Rep.Coef	fr.Rep.T
Width_F2	30.660	2.409	—	—

(Non-signficant values as —)

Table 14: Features with a significant effect of Timepoint in English or French

Features	en.TP.Coef	en.TP.T	fr.TP.Coef	fr.TP.T
Freq_F2	—	—	34.907	4.334
Width_F1	11.132	3.629	37.005	3.998
Width_F2	21.885	2.679	—	—

(Non-signficant values as —)

We see that the models showed a significant effect for timepoint or repetition in only a very few cases, and that in all of these cases, the differences involve small (< 40 Hz) and difficult-to-interpret changes to vowel formants, which differ in degree (or significance) across our two languages.

Thus, although timepoint and repetition do appear to have some effects on these data, the effects shown are small and inconsistent. Out of an abundance of caution, they will remain in the overall models, but clearly, their role is these data is limited, and does not merit further discussion.

4.7.2 Speaker variation in nasality

In strong contrast to the Timepoint and repetition data, the role of speaker in these data is massive. As we can see in Table 15, when we perform a log likelihood ratio test (as described above), evaluating changes and goodness of fit in models with and without speaker and vowel factors, goodness of fit is significantly different for every feature, in many cases, with p values so low as to round to zero in R.

Table 15: Likelihood ratio comparison p-values for models with and without Speaker and Vowel parameters

Features	en.Speaker.p	fr.Speaker.p	en.Vowel.p	fr.Vowel.p
Freq_F1	0	0	0	0
Freq_F2	0	0	0	0
A1P0_HighPeak	0	0	0	0
Duration	0	0	0	0
Width_F1	0	0	0	0.001
Width_F2	0	0	0	0
Freq_F3	0	0	0	0
Width_F3	0	0	0	—
A3P0	0	0	0	0

(p > 0.05 shown as —)

For speaker, this is not unexpected. Nasal cavity anatomy varies from person to person, as do articulatory habits, and as such, we might expect that much like vowel formants, speakers will express the same phonemes in acoustically different ways. Although we expect speaker-by-speaker variation for formant related cues (Formant Frequency and Bandwidth), we should investigate across-speaker differences for our other promising cues.

These cross-speaker differences are particularly evident when looking at data for A1-P0. Table 16 shows the mean values for A1-P0 for all of our speakers in oral and nasal contexts, as well as the difference in means between oral and nasal vowels:

We see immediately that, across different speakers, baseline A1-P0 is vastly different, both for oral and nasal vowels. Clearly, this measure is speaker-specific, as here, one speaker's mean value for expected oral vowels may be lower than another speaker's mean value for expected nasals.

In addition, we see that there are changes in the *amount of change* in A1-P0 from person to person, indicating that different speakers express similar phonological changes in nasality with greater or lesser changes in raw A1-P0. This cross-speaker change-of-change motivates the use of random slopes for speaker, above and beyond random intercepts.

Table 16: Variation in A1P0 HighPeak by Speaker

Speaker	OralMean	NasalMean	ΔFeature
en.amelia	2.767	-1.517	-4.284
en.annie	1.972	-2.301	-4.273
en.daisy	3.047	-1.064	-4.111
en.ellie	-1.264	-3.614	-2.350
en.emily	0.955	-3.804	-4.759
en.greta	7.642	4.345	-3.297
en.hazel	6.025	2.193	-3.832
en.isabella	3.996	2.775	-1.221
en.max	-5.208	-7.350	-2.142
en.molly	4.877	-0.293	-5.170
en.olivia	4.933	-1.923	-6.857
en.sue	3.244	-2.888	-6.132
fr.AZ	6.813	-0.794	-7.608
fr.CB	10.032	3.283	-6.749
fr.CJ	4.731	1.367	-3.364
fr.DD	7.905	0.635	-7.270
fr.JT	4.883	1.115	-3.768
fr.ST	0.854	-1.697	-2.551
fr.SV	7.071	-0.690	-7.761
fr.YH	9.227	2.666	-6.562

We see a similar pattern of cross-speaker variation in both raw values and ΔFeature for P0Prominence (Table 17), A3-P0 (Table 18), and Duration (Table 19).

Table 17: Variation in P0Prominence by Speaker

Speaker	OralMean	NasalMean	ΔFeature
en.amelia	13.692	14.678	0.986
en.annie	12.036	15.646	3.610
en.daisy	10.880	12.745	1.866
en.ellie	15.550	17.437	1.887
en.emily	8.466	11.025	2.559
en.greta	6.840	7.150	0.310
en.hazel	7.537	9.712	2.175
en.isabella	10.451	10.487	0.036
en.max	7.980	8.563	0.584
en.molly	7.799	11.430	3.631
en.olivia	11.707	11.967	0.260
en.sue	11.219	14.907	3.688
fr.AZ	1.579	8.996	7.418
fr.CB	-0.426	1.104	1.529
fr.CJ	8.527	14.665	6.138
fr.DD	1.190	3.297	2.107
fr.JT	2.343	2.358	0.016
fr.ST	4.865	8.490	3.625
fr.SV	2.587	4.572	1.985
fr.YH	3.942	2.533	-1.409

We can't be sure whether this variation is due to cross-speaker differences in these particular features (regardless of nasality) or due to actual cross-speaker variation in degree of nasality, accurately captured by these features.

However, based on these data, it seems that much like with vowel formants, nasality measurement, particularly with A1-P0, is not directly comparable across speakers in raw values, nor in degree of change. Across-speaker comparison would require some form of normalization, and as many years of vowel quality research facing the same issues has shown, this process is *not* straightforward, and the development of such algorithms must be conducted with great care[14]. New methodologies may be required to do direct comparison of the degree of nasality from speaker to speaker.

Perhaps more importantly for the present work, these data provide strong evidence that cross-speaker differences will add further complexity to our eventual perceptual experiments, and any ΔFeature values will need to be modified according to the pattern shown by each speaker, rather than by some generic nasal-oral ΔFeature.

[14]It's worth noting that the vast majority of papers in the literature use A1-P0 to examine nasality in different contexts within the speech of individual speakers. There is no harm in looking at within-speaker or within-word changes in nasality, and comparing the patterns of change across speakers (e.g. "nasality increases more in this context than in that one"), but direct cross-speaker comparison of raw nasality figures would likely be so noisy as to be meaningless.

Table 18: Variation in A3P0 by Speaker

Speaker	OralMean	NasalMean	ΔFeature
en.amelia	-14.632	-17.581	-2.949
en.annie	-15.286	-17.521	-2.235
en.daisy	-17.851	-22.991	-5.140
en.ellie	-23.334	-25.578	-2.243
en.emily	-19.184	-21.320	-2.136
en.greta	-8.105	-10.199	-2.094
en.hazel	-10.558	-13.296	-2.738
en.isabella	-12.550	-14.594	-2.043
en.max	-24.462	-26.992	-2.530
en.molly	-8.427	-13.550	-5.123
en.olivia	-13.660	-18.893	-5.233
en.sue	-16.824	-21.912	-5.088
fr.AZ	-6.660	-17.568	-10.908
fr.CB	-9.055	-20.131	-11.076
fr.CJ	-16.986	-25.425	-8.440
fr.DD	-6.611	-16.210	-9.599
fr.JT	-14.778	-28.862	-14.085
fr.ST	-23.325	-33.895	-10.570
fr.SV	-9.455	-20.591	-11.135
fr.YH	-14.253	-20.788	-6.535

Table 19: Variation in Duration by Speaker

Speaker	OralMean	NasalMean	ΔFeature
en.amelia	199.057	193.631	-5.426
en.annie	174.728	161.337	-13.391
en.daisy	299.027	293.653	-5.374
en.ellie	216.780	197.451	-19.330
en.emily	227.249	199.696	-27.552
en.greta	214.817	191.894	-22.923
en.hazel	209.399	187.107	-22.292
en.isabella	217.609	212.628	-4.981
en.max	186.317	170.642	-15.675
en.molly	246.277	237.776	-8.501
en.olivia	192.596	162.191	-30.405
en.sue	249.481	236.596	-12.885
fr.AZ	117.536	134.900	17.364
fr.CB	136.576	185.586	49.010
fr.CJ	101.809	160.641	58.831
fr.DD	72.653	102.347	29.694
fr.JT	114.608	113.841	-0.767
fr.ST	82.429	114.676	32.247
fr.SV	106.959	142.034	35.075
fr.YH	125.354	144.644	19.290

4.7.3 Feature variation by vowel

We should also discuss these features in the context of different vowel qualities a bit further. The log likelihood comparison tests shown in Table 15 show quite clearly that vowel is a relevant factor in the analysis of these features, and that a model without a random effect vowel will perform rather more poorly than a model including it. We've also already discussed some high vs. low vowel differences in English with regard to A1-P0 and A1-P1 (in Section 4.5.7), stating that A1-P1 is only correlated with nasality for high vowels, while A1-P0 is correlated in all vowel types.

We would expect formant frequency and bandwidth to vary from vowel to vowel (as vowel quality is the primary determiner of formant structure). This formant variation is, then, not terribly exciting for the present work, and we can put it aside.

Given the strong role that vowel height played for A1-P0 and A1-P1, we should examine our other non-formant features for vowel height effects. In Table 20, we see the coefficients and t values for our four non-formant features in English High vs. Non-High vowels:

Table 20: Correlations and Coefficients for Nasality in English High vs. Non-High vowels

Features	HighV.Nas.Coef	HighV.Nas.T	NonHighV.Nas.Coef	NonHighV.Nas.T
A1P0_HighPeak	-3.534	-6.027	-4.347	-6.306
A3P0	-2.280	-2.419	-4.160	-5.985
Duration	—	—	—	—
P0Prominence	—	—	2.070	4.195

We immediately see that all three P0-related features perform better in low vowels (showing stronger coefficients, or for P0Prominence, significance in general). This is to be expected, as with high vowels, F1 is often close enough to P0 to influence or overlap it. P0Prominence suffers particularly as it examines the two harmonics to either side of P0, and in high vowels, the higher of the two is very likely to be F1-affected.

Duration shows no significant change when the model is run on these vowel subsets, but this is most likely due to the massive reduction in N for each model, rather than a specific vowel-by-duration interaction.

We can also compare the variability from vowel to vowel across our different features in our two languages. Table 21 shows the standard deviations of ΔFeature, as calculated from nasal and oral means, across all vowels in each dataset, as well as the standard deviations for each feature by speaker, for comparison purposes:

We see here that P0Prominence varies least from vowel to vowel, that duration varies rather consistently, and that vowels vary more in French (which is itself interesting, given the limited number of nasal vowel qualities).

Finally, we can examine the model output for the Random Factor of Vowel in Tables 22 and 23, showing the by-vowel coefficients (relative to /a/) and the t values for each vowel[15].

[15]Due to difficulties with Unicode characters in R and Python, vowels are here labeled in ASCII, where ae = /æ/, aj = /aɪ/, eh = /ɛ/, O = /ɔ/, uh = /ʌ/, ej = /eɪ/, ih = /ɪ/, and ou = /oʊ/. /i/ and /u/ carry their IPA pronunciations

Table 21: Standard Deviations for ΔFeature across all speakers and all vowels in French and English for non-formant features

Features	Speaker.SD	en.Vowel.SD	fr.Vowel.SD
A1P0_HighPeak	1.972	1.181	2.648
A3P0	3.910	1.810	2.359
Duration	26.537	28.237	24.410
P0Prominence	2.124	0.887	0.932

Table 22: Coeffecients for Vowel fixed effect in English and French for Low and Mid Vowels

Features	en.ae.coef	en.aj.coef	en.eh.coef	fr.eh.coef	fr.O.coef	en.uh.coef
A1P0_HighPeak	-2.744	-2.141	—	-2.775	-1.628	—
Duration	—	44.531	-67.638	—	—	-74.585
P0Prominence	—	—	—	0.910	—	—
A3P0	9.093	6.786	8.897	1.820	-4.714	2.833

(All coefficients relative to vowel /a/, non-signficant vowel effects as —)

Table 23: Coeffecients for Vowel fixed effect in English for High Vowels

Features	en.ej.coef	en.i.coef	en.ih.coef	en.u.coef	en.ou.coef
A1P0_HighPeak	-6.318	-7.051	-2.681	-4.620	-2.039
Duration	—	—	-95.337	—	—
P0Prominence	—	—	-1.068	—	-1.038
A3P0	7.180	—	7.702	-3.282	-2.569

(All coefficients relative to vowel /a/, non-signficant vowel effects as —)

Although the coefficients from random effects in a mixed model are very difficult to interpret, we see that the sole significant duration effects either show the increased length of the /aɪ/ diphthong, or show the known shortening of lax vowels in English (reported as far back as in Rositzke (1939)). We also see that P0Prominence shows relatively less by-vowel variability, compared to A1-P0 and A3-P0.

Most importantly, though, we see that vowel quality *does* cause significant changes in A1-P0 and A3-P0, and thus, that vowel is a factor when conducting nasality experiments. This means that cross-vowel comparisons of raw nasality measurements are simply not prudent, and that, when modeling nasality in our perception experiment, we will need to describe and reproduce ΔFeature *in each particular vowel.*

With our within-model sources of variability discussed, we should now move on to our largest source of variability in nasality in these data: the language being spoken.

4.8 Discussion: Nasality in English vs. French

Vowel nasality, no matter its function or phonological significance, is ultimately a speech gesture, and *a priori*, there is no reason that lowering the velum should result in different acoustical consequences in our two different languages, no matter the function of the nasality. Yet, based on the data collected thus far, we must acknowledge that nasality does differ in English and French. In this section, we'll examine the differences (and the similarities) explicitly.

4.8.1 Characterizing the Differences

First, we must acknowledge that although nasality appears to function differently in these languages, there are still more similarities than differences. A single set of 6 features (proposed in 4.6) appears to capture the most promising acoustical consequences of nasality in both languages, and the features (even outside of our final set) which showed statistically significant links with nasality in one language, also showed links in the other. These differences are differences of degree of change from oral to nasal for a given feature, and differences of the strength of the coefficients, rather than evidence that French and English speakers are expressing nasality by wholly different articulatory means.

That said, the presence of *any* difference in the two languages' expression of nasality is interesting, and merits careful consideration. Although many of the features showed cross-linguistic differences, we will focus our comparisons on the preliminary feature set given above. Table 24 shows coefficients in English and French for those features *which showed a statistically significant ΔFeature in both English and French* [16].

[16]It is worth noting that the author was unable to find a statistical test which could evaluate the true difference between coefficients representing the change from one condition to another as part of a complex model, without sacrificing the noise-reducing influence of the fixed and random effects. We will assume, for the purposes of discussion, that differences in coefficient are meaningful and interpretable, particularly in light of the machine learning data, but we cannot rule out the possibility that these "differences" may vary in degree (or reality) relative to the figures below

Table 24: Correlations and Coefficients for Nasality in English vs. French, sorted by magnitude of difference

Features	en.Nas.Coef	fr.Nas.Coef	EnFrDiff
A1P0_HighPeak	-4.065	-5.665	1.600
A3P0	-3.580	-10.288	6.708
Duration	-17.553	29.994	-47.547
P0Prominence	1.868	2.676	-0.808
Width_F1	96.091	157.591	-61.500
Width_F3	98.559	359.829	-261.270

First, we must note that of course, different measurement types will show different degrees of variation, partly due to different units. It should be unsurprising that formants have a larger magnitude of difference than our relative amplitude measurements, and this alone provides no useful information for our analysis.

Within our relative amplitude measures, we see immediately that P0Prominence and A1-P0 show differences between English and French, but not particularly strong ones. ΔP0Prominence is around 1dB higher in French than in English, and ΔA1-P0 is ~1.6dB higher in French than English.

We see that Duration is profoundly different in our two languages, with English nasal vowels nearly 20ms shorter than oral vowels, and French nasal vowels almost 30ms *longer*. Put differently English and French differ not just in terms of *degree* of duration, but in *direction* of duration change, and based on the per-speaker results in Table 19, this appears to be a near-universal pattern among the speakers.

Most interestingly, A3-P0 shows a much larger difference, around 6dB lower in French, implying that spectral tilt is much more strongly linked to the production of nasality in French than in English. This in turn affects formant bandwidths (as a damped formant will appear wider to an LPC analysis), causing French Width_F1 and Width_F3 to grow sharply in French relative to English.

This is, perhaps, the strongest difference between the two languages, both in terms of degree and in terms of expectation. English shows some spectral tilt differences, but nowhere near the magnitude of French, and the degree of difference is much greater than found with A1-P0 and P0Prominence, suggesting that this may go above and beyond simple variations in degree.

Finally, we should note that although we've focused on our most promising features here, many other features did show differences in both baseline and ΔFeature in our earlier analyses (particularly in Section 4.4), and, as we will see in Section 5, these differences are sufficient to affect the classification of speech.

Thus, it does appear that English and French differ not just in the phonological nature of their nasality, but in terms of the features used to express it. With this in mind, we must ask ourselves not just how two languages can differ in their realization of nasality, but *why* they might do so.

4.8.2 Potential sources of cross-linguistic variation

Cross-linguistic variation in duration and formant frequency are easily understood. Neither of these features are *necessarily* affected by nasality, and variation in formant structure or duration could easily be attributed to speaker efforts to heighten (or downplay) the oral-nasal contrast. Put differently, nasality need not change formant structure, and *cannot* affect duration, anatomically speaking, so any differences in realization can easily be attributed to speaker or language choice, conscious or otherwise.

Given that the measurements themselves were taken automatically, the data were aligned and annotated similarly, both sets were similarly balanced, and that the exact same analyses were applied to both languages, it is unlikely that any part of the data collection procedure could have varied *systematically enough* to skew the results and cause a seeming cross-linguistic contrast. Similarly, because we're looking at changes to these features ("How does spectral tilt differ in oral and nasal sounds?") rather than overall values for the features ("What is spectral tilt like in this speaker's speech?"), we can be more confident that these are not simple speaker baseline differences, but differences in the way nasality is being performed. However, other types of noise are still possible.

First, it could be that the speakers used were poorly representative samples in either or both languages, and that on both sides, oppositely-atypical groups were used. We have shown that different speakers differ in terms of both their baseline A1-P0 and their oral-to-nasal ΔA1-P0. It is conceivable that all eight French speakers showed higher-than-normal ΔA1-P0 values, and that all 12 English speakers showed lower-than-normal deltas, resulting in artificially inflated ΔFeature values. This cannot be conclusively ruled out with these data, but given that all measurements were within-speaker and all comparisons controlled for context, the risk of this being the *sole* difference seems low.

Second, this could be entirely an issue of gestural degree. If it is the case that French speakers produce articulations which result in a greater degree of nasality for nasal vowels than English speakers produce in coarticulatory contexts, then the increased ΔFeature values for every category are fairly understandable. Ken Stevens suggests this possibility in *Acoustic Phonetics* (Stevens (1998)):

> These examples provide some indication that there can be substantial differences in the spectra of nasal vowels in a language like French, where these vowels are heavily nasalized, and English, in which the nasalization that occurs in vowels before nasal consonants is apparently produced with a smaller velopharyngeal opening.

Such a hypothesis would require us to believe (and prove) that English speakers are making less complete velopharyngeal port gestures than French speakers are. This is, to the best of the author's knowledge, impossible to prove or disprove given the data at hand, as we can't separate across-speaker variation in ΔFeature from across-language differences in degree of nasality without more concrete data (airflow, acoustic nasalance, etc).

However, if the primary factor at play here is degree, we would expect that the two languages' expressions of nasality would differ only by degree, and that speakers would use the same percep-

tual cues to identify nasality, regardless of language. Although cross-linguistic perception is not feasible here, machine learning will allow us to compare the classification of both languages, to address this very question.

Finally, it's possible that although French and English speakers are making the same velar gestures (opening the VP port to similar degrees) and thus, producing "equal nasality", but that French speakers are able to, using oral articulation or other articulatory means, produce nasality in such a way to enhance some of the acoustical features we're examining here. Given that feature enhancement is a well-known phenomenon in speech, and the increasing literature discussing differences in the oral articulation of nasal vowels (c.f. Krakow et al. (1988), Delvaux et al. (2002b), Shosted et al. (2012)), Carignan et al. (2011)), the idea that French speakers are producing nasal vowels differently in a feature-enhancing way is very possible.

This final possibility, that nasality could be performed in more and less cue-enhancing ways, would provide both explanation *and motivation* for our cross-linguistic differences. In a language like French where vowel nasality carries a relatively higher functional load, the perception of nasality is relatively more important, and it's not unreasonable that speakers might learn particular means of articulation (or feature-enhancing secondary articulations) which give listeners the best chance possible to pick up a complex phenomenon. On the opposite side, in a language like English where the perception of nasality is often helpful but seldom critical, there's little reason to believe that speakers would learn to (or bother to) produce nasality in an optimally perceptible way. If this were the case, then we would expect French nasality to be objectively easier to perceive and differentiate, even by an untrained (or computerized) listener. This will be tested in Section 5.

Regardless of the source of these differences, they do appear to exist, and this raises a variety of interesting questions. We will address these cross-linguistic differences again in Section 5.9, and then in our discussion of the perceptual experiment results, once additional data is available.

4.9 Discussion: Statistical Analysis of Nasal Features

Our statistical tests have given us several important perspectives on the acoustical nature of nasality.

First, they lend further support to the validity of much previous work. We can see from these data that although A1-P0 may not be the sole acoustical feature of nasality, and it may not be the best, it *is* strongly linked with nasality even in this large dataset. Similarly, several other findings (formant frequency and bandwidth effects, A1-P1 for high vowels, damping of A1, and duration as a secondary cue) are supported here, and these reproductions, in turn, support the idea that the present study's methodology is effective for finding linked features.

We also find several new promising features, chief among them spectral tilt (as measured by A3-P0) and P0Prominence. Where many of the 29 features did not show particular links to nasality nor perceptual promise, these two did. Thus, we see the possibility of improving the state of the art, if these features prove as perceptually useful as they appear.

We also see these features as deeply variable. Differences were found here by vowel, by speaker,

and by language, for all of our promising features. This advocates caution in the use of these features in some experimental designs, particularly in cross-speaker, cross-vowel, or cross-language scenarios, and suggests that these measures are only useful for within-speaker and within-vowel comparisons of condition.

Finally, and most practically, we now have a preliminary feature set, consisting of those six features which seem *a priori* and based on these results to be the most promising potential perceptual cues. We will test these features in a perception experiment of sorts with Machine Learning in Section 5, and, if they still appear useful, carry them forward into our human perception experiment.

So, with knowledge of the overall performance of these features in hand, we will now shift approaches, and examine the utility of these features for token-by-token evaluation of vowel nasality.

5 Computational Perception: Machine Learning of Nasality

As mentioned previously, mixed effects models are excellent for finding *trends* of predictability within a dataset. Perception, though, is a different task. In perception, decisions about the nasality of a vowel are made *token-by-token*, rather than across an entire dataset. So, although both machine learning and our mixed models above rely on statistical processing, machine learning will give us greater insight into the utility of each feature for classifying *any given vowel*, rather than every vowel at once.

So, in this study, machine learning will be used here as a sort of "perceptual proxy", allowing us to test the classificational utility of more features in the signal than time, money, and complexity would allow us to do with humans. Although machine learning will not map directly to human perception, the underlying task is the same: to take in concrete information about the signal and, based on that information, classify the sound as "oral" or "nasal" (or, for English, "oral" or "nasalized").

One of the biggest advantages to this approach to studying perception is isolation from human linguistic context. We are able to evaluate the token-by-token predictive power of these features *in total isolation*, without any contextual or whole-signal information. An actual listener has a great deal of information available for speech perception, ranging from lexical information, to probability of a word's mention based on prior context in conversation, to the sorts of very fine grained, whole-word data stored in the mind. But clearly, speech perception based on acoustics must still be able to work when all other information fails. Feeding the feature data into a computer (which has no other information at all) will allow the exploration of one key question: are the acoustical features discovered above *sufficient* for nasality perception in a vacuum, or are they simply helpful on top of more important contextual and whole-signal information?

This is the fundamental question we will evaluate in Experiment 2. However, before we conduct the experiment, a word on the techniques used is prudent.

5.1 A Brief Introduction to Machine Classification

A "classifier", in the machine learning sense, is a software program designed to place a series of observations into one of several pre-specified categories. There are many types of classifier available, but all require the use of "training" data to conduct analyses which attempt to place data into difference classes, based on the combination of features (and values for those features) present in each data point.

For these analyses, we will be using two types of classifiers. For evaluating feature performance, we will be using Decision Trees, both on their own and as part of "RandomForest" ensemble classifiers. Then, to more closely approximate the state of the art in machine learning, we'll be using Support Vector Machines (SVMs). Both of these classifiers will be discussed in detail below, but for now, a word on the general nature of classification.

Regardless of classifier, the practical procedure for machine learning is the same. First, one extracts from the data the features to be used in classification (here, performed in Section 3.5). Then,

one feeds a subset of the observations (here, the measured vowel timepoints) into the classifier, each labeled according to the desired categories ("Oral" vs. "Nasal", as discussed above) and specifying the features to be used, in a process referred to as "training" the classifier. In this process, a "model" of the data is built which includes information about each feature's interaction with the categories as well as the weight of each in the eventual classification.

Finally, with a model now built, one feeds the remainder of the data (not used for training) into the classifier without labels for the categories, and asks the classifier to assign categories based on the previously-generated model. In these experiments, we will make extensive use of "10-fold cross-validation", where instead of specifying a specific subset as "training" vs. "test" data, the analysis is run 10 times on the data, each time using a different 9/10ths of the data as "training" and the remaining 1/10th as "test". The final accuracy score is based on the aggregate accuracy across these iterations.

Ultimately, then, each machine learning study will involve training models using carefully chosen subsets of features, and then evaluating the accuracy of the classification as a clue to the usefulness of the specified features.

5.1.1 Choosing Machine Learning Algorithms

Ultimately, the choice in learning algorithms is determined by the unusual nature of this task. In most machine learning tasks, the goal *is* the classification, and success is defined as accuracy on novel data, ability to handle massive datasets, and maximal computational efficiency.

For instance, a more conventional machine learning task might be to examine the readings of many different vehicle sensors, each showing variability and inconsistency, and use them to determine whether a driver is simply applying the brakes, or whether she's "panic braking", such that airbags can be readied, additional brake assistance can be applied, and data can be stored for later analysis. For this classification task, accuracy is crucial (as one doesn't want a machine-assisted "panic stop" when simply slowing for an upcoming curve, nor to have no assistance when rapidly approaching a stopped bus), the cost of a false positive is high, and it's not a problem if this accuracy comes from a "black box". For these situations, there are many conventional classifier algorithms (Perceptrons, SVMs, MaxEnt) which excel at accurate classification and allow tuning to avoid false positives, but provide relatively little (or tough-to-interpret) information about *how* the eventual classification choices were made.

However, this is not why we're using machine learning in the present work. Our primary goal is to gain information *about* the machine learning process. We want to know not only the predictive power of features and feature sets, but the predictive *utility* of each individual feature, and we'd like a firm idea of the exact criteria. This means that a black box is of little use to us. Given also that there is no need for accurate classification of novel data, the dataset is quite small (by machine learning standards), and that the efficiency of the algorithm isn't particularly important, clearly, our machine learning needs are not conventional.

Given these needs, I've elected to use a combination of two approaches. To determine the relative importances of the different features towards "perception" of nasality, we will use a series of Ran-

domForest models. Then, we'll test the individual features (and our final group) in an "enterprise grade" Machine Learning algorithm, a Support Vector Machine, to get an idea of the best-case predictive power of each feature and our groupings.

Let's briefly discuss these models now.

5.1.2 Decision Trees and Random Forests

Imagine you're your family's sole birdwatcher, and you get a text message, no picture attached, from a relative kayaking at a nearby nature preserve. "There's this beautiful bird on the lake, what kind is it?". You have been given a classification task.

As a birder in the area, you're familiar with many of the birds in the area, so you might start asking questions about individual features of the bird. You'd start with asking a broadly diagnostic question ("Is it sitting in the water, or perched someplace?"). The answer to this question would change the next question asked (e.g. if it is on the water, is it duck-sized, or swan-sized?). Through this process, and progressively more detailed questions ("Is the bill pointy or flat?", "Is it diving under the water, or just dipping its head?", "What color is the duck's eye?"), you might eventually arrive at the right answer.

This process is, effectively, a decision tree classification. One asks a series of feature-based questions whose answers lead to other questions, whose answers each lead to other questions, which, after any number of iterations, lead to a confident classification.

In machine learning, Decision tree classifiers train by taking a labeled dataset and list of features, and analyze the data to find the sequence of questions (as well as the criteria for evaluating the answers) which lead to the greatest accuracy on the training data. Then, for novel data, the questions are "asked", the answers evaluated, and depending where in the tree one ends up, the most likely classification is known.

The chief advantage of Decision trees for the present work is that they are transparent by design: not only do we know the error rate, but we also know the features which the model has found useful *and* the exact criteria used for the classification (e.g. "If Amp_P0 $>$ 4.38…"). This makes decision trees an excellent tool for our present purposes. There is one problem, though.

For a decision tree to work, we must find the most meaningful subset and ranking of features. An experienced birder knows, through years of waterfowl-based classification tasks, the important characteristics of the local lake-dwellers. We have intuitions about the most *important* features. For instance, if the bird is sitting in the water, we can significantly narrow the field, and we no longer have to ask questions which might be crucial to properly classify owls. Similarly, we might understand which features *have little classificational power*, questions like "Is it swimming in the sun or the shade?" provide little information, and questions like "Does the bird have any feathers?" or "Does the bird have two legs?" provide absolutely no classification power at all. Finally, we might understand when we're *overfitting*, that is, being led astray by coincidence: even though all the swans *on this particular lake* happen to be on the Western side, "cardinal location" is a poor predictor of swan-hood.

In extremely large datasets, Decision Trees have a good chance of finding the optimal subset and ranking of features on their own, that is, maximizing the use of important features, skipping features without power, and avoiding coincidence, simply by comparing the accuracy of different arrangements. However, the dataset used in these experiments is *not* particularly large by machine learning standards, and thus, decision tree classifiers won't reliably find the *best* set of features. Thus, we must use a different approach.

Although arriving at optimal feature sets isn't generally a problem for humans, imagine that, after you gave your relative the right answer, this relative texted the entire 500+ person membership of a local birding group asking for a classification, and gave each person a different subset of the features. One birder might be able to ask questions only about color and size. Another only about shape and color. Another might learn only about behavior and color (and so forth). Because of these feature limitations, many of them might give the wrong answer. But by evaluating which features, when given for classification, resulted in the highest accuracy of classification across all the different birders, we'd end up with an *excellent* understanding of the importance of these different features, even with a relatively small dataset. And we could then use the data obtained from this process to create a final tree which does classification with the very best features, in the very best ordering. This basic process describes RandomForest classification (albeit with fewer waterfowl).

RandomForests (a trademarked name) were originally described in Breiman (2001) and are a type of ensemble classifier, where many Decision Trees are run, but only the consensus is used as the final output. In a RandomForest model, a large number of decision trees are made (often 500-1000), each training and testing on randomized subsets of the data (to avoid overfitting and contribute entropy) and using different sets of features in different rankings. Then, once all models are run, the classifier examines all of the outputs and their accuracy for classification, evaluating which features provided the greatest boost to accuracy, as well as providing overall error rates based on the ensemble classifier's output.

This direct measure of feature utility, called "Importance", is an excellent means of evaluating feature strength (as shown by the reduction in error allowed by its presence), and this feature selection data can be fed back into a single-model Decision Tree, which will give us an exact ranking, and which will give us the all important criteria for evaluating these features.

These two approaches together allow us to perform classification in a principled, reasonably accurate, and transparent way, such that we can learn not only how well we can classify, but how we can classify well.

5.1.3 Support Vector Machines

Unfortunately, RandomForests do generally sacrifice some accuracy in classification for their interpretability. Thus, we'll also want to double-check their output using a more conventional classifier, and for this, we've chosen SVMs.

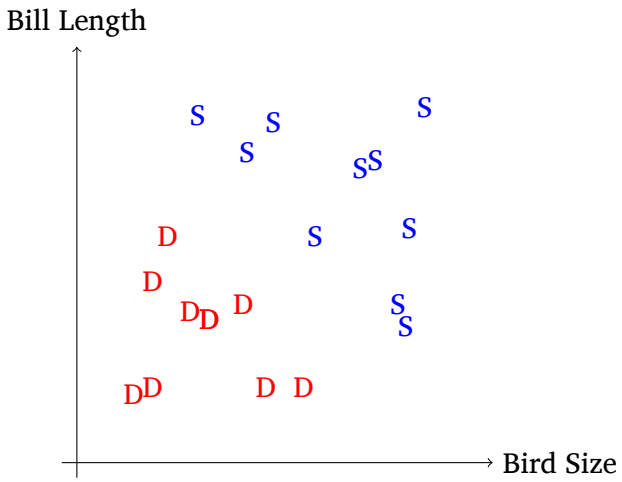
Support Vector Machine ("SVM") classifiers are a specific type of classifier which examines a field of data and attempts to find a line which best separates the different classes. Let us return to

waterfowl to explain the difference.

Your kayaking relative has taken to a quantitative approach to bird identification, and has begun snatching up innocent waterfowl from the lake and measuring the length of their bills, as well as their bill-to-tail length[17]. She sends you a list of measurements, along with pictures, and asks you to help her identify the birds as either "ducks" or "swans".

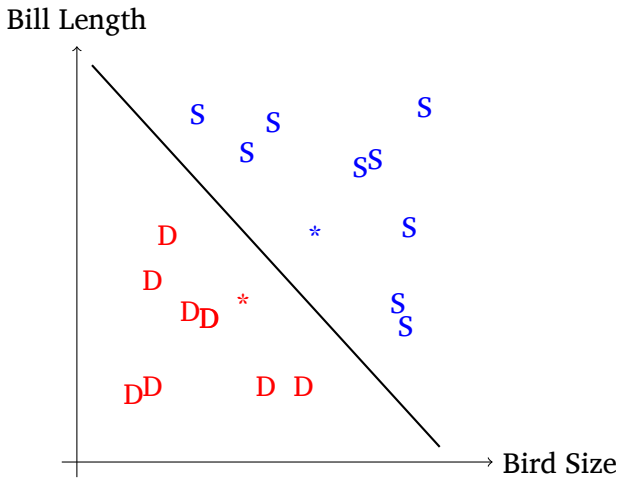
The RandomForest approach would be to first establish a criterion point for bill length, and then for bird size, and to work through the dataset, making decisions for each. "If the bill length is greater than 2 cm, and then the body length is greater than 55cm, it's a swan", and so forth. The SVM approach, on the other hand, is to create a single duck-swan line, against which all points are compared.

To do this, you plot the birds on a bill-length by bird-size plane, and then label each as either a "swan" or a "duck" based on the photo, in an initial "training" set. The resulting fowl plot is below, with D(ucks) and S(wans) labeled red and blue, respectively:



As humans, we see a single clear delineation between the two groups. The decision need not be made sequentially, feature-by-feature as in decision tree classifiers, but can be made by drawing a single line which optimally divides the two classes with the greatest distance from any point. A Support Vector Machine is just a specialized algorithm dedicated to finding the points which effectively dictate the path of the line with maximized margins, which are referred to as the "support vectors". The delineation (and the two support vectors, as *) are roughly represented below:

[17] All datapoints below are fabricated, and no waterfowl were measured or otherwise harmed in the making of this explanation.



Knowing this line, we can simply send our relative information about the line which divides ducks and swans (by identifying the support vectors). With this single piece of information, your relative can compare her latest measurements to this boundary, identifying each bird by seeing which side of the line the poor creature falls on, and duck/swan classification becomes tractable in the absence of other knowledge.

This same approach works when there are more than two features being used. Instead of a simple X-Y plane as used before, additional dimensions are added for each feature used, and a single maximized-margin hyperplane (effectively a flat surface in N dimensions) is created from two support vectors which delineates the classes on all of these dimensions. If we added "neck length" as a feature for waterfowl classification, all classification would be done in a three dimensional plane. If we added "leg length" and "wingspan", we'd move to a five dimensional plane, and so forth.

When the data is not linearly classifiable, or to allow further flexibility for complex feature/class relations, a "kernel trick" can be employed. In this approach, we use a different function to compute the similarity of new items to previous data. In the default, "linear" kernel, the classifier determines weightings for each feature (which determine the line described above), and classifies items based on the product of feature and weight in relation to the existing support vectors. Although many different kernels can be used, each with different mathematical properties, they all compare each new example to prior examples or support vectors, *using the data itself to determine similarity*, rather than creating abstracted feature weightings.

Thus, in a kernelized approach as described here, each novel item is compared against *each meaningful support vector*, yielding similarity to the data itself. This shows some similarity to exemplar-based models of speech perception, where new input is compared not based on parameters and cues, but based on similarity to previously-heard words and sounds[18].

Use of a kernel trick also has the advantage of better handling of non-linear feature-class relationships. Although a given kernel may be computing similarity using a very specific formula which cannot be easily understood in linear space, Mercer's Theorem dictates that any kernel classifier

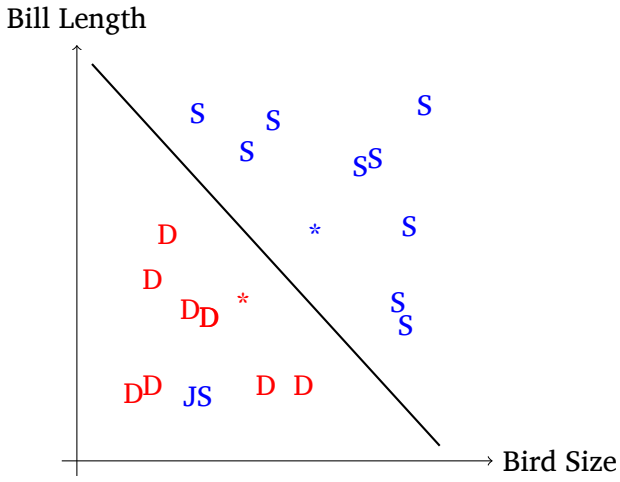
[18]This strong similarity further motivates the use of a kernelized SVM in this experiment, whose tacit goal is to optimally simulate human speech perception.

(using a well-constructed kernel) will be equivalent to a linear classifier in some expanded feature space. Thus, by using a kernel, data which is non-linear in a conventional feature space can still be linearly separated.

For a radial kernel (as used in our experiments), the feature space is seemingly projected onto a paraboloid, and thus, a single flat hyperplane (which intersects the paraboloid in different places) can classify data which seems in a conventional plane to require a curved or circular line[19]. Given these two advantages, we will be using the radial kernel throughout our SVM classifications.

Using a kernel does have one major disadvantage: a considerable loss of transparency in classification. In a linear SVM, we can examine the feature weightings explicitly to get a sense of the importance of different features for classification (much like the previously-discussed RandomForest importance measure). When an SVM is kernelized, these weightings are no longer used, and thus, we have to rely on RandomForest output to understand the usefulness of each individual feature.

One final aspect of SVM classification deserves mention. Occasionally, there are datapoints which simply don't fit the pattern. Imagine, returning to waterfowl, that we've been sent one more datapoint, a juvenile swan, marked "JS" below:



This swan, while clearly a member of the "swan" class, does not fit the overall pattern for swans based on these two size parameters. In order to make SVMs more tolerant of these odd points, SVMs include a "Cost" parameter, "C", which allows the user to tune how much these points on the "wrong side" of the hyperplane will affect the eventual model.

A higher C parameter increases the cost of being "wrong", and will result in a model which sacrifices some margin space in order to avoid known-misclassified datapoints. Although this results in a more conservative plane, this can lead to poor performance or overfitting when taken to extremes (as no size-based classifier should be expected to gracefully handle mini-swans or mega-ducks). On the other hand, a C parameter which is too low will result in a model which picks a hyperplane with very wide margins, but which mis-classifies many datapoints. Using a lower-than-infinite value for C results in a "soft margin" SVM, which explicitly accepts a certain amount

[19]The transformative power of the "kernel trick" for non-linear data is neatly visualized in a 42 second video which is worth several thousand words. - `https://www.youtube.com/watch?v=3liCbRZPrZA`

of error in order to avoid overfitting.

This C parameter must be "tuned" for individual models, and can greatly affect the classification accuracy of a given model[20]

Details aside, we can see that SVMs take a *very* different approach to machine classification. Whereas decision trees classified our birds on the basis of many cascading criteria[21], SVMs use the training data to find a single hyperplane which divides the feature space according to the labeled categories, and against which all future datapoints can be compared.

In addition to being the same approach chosen in Pruthi and Espy-Wilson (2004) for classifying nasal consonants, SVMs are extremely widely used in machine learning. More importantly, at the time of writing, they're considered the gold standard of machine learning algorithms: that is, even though some other algorithms (neural networks, particularly) can sometimes incrementally improve on the performance of SVMs, the improvement is seldom worth the additional complexity, tuning and processing power required.

So, in experiment 2, we'll be using RandomForests to help narrow and evaluate the list of features, and kernelized SVMs as a sort of "second opinion", using a very different algorithm with different strengths to evaluate the peak performance of these features (and groupings) for classification.

Now, with these two approaches in mind, we must discuss the precise nature of the task.

5.2 Experiment 2: Structure and Goals

As with Experiment 1, our fundamental goal is to evaluate the features, and to find the most promising features (and in this case, feature groupings) for perceptual testing in Experiment 3. We will accomplish this in three steps.

First, we will take the features extracted and discussed in Section 4, and for each one, train a single-feature model to classify "nasal" vs. "oral" in the English and French data using both the RandomForest and SVM algorithms. We will use the accuracy results from this as a baseline for our multi-feature studies, as well as to provide useful information about the utility of the features themselves.

Then, we will train two all-inclusive models (again, a RandomForest and an SVM), incorporating all 29 features, and we will examine some of the models' output to capture the relative importance of each feature. These all-inclusive models will also provide the "gold standard" of accuracy, the performance level incorporating as much information as is available.

Finally, we will create and implement several multi-feature models, designed to model potential subsets of features which could be used in the human perceptual experiment. Comparison of these models to one-another, alongside single-feature model results, will help inform our feature

[20] In our multi-feature models discussed in Section 5.8, tuning the C parameter resulted in an accuracy gain of nearly 5% for some models.

[21] Resulting in a waterfall waterfowl workflow for classification.

choices, and provide us with more useful information to be used in determining a final feature set in Section 5.8.4, as well as for comparison with human perceptual results later on.

5.3 Methods: Machine Perception (Experiment 2)

Although both machine learning and statistical analysis are fairly similar, there are a few key methodological differences which merit discussion.

5.3.1 Software used for classification

As with the statistical processing, all classification and generation of tables and graphics was conducted within the R Statistics suite, for both ease of use and ease of reproducibility.

Within R, for RandomForest models, the `randomForest` package was used (Liaw and Wiener (2002)). For Support Vector Machines, the *e1071* package was used (Meyer et al. (2014)). `e1071` is an R interface to the well-known LIBSVM library (Chang and Lin (2011)), which is a widely-used standard in SVM classification.

5.4 Preparing the data for classification

Although the data used in machine learning are drawn from the same master file as those used in the statistical experiments in Section 4, the nature of the machine learning task requires that some changes be made.

5.4.1 Scaling the Dataset

One problem with these data is that many different units and scales are in use. Formant heights and bandwidths vary on the order of 100s of Hertz for oral vs. nasal vowels, whereas AX-PX measurements vary within 3-7dB. For machine learning, these variations in meaningful difference cause difficulties, as the properties and nature of these different scales and unit types aren't interpretable or known to the algorithms.

To simplify the classification, and reduce the problems inherent with speaker variability, all features here are z-scored by speaker prior to machine learning use. In this process, the mean for the feature is taken, and each feature's unique value (e.g. In Hz or dB) is converted to a value based on the direction and number of standard deviations it varies from the mean. Thus, if for a given feature a speaker had a mean of 4 and a standard deviation of 2, a token measuring 6 would be Z-Scored to "1", and a token measuring 3 would be "-0.5".

This process removes variation from both speaker variation (e.g. different baselines for A1-P0) and unit variation (a 1 SD difference is always "1", no matter whether it corresponds to 2.4 dB or 107 Hz), reducing (but by no means eliminating) the variability in the data stemming from variable baselines and ΔFeature values for different speakers.

Note, however, that this scaling is the sole place where speaker will play a role in our machine learning studies. Speaker itself cannot be included in the model, as it is in no way directly predictive of oral vs. nasal distinctions, and there is no way in these models to include speaker*feature interactions, short of creating and training a different model for each speaker (which would not be particularly useful for general classification).

5.4.2 Balancing the Dataset

One important difference between conventional statistics and machine learning is that some machine learning tasks are adversely affected by unbalanced data sets.

This makes sense: If asked to classify a pen with 99 horses and a single zebra, one can achieve 99% accuracy simply by guessing "Horse" for all 4 legged creatures, and it would be a poor learning experience on top of that.

Imbalance also complicates the interpretation of accuracy in classification, as although there are two answers, one is far more likely. Thus, "chance" for a two-class is roughly the percentage-of-dataset for the most common category.

Table 25 shows the number of "oral" vs "nasal", tokens (and their percentage) in several different possible datasets, after bad tokens (i.e. tokens flagged by the measurement script) have been removed:

Table 25: Number of clean oral vs. nasal measurements

Dataset	NumOral	NumNasal	PercentNasal
French	$1,356$	$1,371$	50.275
English	$2,571$	$7,670$	74.895
All Data (All English)	$3,927$	$9,041$	69.718
English (CVC vs NVN)	$2,571$	$2,765$	51.818
All Data (English CVC/NVN)	$3,927$	$4,136$	51.296

In French, this is not a problem, as we have roughly equal numbers of oral and nasal vowels. However, English is poorly balanced, as we've collected three nasal tokens (CVN, NVC, NVN) for every oral token (CVC), and this in turn affects the entire dataset ("All Data (All English)").

The solution here is to remove CVN and NVC tokens from the classification task for English. We've already shown that patterns of correlations are similar for CVC/CVN-NVC-NVN and CVC/NVN, and the degree of nasality is stronger for the CVC/NVN task, maximizing classification power. Most importantly, if we exclude NVC and CVN data, we can see that the oral/nasal split for English and for the dataset overall returns to near-even.

Thus, for all machine learning tasks, our "English" dataset will compare only NVN to CVC.

5.5 Experiment 2: Criteria for Feature Evaluation

Our goal here is again to evaluate the features (and various groupings), and to identify the most promising features for perceptual experimentation. To do this, we will use two different types of data that machine learning experiments generate: accuracy and importance.

First, and most important, both SVMs and tree-based classifiers will return an accuracy for each set of features. This indicates how often the classifier was correct (i.e. it called an oral vowel oral or a nasal vowel nasal) using that particular feature or feature set.

Accuracy, naturally, will be the most important feature for this work, as it provides a straightforward analog to the accuracy measures made in perceptual studies. By comparing relative accuracy among features and feature groups, we can get an understanding of the amount of information each provides towards nasal classification, and identify the "best" features, that is, those most likely to also be meaningful to human listeners.

For RandomForest models with multiple features, we can extract another type of data, called "importance". Although it is described in more detail in Section 5.7, Importance provides a measure of the contribution of each feature towards reducing the overall error rate. This measure allows us to "peek behind the curtain", and to see not just the results (accuracy) of the different multi-feature models, but to see which of the features is providing most of the power. In conjunction with accuracy (for evaluating the model as a whole), importance provides a second data point which allows us to make better statements about the nature of these features.

Thus, we can establish two criteria which we can use for identifying the most promising features and feature groups:

1. The feature or grouping must display good *accuracy*, relative to other features and models.
2. Any given feature should show good *importance* relative to other features in multi-feature models.

Using this information, we can now proceed to our first sub-task within the experiment, Single-feature Classification.

5.6 Experiment 2: Single-feature models

We must first begin by capturing a baseline, understanding how useful *any particular feature* is for machine classification. To do this, we will generate single-feature models for each of our features for each language, classifying the data as oral or nasal using decision trees and SVMs using *only that particular feature*.

This will allow us to evaluate the features from a second perspective, seeing not only their relationship with nasality in the data, but their token-by-token utility for classification.

5.6.1 Methods: Single-feature classification

For consistency and ease of comparison, we will use `randomForest` for both single-feature and multi-feature classification. So, for a single-feature model testing P0Prominence in English:

```
randomForest(nasality ~ P0Prominence.z, ntree=500, data = en)
```

Put into plain English, we are predicting nasality using a Z-scored version of P0Prominence (P0Prominence.z), and running on the English dataset. We are generating 500 trees, allowing the model 500 chances to find the proper decision points[22].

For Support Vector Machine testing, the code is similar:

```
svm(nasality ~ P0Prominence, kernel="radial", cost="1", cross = 10,data=en)
```

Here, we are again predicting nasality using P0Prominence, in the English data. We are using a Radial Basis ("RBF") Kernel as recommended by Hsu et al. (2003), to account for potentially non-linear boundaries for categories, and as doing so resulted in the best accuracy. A variety of values were tested for the `cost` (or "C") parameter, which sets the penalty associated with missed classifications, but a value of 1 ultimately yielded the best results. The gamma parameter used is the default for `e1071`, that is, 1/[number of features]. "cross = 10" states that we will use 10 fold cross-validation, as described in Section 5.1.

Finally, all models were run on two separated datasets, English (CVC-NVN) and French. Comparison of these two language-specific models will provide useful data for the cross-linguistic comparison to be conducted in Section 5.8.4.

5.6.2 Results: Single-Feature Classification by Algorithm and Dataset

First, we will examine the classification in summary, comparing the accuracy of SVM and RandomForest classification for all 29 features, for all the datasets in Table 26:

Table 26: Accuracy by Algorithm and Dataset in English and French

Statistic	N	Mean	St. Dev.	Min	Max
en.svm.acc	29	59.218	4.503	51.237	67.635
en.rf.acc	29	55.417	5.486	50.862	79.704
fr.svm.acc	29	63.731	7.586	51.118	76.824
fr.rf.acc	29	57.705	6.965	49.065	79.208
avg.svm.acc	29	61.474	5.293	53.011	71.529
avg.rf.acc	29	56.561	5.756	51.040	79.456

We can immediately see that single-feature accuracy across the entire feature set is not particularly good, with average accuracies at 61% for SVMs and, 56% for RandomForests. Recall that 50%

[22] Although we don't expect a single-feature model to require 500 iterations to find an optimal set of decision points, using the same parameters allows more straightforward comparison to the models we will use for multi-feature comparisons.

accuracy could be obtained with a coin-flip in this task, and ~51% accuracy can by had by always guessing "nasal". Although some individual features do perform better (as described below), this implies that any given feature alone may have little predictive power.

We also see that SVMs seem to generally have higher accuracy than RandomForests, but the difference is not night-and-day. This finding is not particularly surprising given the nature and tuning of the algorithms, but their relative closeness is helpful, allowing RandomForests a respectable place in the overall analysis.

Finally, for both algorithms, classification accuracy is higher in French than in English. The significance of this will be discussed in greater depth in Section 5.9, but for now, it will suffice to say that classification by feature is *not* language-agnostic, and that classification of French nasals appears to be somewhat easier than classifying English.

Now, let's examine the individual features' accuracy.

5.6.3 Results: Single-Feature classification by feature

The by-feature classification results for English (Table 27) and French ((Table 28) are displayed below, following the below format:

- **Features**: The short names of the features, as listed in Table 5.
- **language.rf/svm.acc**: The accuracy of a randomForest or svm classifier, as a percentage, for English (en) or French (fr).
- **Avg.RF/SVM.acc**: The mean accuracy for both classifiers.
- A "**- % Nasal -** row is included for each dataset. This represents the accuracy of a"dumb" model which always guesses "nasal". This forms the low-end baseline, at or beyond which a model can be understood to be completely worthless.

As we examine the features in both datasets, a few promising candidates immediately emerge.

Immediately, we see that Duration is the best overall cue for classification, particularly in a decision-tree style classifier. In both English and French, ~79% accuracy can be obtained simply by looking at the duration of the vowel. This performance is quite reasonable, given that it should be fairly speaker-independent, and that there is a good separation between the categories. Thus, it's even more clear that Duration cannot be ignored.

This is also one of the few cases where RandomForests and SVMs diverge sharply: in English, the RandomForest model using Duration performs 21% better than the SVM model. This is testament to the fact that although these algorithms do the same thing, they do so in a fundamentally different way. RandomForests consider single features sequentially, and when there's one feature with a very sharp boundary ("Vowels longer than X are likely to be nasal" or vice versa), the sorts of "snap-judgement" decision making favored by RandomForests will have the advantage over the "all points considered" approach of SVMs.

As we might expect, the A1-P0 family also does exceptionally well in both languages, and in both algorithms. In terms of its components, Amp_F1, closely related, performs well in both French and

Table 27: SVM and RandomForest accuracy by feature in English (CVC/NVN)

Features	en.svm.acc	en.rf.acc	avg.en.acc
Duration	59.220	79.704	69.462
A1P0_HighPeak	67.260	61.900	64.580
A1P0_Compensated	67.448	60.401	63.924
Width_F1	67.635	58.358	62.997
LowZeroDiffP0	65.892	57.046	61.469
LowZeroDiffA1	64.037	56.578	60.307
Amp_F1	62.556	56.709	59.633
Freq_F1	59.145	59.239	59.192
Amp_P1	62.425	55.885	59.155
A1P1	60.495	56.353	58.424
Freq_F3	61.057	54.873	57.965
A3P0	61.675	54.085	57.880
Amp_P0	61.057	54.254	57.656
P0Prominence	59.876	54.760	57.318
MidZeroDiffP1	59.539	54.835	57.187
A1P1_Compensated	59.820	54.348	57.084
Width_F3	58.433	53.017	55.725
Amp_F3	57.815	52.999	55.407
A1P2	56.540	53.542	55.041
Freq_F2	56.128	53.017	54.573
SpectralCOG	56.859	51.987	54.423
Ratio_F1F2	56.091	52.343	54.217
H1MinusH2	55.941	51.443	53.692
Width_F2	54.310	52.736	53.523
Amp_P2	55.660	51.068	53.364
Ratio_F2F3	52.867	52.661	52.764
Amp_F2	53.355	50.862	52.108
P1Prominence	52.942	50.918	51.930
- % Nasal -	51.818	51.818	51.818
MidZeroDiffA1	51.237	51.181	51.209

Table 28: SVM and RandomForest accuracy by feature in French

Features	fr.svm.acc	fr.rf.acc	avg.fr.acc
A3P0	76.824	68.317	72.571
Duration	65.787	79.208	72.497
Amp_F3	76.788	67.657	72.222
A1P0_HighPeak	75.798	67.070	71.434
A1P0_Compensated	75.138	66.300	70.719
Amp_F1	74.697	64.173	69.435
LowZeroDiffP0	72.094	60.946	66.520
Width_F1	69.380	60.946	65.163
Width_F3	68.720	58.306	63.513
Amp_F2	66.630	59.406	63.018
Freq_F3	66.813	55.996	61.404
SpectralCOG	64.356	57.462	60.909
H1MinusH2	63.806	57.242	60.524
P0Prominence	64.210	56.766	60.488
Width_F2	64.173	56.142	60.158
Amp_P2	63.623	53.905	58.764
Freq_F1	56.032	59.369	57.701
LowZeroDiffA1	61.643	53.722	57.682
Amp_P1	59.699	54.969	57.334
Ratio_F2F3	60.249	54.089	57.169
Freq_F2	59.773	54.309	57.041
A1P1	57.902	53.319	55.611
P1Prominence	55.996	52.219	54.107
A1P1_Compensated	58.379	49.065	53.722
MidZeroDiffP1	56.106	51.192	53.649
MidZeroDiffA1	54.785	50.898	52.842
A1P2	54.235	49.615	51.925
Ratio_F1F2	53.429	49.872	51.650
Amp_P0	51.118	50.972	51.045
- % Nasal -	50.275	50.275	50.275

English, but Amp_P0 performs rather poorly in all models. We also see that A1P0_HighPeak outperforms A1P0_Compensated overall, and that several other related measures (LowZeroDiffP0/A1) are similarly well ranked.

P0Prominence performs disappointingly for classification in both languages, towards the middle of the pack, and averaging 57% and 60% overall accuracy in English and French, respectively. Although this bests many features, based on these data, P0Prominence alone is unlikely to be a useful cue.

A3-P0 continues its exceptional performance in French, as the best performing SVM feature, and second-best for RandomForest (behind the sharply-boosted Duration feature). This is again due to Spectral Tilt (which also explains Amp_F3's performance in French). In English, it performs considerably more poorly (11th overall), but is still towards the top of the pack.

Formant bandwidth performs notably well in both languages, particularly F1's bandwidth, although formant frequency performs rather poorly, which is not surprising from an algorithm which is fundamentally unaware of vowel categories.

P1-related features continue to perform poorly in English (58% accuracy overall for A1P1), and not at all in French (55% accuracy for A1P1), while P1Prominence performs more poorly still. The strong by-vowel-height variability (described in 4.5.7) certainly plays a role here, with P1 being even potentially useful only for high vowels (which aren't found at all in French in this dataset), and explains the difference in performance between the languages. However, we again see evidence that P1, if useful at all, is only useful in high vowels, and should not be used for more generalized evaluation of nasality.

Finally, we do see that alongside our strong performing features (in the high 60s to low 70s), there are some features which provide next-to-no information about nasality, with overall accuracies in the low 50s[23]. This simply serves to confirm that feature selection *does* matter, and not all features, even in a group chosen *a priori* to provide useful information about nasality, will actually do so in isolation.

5.6.4 Discussion: Single Feature Classification

There are three main takeaways from the Single Feature classification experiment described above.

First, we see that single feature classification is simply not a great idea. The highest accuracy possible in these data, using Duration in a decision tree classifier for English, only provides 79.7% accuracy, and the average accuracy was in the high 50s to low 60s, depending on language and model. Thus, even our best-performing model does not provide acceptable accuracy for speech recognition (human or otherwise), and rather neatly rules out a perceptual model where only one feature is attended to.

Second, we see that although RandomForest and SVM classification produced similar outputs and feature accuracies, they are not identical. SVM classification proved more accurate overall, and

[23] In a rather impressive feat of uselessness, MidZeroDiffA1 actually manages to perform more poorly than the "It's all nasal" dumb model in English.

each classifier worked better with some features than the other (e.g. RandomForests' affinity for duration). Thus, the use of both classifiers is neither redundant nor unnecessary.

Finally, we see that our feature choices from Section 4.6 (with the exception of P0Prominence) tended to be towards top of the pack, suggesting that our choices were not poorly made. These may constitute a useful grouping of features for group classification, both for the perception task and for multi-feature evaluation.

5.7 Experiment 2: Evaluating Feature Importance

At this point, we know that no single feature will accomplish the classification task with acceptable accuracy, we must generate promising groupings of features, both for classification and for our eventual perceptual task.

Although we know the relative accuracy rates of these features in single feature models, this is a particularly poor analogy to human perception, as humans are highly unlikely to use a "single feature model", unless one feature provided such good information as to negate the usefulness of all others (and this is clearly not the case based on these data). When identifying oral and nasal vowels, humans have access to the entire signal, and thus, access to any (and every) possible "acoustical feature". The choice of cues in perception is not so much a choice of which cue or cues to use, but instead, which to *focus on*.

The closest analogy to this process in machine learning is the development of feature rankings, which show the *relative utility* of different features for classifying a particular category. This provides (limited) insight into the functioning of the model, will help us to eliminate those features which provide little classification power not just alone, but as a part of the greater analysis.

Because the Kernelized SVMs used here provide no feature weights, we will examine the 'importance' of each feature as given by an all-inclusive RandomForest model, and use these to find those features which seem most useful for the classification task. This will provide additional data for generating possible feature groupings.

5.7.1 Methods: Using and Interpreting RandomForest Importance Measures

The same data and fundamental method is used for multi-feature RandomForest classification as was used for single-feature classification, except for the number of features. So, to create an all-inclusive RandomForest model (called en.allinclusive.rf), we would use:

```
en.allinclusive.rf <- randomForest(nasality ~ Amp_F1.z + Amp_P1.z + …[ and the
   rest of the features], ntree=500, data = en)
```

We must then extract Importance values. In RandomForest models, importance is determined by calculating the effect of permuting each feature throughout the dataset (such that each token now has a different token's value for the feature), and then measuring the increase in classification error that comes with this permutation. To give a more concrete example, if measuring the importance of color for classifying 100 birds into species, you would randomize the "color" column, such that

a bird which was formerly "black" is now labeled "white" or "brown", and so on, and then re-check the accuracy of classification given by the birders with this purposefully faulty information. Whereas permuting bird color might have a major negative effect on accuracy, permuting bird location-on-lake would likely not. Thus, color is "more important" than location-on-lake.

Features with little relation to nasality will show little increase in classification error when permuted, and permuting strongly related and useful features would cause significant reductions in accuracy.

To extract the importance values, we simply use the `importance` method within the `randomForest` package, running it on a previously generated model:

```
importance(en.allinclusive.rf)
```

This will display a table of features and importance measures for the model, which can then be compared to develop a clearer understanding of what these features contribute to the overall models.

5.7.2 Results: Evaluating Feature Importance

Table 29 shows the top 15 features in English and French, ranked by importance in each language, giving raw values as well as that feature's percentage relative to the total importance assigned.

Table 29: Top-Ranking 15 Features by importance in an all-inclusive model for English and French

	en.TopFeatures	en.imp	Percentage	fr.TopFeatures	fr.imp	Percentage.1
1	Width_F1.z	203.163	7.626	A3P0.z	154.236	11.317
2	Duration.z	178.329	6.694	A1P0_HighPeak.z	117.563	8.626
3	A1P0_Compensated.z	161.119	6.048	Amp_F3.z	109.633	8.044
4	LowZeroDiffP0.z	149.373	5.607	A1P0_Compensated.z	88.651	6.504
5	A1P0_HighPeak.z	131.394	4.932	Amp_F1.z	72.183	5.296
6	Width_F3.z	115.274	4.327	Width_F1.z	66.186	4.856
7	A3P0.z	98.705	3.705	Amp_F2.z	61.176	4.489
8	A1P1.z	93.531	3.511	LowZeroDiffP0.z	60.965	4.473
9	LowZeroDiffA1.z	93.207	3.499	Ratio_F2F3.z	56.555	4.150
10	Freq_F1.z	92.656	3.478	Freq_F3.z	50.727	3.722
11	Freq_F3.z	84.756	3.182	Freq_F2.z	50.207	3.684
12	Amp_P1.z	84.246	3.162	Width_F2.z	47.144	3.459
13	Amp_P0.z	83.341	3.128	Duration.z	45.684	3.352
14	MidZeroDiffP1.z	81.903	3.074	Width_F3.z	41.831	3.069
15	P0Prominence.z	81.632	3.064	Ratio_F1F2.z	33.616	2.466

First and foremost, we see that the rankings are *not* the same across languages. Whereas A3-P0 is the most important feature for French (along with Amp_F3, another marker of spectral tilt), it is only ranked 7th in English. Similarly, Duration proves a much more important feature for

English (2nd) than French (13th). This further suggests that although there are many similarities, the classification of nasality in English and French are two *very* different tasks.

A1-P0 (in both forms) appears within the top 5 for both languages, along with other related measures. Interestingly, in English, neither A1 nor P0 alone show strong importance, and in French, P0 is not even present in the top 15. This indicates that the composite measure serves the purposes of classification considerably better than its components. P0Prominence shows very little importance, 15th in English, off the chart in French.

Formant Bandwidth again shows strong utility, with F1's bandwidth as the top ranked feature in English (6th in French), and F3's width ranked 6th and 10th in English and French, respectively. And yet again, Formant frequency performs relatively more poorly than the other features, again suggesting that absolute formant changes (independent of vowel) aren't terribly useful for classifying nasality.

Finally, the numbers so far have reflected our entire feature set, where there is considerable redundancy, as multiple measures all reflect the A1/P0 pole-zero complex, multiple measures show spectral tilt, etc. Although this is good for finding the best exemplar of each phenomenon, this distributes some of the classification importance across multiple features all representing the same acoustical consequence. Table 30 shows the same analysis, but when run on a reduced-redundancy grouping (described in Section 5.8.2) which reduces the feature set, leaving a more independent set of features:

Table 30: Features by importance in a Reduced Redundancy model for English and French

	en.TopFeatures	en.imp	Perc	fr.TopFeatures	fr.imp	Perc.1
1	Width_F1.z	345.451	12.968	A3P0.z	229.867	16.865
2	A1P0_HighPeak.z	311.005	11.675	A1P0_HighPeak.z	203.614	14.939
3	Duration.z	273.031	10.249	Width_F1.z	140.347	10.297
4	A3P0.z	200.558	7.529	Amp_F2.z	117.135	8.594
5	A1P1.z	194.796	7.312	Freq_F2.z	99.098	7.271
6	Freq_F1.z	181.496	6.813	Freq_F3.z	95.919	7.038
7	Width_F3.z	172.531	6.477	Width_F3.z	82.292	6.038
8	P0Prominence.z	163.434	6.135	Duration.z	79.967	5.867
9	Freq_F2.z	158.767	5.960	Width_F2.z	79.783	5.854
10	Freq_F3.z	157.541	5.914	P0Prominence.z	62.744	4.604
11	A1P2.z	146.892	5.514	A1P1.z	53.751	3.944
12	Width_F2.z	127.970	4.804	A1P2.z	43.531	3.194
13	P1Prominence.z	115.422	4.333	Freq_F1.z	39.726	2.915
14	Amp_F2.z	115.043	4.319	P1Prominence.z	35.173	2.581

This brings the importance of each phenomenon into sharper focus. Those features which were important in the all-inclusive (yet highly redundant) model (F1's bandwidth, A1-P0, Duration, and A3-P0) are now much more so. In addition, we see more clearly a jump between the "very useful" top few features, and the less useful features below. Although practically speaking, the results do not change, they become far more easily interpretable.

5.7.3 Discussion: Evaluating Feature Importance

From this sub-experiment, we learn three main lessons.

First, we see that in a greater multi-feature model, not all features are created equally. The top five features by importance in English represent ~29% of the total importance in a composite model, and almost 40% in French. When feature redundancy is reduced, his goes up to 44% and 51% respectively. Thus, we see that there is ample opportunity for feature-set reduction, without major loss of classification power.

Second, we see very strong showings for Formant Bandwidth, A3-P0, Duration, and the A1-P0 family. We already knew that these features are correlated with nasality and useful for predicting it on their own (given their single-feature performance), and now, we know that these features are critical parts of a larger, multi-feature model.

Finally, and again crucially, we see that English and French are different, and classification of them depends on different factors. A3-P0 is far more important in French than in English, and Formant Bandwidth is far more important in English than in French. So, although we are doing the same fundamental task, we must go about it in different ways.

5.8 Experiment 2: Multi-feature models

At this point, we have four different perspectives on the individual features. We have evaluated each feature in terms of whole-dataset statistical relationships in Section 4, we have seen the solo performance of each feature in both RandomForest and SVM classifiers in Section 5.6, and we have seen the importance of each feature as part of a larger composite model, above in Section 5.7.

Ultimately, though, the goal of the statistical and machine learning studies is to help us to select a subset of features for modification and testing in the perception study. Effectively, we want to choose a group of features which is both minimally complex and maximally informative in terms of nasality.

Unfortunately, in machine learning, it is generally true that more (reasonably meaningful) features leads to greater performance, and thus, our strategy for selecting features for evaluation must be to find a balance between these two extremes, and identify the smallest grouping possible without sacrificing too much predictive power.

To this end, we will now test not just individual features, but subsets of the greater feature set, with the the hopes of finding a set which fits this goal.

5.8.1 Method: Multi-Feature Classification

Multi-feature models will be conducted using the same software and settings as the single-feature models, discussed in Section 5.6.1. The principal difference will be the addition of other features to the model, creating commands such as the below, where [features] is a list of the features included in the model:

```
svm(nasality ~ [features], kernel = ""radial, cost = 7, cross = 10,data=en)
randomForest(nasality ~ [features], ntree=500, data = en)
```

In addition, after tuning, the best accuracy for these multi-feature SVM models was found to come from using a C value of 7 (although C = 1 was still used for the two single-feature comparison points). The gamma parameter used (in all models) is again the default for `e1071`, that is, 1/[number of features]. For consistency, all other parameters (ntree, and the number of cross-validation slices) will remain the same as elsewhere in experiment 2). The accuracy of each group model will be compared and discussed below.

It is worth noting that although the models here seem quite complex and unwieldy, both RandomForest and SVM classifiers are built to work with incredibly large feature sets. For RandomForest models, the many trees grown will vary not just in terms of decision points, but in terms of feature ordering ("First check A1-P0, then duration…"), and the best combination of decision points and feature ordering will serve as the final model.

For SVMs, when creating our 29-feature all-inclusive model, a single hyperplane is drawn which separates the two categories in a 29-dimensional space. However tough to visualize, the capability is built in to the models, and it is not uncommon in some fields (natural language processing, particularly) to conduct classifications with many millions[24]) of automatically extracted features.

So, although classification on the basis of this many features may seem intractably complex on the surface, this is the sort of task that these algorithms were designed for, and even our most complex 29-feature model comes nowhere near the practical or useful limits of these classifiers.

5.8.2 Multi-Feature Classification Groups

First, we must describe the groupings to be formally tested, and the data and ideas which led to their creation[25]. Each feature grouping is listed below, with the number of features used, a short description of the motivation behind the grouping, and a list of the included features.

All-Inclusive The all-inclusive grouping represents our "upper bound", and should offer the highest performance of all. It contains all 29 features investigated in the present work, as listed in Table 5.

Reduced Redundancy This 14 feature grouping was designed to capture all *phenomena* being investigated, using the best features for each. Although there remains some redundancy (e.g. A1 is referenced by many of the features), and changes to some of the features below could affect others, each one could, plausibly, be independently useful for the perception of nasality.

[24]For instance, in a topic identification task, we might automatically generate "Sentence contains word X" and "Sentence contains bigram Y" features for *every word and combination of words* in a large corpus of text.

[25]Additional groupings were investigated, and the below represent the best performing or otherwise most desirable of those informally tested.

It should also be noted that removing the redundant features here also removes the atomic features which compose A1-P0, A1-P1, and A3-P0. The difference in performance between the all-inclusive model and the reduced-redundancy model can be, at least in part, attributed to the difference between using composite features and atomic features.

The components of this group are listed below, along with their reasoning for inclusion and selection:

- **A1P0_HighPeak** - Our chosen feature for the A1-P0 pole-zero complex, capturing Amp_F1, Amp_P0, etc.
- **A1P1** - Our chosen feature for the A1-P1 pole-zero complex
- **A3P0** - Our chosen measure of spectral tilt, supplanting SpectralCOG and H1MinusH2.
- **Amp_F2** - The sole formant whose amplitude is not incorporated into other measures.
- **A1P2** - Our chosen feature to capture the effects, if any, of P2.
- **Duration** - Duration is surely unrelated to our other acoustical features
- **Freq_F1/F2/F3** - We can consider vowel formant shifts as potentially independent phenomena, and these raw values supplant RatioF1F2 and RatioF2F3
- **P0/P1Prominence** - Again, although we've used P0 and P1 as part of greater pole/zero measures, it's possible that their relative prominence is separately indicative
- **Width_F1/F2/F3** - Formant bandwidths, although affected by spectral tilt, could easily serve as independent cues.

So, our reduced redundancy model consists of:

A1P0_HighPeak - A1P1 - A3P0 - Amp_F2 - A1P2 - Duration - Freq_F1 - Freq_F2 - Freq_F3 - P0Prominence - P1Prominence - Width_F1 - Width_F2 - Width_F3

Preliminary Features This 10 feature grouping corresponds to our preliminary feature set derived from the statisticalarchives data, as discussed and outlined in 4.6.

A1P0_HighPeak - A3P0 - Duration - Freq_F1 - Freq_F2 - Freq_F3 - P0Prominence - Width_F1 - Width_F2 - Width_F3

Single-Feature Top 10 This grouping contains the top 10 non-redundant features, in terms of overall Single-Feature performance across both languages. As the same 10 features are indicated by both SVM and RF classification, only one grouping is required. Included are:

A1P0_HighPeak - A3P0 - Amp_F1 - Amp_F3 - Amp_P1 - Duration - Freq_F3 - P0Prominence - Width_F1 - Width_F3

Best Importance Top 10 This grouping contains the Top 10 features *by importance in an all-inclusive model.*

A1P0_HighPeak - A3P0 - Amp_F2 - Duration - Freq_F1 - Freq_F2 - Freq_F3 - P0Prominence - Width_F1 - Width_F3

ADW3 (A1P0, Duration, Width, A3P0) This is a minimal six feature grouping, containing the four top feature types, based on importance and single-feature results.

A1P0_HighPeak - A3P0 - Duration - Width_F1 - Width_F2 - Width_F3

A1-P0 Only This is the accuracy using just A1P0_Highpeak, for comparison.

Duration Only This "grouping" contains only our top performing single feature, Duration.

5.8.3 Results: Multi-feature classification

It is worth noting that, barring completely meaningless features, more features leads to greater accuracy. We should fully expect the All-Inclusive model to outperform all others, and similarly for the 17 feature Reduced Redundancy model to best those with fewer features. So, when making comparisons among models, we should consider not just absolute performance, but performance in the context of the number of features used to achieve that performance.

With this in mind, Table 31 shows the results of classification using each tested feature group, ordered by overall SVM performance.

Table 31: Accuracy by feature grouping in English (CVC/NVN) and French

Grouping	N	en.svm.acc	en.rf.acc	fr.svm.acc	fr.rf.acc	avg.svm	avg.rf
All-Inclusive	29	84.764	84.539	93.729	91.566	89.247	88.052
LowRedundancy	14	83.171	84.539	91.896	91.309	87.533	87.924
Preliminary	10	82.290	83.902	91.713	90.906	87.001	87.404
Importance10	10	82.084	83.883	91.749	90.942	86.917	87.413
SingleFeature10	10	80.922	82.046	88.009	88.045	84.465	85.046
ADW3	6	75.787	78.017	84.745	86.432	80.266	82.225
A1-P0 Only	1	67.279	61.863	76.128	66.923	71.703	64.393
Duration Only	1	65.640	79.171	65.640	79.171	65.640	79.171

First, we see the first truly impressive results from our machine classification, with a 93.7% accuracy rate for oral vs. nasal vowels French in an all-inclusive SVM model. Clearly, we have captured a useful set of nasal features, which, combined, is able to approach usable accuracies.

Again, English proves more difficult to classify than French, with the sole exception of Duration-only classification. We also see that RandomForest and SVM classification are often within a few percentage points of one-another, occasionally trading off for "best performance", particularly for English, where Duration is a very useful feature for decision trees.

We must now return to our two principal requirements for a feature grouping: We want a grouping which includes as few features as possible, while still selecting all of the features strongly associated with nasality.

As we compare the Preliminary grouping with the All-Inclusive grouping, we see that by adding 19 additional features, we only gain around 2% accuracy overall. This would seem to indicate that those features which are missing in the Preliminary model aren't terribly useful for classification overall.

At this 10 feature level, it's clear that feature choice is important. Rather conveniently for the author, the "Preliminary" group proposed in Section 4.6 (containing A1P0, A3P0, Duration, Formant Width and Frequency, and P0Prominence) outperforms every other 10 feature model (except in English SVMs), and thus, appears the best choice, especially given its overall simplicity and concordance with the previous statistical analysis.

On the other hand, when we omit formant frequency and P0Prominence altogether and move to only four features in the ADW3 model (A1P0, Duration, Bandwidth and A3P0), we lose nearly 10 points worth of accuracy. At this level, we're still outperforming the single-feature models (with Duration as an exception), but significant accuracy has been lost.

Based on these data, the "Preliminary" feature grouping appears to be the best balance of complexity (using only 10 features representing six phenomena) and accuracy (with 87% accuracy overall)[26]. However, there may be room for further improvement.

5.8.4 Discussion: Finalizing the Feature Set

At this point, it appears that the best combination of performance and simplicity comes from the "Preliminary" model, which uses:

- A1P0_Highpeak
- A3P0
- Duration
- Formant Bandwidth (Width_F1, Width_F2, Width_F3)
- Formant Frequency (Freq_F1, Freq_F2, Freq_F3)
- P0Prominence

However, we may be able to "tune" this model further. To do so, we'll first consult a ranking of these features by relative importance in our Preliminary RandomForest model, following the same methods outlined in 5.7.1. Importance data is shown in Table 32.

Even in this small feature set, we see again the striking differences between English and French, and specifically, that importance in this model is not uniform. For instance, A3-P0's relative

[26]It is certainly possible that there is another combination of ten features which can obtain higher performance through some unexpected feature synergy. However, due to strong performance of these existing models, as well as the roughly 6.3 years of computing time required to test the 20,030,007 other possible ten-feature models, the author decided to constrain the search space to the most probable *a priori* candidates, the best of which are reported above.

Table 32: Features by importance in the Preliminary Feature Set model for English and French

	en.TopFeatures	en.imp	Perc	fr.TopFeatures	fr.imp	Perc.1
1	Width_F1.z	416.701	15.642	A3P0.z	300.872	22.076
2	A1P0_HighPeak.z	366.206	13.747	A1P0_HighPeak.z	237.802	17.448
3	Duration.z	340.037	12.764	Width_F1.z	151.255	11.098
4	Freq_F1.z	296.712	11.138	Freq_F3.z	128.672	9.441
5	A3P0.z	238.597	8.956	Width_F2.z	112.591	8.261
6	Width_F3.z	219.978	8.258	Freq_F2.z	109.118	8.006
7	Freq_F2.z	211.622	7.944	Duration.z	97.224	7.133
8	Freq_F3.z	200.300	7.519	Width_F3.z	94.808	6.956
9	P0Prominence.z	200.150	7.513	P0Prominence.z	72.780	5.340
10	Width_F2.z	173.672	6.519	Freq_F1.z	57.798	4.241

importance in French is far higher than in English, and Duration plays a far greater role in English than in French.

This complicates the process of further thinning the dataset. While F2's bandwidth is the least important in English, it is in the top 5 features for French. Similarly, while F1's frequency is at the bottom of the rankings in French, it's ranked 4th in English.

Some poor performers do emerge. Both Freq_F2 and Width_F3 are consistently in the bottom 5 (trading for 6th and 7th place in English and French), and P0Prominence is ranked 9th out of 10 in both languages.

In Table 33 are the results of three models, the All-Inclusive model (as a baseline), our base Preliminary model, as well as a model lacking P0's Prominence ("noprom").

Table 33: Accuracy by feature grouping in English (CVC/NVN) and French for Preliminary Sets

Grouping	N	en.svm.acc	en.rf.acc	fr.svm.acc	fr.rf.acc	avg.svm	avg.rf
All-Inclusive	29	84.764	84.539	93.729	91.566	89.247	88.052
Preliminary	10	82.290	83.902	91.713	90.906	87.001	87.404
Pre-NoProm	9	82.271	83.564	90.869	90.869	86.570	87.217

Models without Freq_F2 (Pre-NoFreqF2) and Width_F3 (Pre-NoWidthF3) suffer a noticeable performance reduction, losing ~0.75% and ~1% accuracy respectively. Given this drop in accuracy, as well as the fact that we will already be manipulating formant frequency and bandwidth for other, higher ranked features, removing these features would do little to reduce the overall complexity of the process, and would do so at a non-trivial cost to accuracy.

On the other side, We can see here that "Pre-NoProm", the grouping without prominence, performs almost as well as the full "Preliminary" grouping, losing only ~0.2% accuracy overall. Although our statistical studies have shown that P0Prominence is clearly associated with nasality (See Section 4.5.2), it has shown poor classification performance overall, has a very close relationship to A1-P0, and in these studies, the effects of its removal from the models was negligible.

Given the already sizable number of conditions to test in the perception study (five types of feature, plus a combined grouping), the chance of P0Prominence being perceptually crucial enough to merit the additional experimental complexity seems low, and thus, we will not carry it through to the final stages of this project.

Thus, from a combination of excellent performance in the statistical studies and selection through machine learning, our final, most promising feature set emerges:

- A1P0_Highpeak
- A3P0
- Duration
- Formant Bandwidth (Width_F1, Width_F2, Width_F3)
- Formant Frequency (Freq_F1, Freq_F2, Freq_F3)

In addition to showing strong statistical links to nasality in both French and English, these features also show excellent classification power, and seem to represent the smallest feature group which provides the most information about nasality.

Although there may be additional features which can contribute to the perception of nasality, based on our first two experiments, these features combined constitute the most likely set of perceptual cues for nasality overall.

However, there are still some differences in how they are best used for classification in French vs. English, which will be crucial in planning our perception experiment.

5.9 Discussion: Classifying Nasality in English vs. French

As we can see from the results of our multi-feature classification experiment (see Table 31), for both algorithms and in all models, French was more easily classified as “oral” or “nasal” than English. These differences are not trivial either: for our best performing model (an SVM model including all of the features), classification accuracy was nearly 9% better in French than in English. This pattern holds universally, and is particularly striking because more training data was available for English (12 speakers vs. 8).

Thus, we must conclude that French nasality is easier for machine classification. But again, the more interesting question is “Why?”. When discussing the differences between English and French in terms of the statistical links between our features and nasality, we discussed three possible sources of the English vs. French difference in nasality.

First, there could be some subtle imbalance in the dataset which somehow expresses itself as a cross-linguistic difference in ΔFeature. This is difficult to test, and not particularly useful for explanation, and thus, will be put aside for the moment.

Our second hypothesis was that French and English speakers were actually producing different *degrees* of nasality, causing differences in ΔFeature in the two languages. Towards this point, we should recall that to ensure balanced data for our English machine learning experiment, only CVC vs. NVN comparisons were made. Thus, we are comparing only the most nasal context for nasality

in English with our French nasal vowels, and difference due to degree should be minimized as much as it can be across the two languages. It is still possible that English speakers are "holding back" in terms of VP port aperture, but there is nothing in the data or literature which specifically suggests this.

Degree of feature change between categories is, doubtless, helpful for classification. In the same way that it's far easier to separate black and white stones than it is two different shades of gray, increased ΔFeature values would improve classification accuracy, as we have seen here. If English and French nasality differed *only* by degree, although accuracy might vary, we would expect that the models would consider the features to be similarly important in both languages. But, they do not.

In Table 34, reproduced below, we see that our RandomForest models arrived at the accuracy they did in fundamentally different ways for each language:

Table 34: Top 10 Features by importance in a Reduced Redundancy model for English and French

	en.TopFeatures	en.imp	Perc	fr.TopFeatures	fr.imp	Perc.1
1	Width_F1.z	345.451	12.968	A3P0.z	229.867	16.865
2	A1P0_HighPeak.z	311.005	11.675	A1P0_HighPeak.z	203.614	14.939
3	Duration.z	273.031	10.249	Width_F1.z	140.347	10.297
4	A3P0.z	200.558	7.529	Amp_F2.z	117.135	8.594
5	A1P1.z	194.796	7.312	Freq_F2.z	99.098	7.271
6	Freq_F1.z	181.496	6.813	Freq_F3.z	95.919	7.038
7	Width_F3.z	172.531	6.477	Width_F3.z	82.292	6.038
8	P0Prominence.z	163.434	6.135	Duration.z	79.967	5.867
9	Freq_F2.z	158.767	5.960	Width_F2.z	79.783	5.854
10	Freq_F3.z	157.541	5.914	P0Prominence.z	62.744	4.604

Whereas in English, F1's Bandwidth, A1-P0 and Vowel Duration were the most important features, in French, A1-P0, A3-P0, and Bandwidth were most important for classification, with Duration far below. Put differently, speakers of French and English seem to perform nasality in a different enough way that a computer model whose sole agenda is accurate classification "chose" to approach the problem using not just two different sets of criteria, but two different feature rankings.

We can test for the hypothesis that French and English nasality is different in nature, not just degree, in another way. Machine learning models are built from training data. In all previous results, we have used the same dataset for both training and testing (via 10-fold cross-validation). However, this is not a requirement, and one can test a model on *any* set of data, so long as the same features are used.

If these two languages are doing nasality similarly *except* in degree, then we would expect an English model to show *greater* accuracy when used to classify French (where the ΔFeature values are higher), and we would expect a French model to perform more poorly on English (where the ΔFeature values are lower).

Table 35 shows the accuracy results of training all-inclusive SVM models on our English and French data, and then using both models to classify the both English and French data[27]:

Table 35: Accuracy of an all-inclusive SVM model in same-language and cross-language testing

	english.test.acc	french.test.acc
Model trained on English	84.764	72.901
Model trained on French	65.499	93.729

We had predicted that, if the difference were entirely a matter of degree, models would perform uniformly, with English models showing an improvement in French (where ΔFeature is higher). This is clearly not the case.

Instead, we see that the English-trained model shows decreased accuracy in French, and that a model trained on French performs very poorly on English data, far more so than a simple reduction in degree could account for[28].

Thus, we see that not only is French easier to classify than English, but that French is *different* to classify than English.

This, in turn, supports our third possible explanation for the differences: that French and English speakers are producing a similar *degree* of nasality, but that French speakers are producing nasality in a different way. It would appear that French speakers are producing nasality which is, at least in terms of our features, easier to detect and classify. If we accept that classification is at all functionally analogous to human perception, this increased ease of classification in French would likely correspond to greater perceptibility.

This does make sense because, as mentioned in 4.8, nasality carries a much higher functional load in French than in English, and thus, increased perceptibility may be important enough to merit extra articulatory effort or precision.

So, we again see evidence that although nasality is expressed using the same five acoustical features, English and French differ in terms of the degree and relative importance of these features. And, through comparison of classification results, we can hypothesize that because of these differences, English and French speakers would likely show differences not just in production, but in the perception of nasality as well.

5.10 Discussion: Machine Learning and Nasality

Machine Learning has helped us to understand three crucial aspects of this problem.

[27] The same-language classification accuracy figures come from 10-fold cross-validation, whereas cross-language figures are trained and tested on two entirely different datasets.

[28] As an aside, this finding also indicates that we have *not* trained our models based on speaker idiosyncrasies in each language group. Were it the case that each model was developed not based on language-specific patterns of nasality, but on the specific interactions and degrees of each feature *used by each individual speaker*, we would expect near-chance performance here, as the speaker idiosyncrasies it would rely on would be absent in the other language. Instead, each model is able to classify the other language to some meaningful degree, albeit with a severe penalty for the reasons discussed above.

First, we now have a far better idea what aspects of the speech signal are not just correlated with nasality, but *practically useful identifying it*. We have seen that although no feature alone will yield particularly promising classification accuracies, some relatively small groupings of features provide sufficient information for accurate classification. In addition, we now know *which* grouping of features is most accurate, and have an understanding of how each element is useful. This is invaluable information, useful for establishing a hypothesis about which features humans are using to help us prepare for the human perception experiment in Section 6.

Second, we see that from a classification perspective, English and French are doing "nasality" differently. The classification models for English perform relatively more poorly than those for French, and models for both rely on fundamentally different features. Coupled with similar results from our statistical studies, it seems very likely that the acoustical realization of nasality is language-specific, and that there is no one-size-fits-all solution for measuring or detecting nasality.

Finally, and most importantly, we see that the classification of nasality from these acoustical features alone is a tractable problem. This may seem obvious at first, but a great deal in speech and speech perception depends on top-down processing, context, and nuanced analysis of the whole signal. One could easily picture a world in which the perception of nasality makes reference to some features, but where the lion's share of perceptual "work" is done by guesswork ("Is it more likely that this loud Bostonian yelled "Go Pats!" or "Go Pants!" just a moment ago?") and top-down processing based on the signal and the greater communicative context. However, this no longer appears to be the case.

Although context and top-down processing likely do play a role, the machine has no understanding of "word", "context", "vowel", or even "nasality". It has no expectations, it has no "understanding", and it has no "top" to process down from. It only has numbers representing features, given to it by an imperfect feature extraction script. Yet, these computer models can recognize nasality, in isolation, with 85-95% accuracy. So, if we assume humans to be at least as skilled at the perception of nasality as machines, we know that the premise of this work, that nasality is carried on a subset of acoustical features, is tenable.

This gives us all the more hope as we move towards the final part of this task: to test the role of these features in the perception of living, breathing humans.

6 Testing Human Perception of Nasality

At this point, we have a firm understanding of the features present in produced nasality (Section 4), and of their relative utility for machine classification of nasality (Section 5). On top of this, we have a finalized set of features which seem particularly closely linked to nasality in terms of correlation and classification, along with some understanding of how they differ in French and English.

The next step towards improving our understanding of the perception of nasality is to test the perception of these features by actual humans.

Due to practical and temporal constraints, mostly concerning the availability of speakers, we will only be testing the English language here. However, the results of the English study should shed some light on the perception of nasality overall, and, when compared with the machine learning models (as will occur in section 7), these results should allow the formation of very specific and testable hypotheses about the perceptual acoustics of French nasality.

6.1 Experiment 3: Structure and Goals

The goal of this experiment is to bring all of the knowledge gained thus far back to our true interest, the human perception of linguistically useful differences, and to gain a better understanding of which of these promising features contribute to the perceptual phenomenon we call "nasality".

There are three *a priori* possibilities for how a feature can contribute to the perception of a phonological (or phonetic) phenomenon.

First, there could be features which *must* be present in order for a sound to be considered nasal: that is, any vowel which lacks these features will *not* be interpreted by listeners as nasal, no matter what else is present in the signal. These features will be referred to here as being *necessary* for the perception of nasality in a given token.

There may be some features whose presence alone is enough to trigger a functional percept of nasality, even in an otherwise oral vowel. These features, here termed *sufficient* for the perception of nasality, can be viewed as primary cues to nasality which serve not just as an indication of nasality (alongside other features), but serve to "prove" nasality by their presence alone (much like smoke and heat are indicators of fire, but only visible flame proves its presence).

Finally, it could be that no feature alone is *necessary* or *sufficient* for the perception of nasality. Instead, the perception comes from the sum total of many different features, and thus, only the presence of several of these various features forms the necessary and sufficient evidence for the perception of nasality.

To identify which of these possibilities applies to the features under consideration here, two different tactics will be used.

First, to identify features which are *sufficient* for the perception of nasality, we will take a word containing a vowel originally produced as oral, and modify it such that it now has a particular

feature associated with nasality. Those features whose inclusion is enough to change an oral vowel to 'nasal' for our listeners will be considered *sufficient* for the perception of nasality.

To find *necessary* features, we will take the opposite tack. We will modify natural tokens of nasalized vowels in each language to *reduce* that feature to the level found in oral vowels, and then present the stimuli to listeners. We will take as *necessary* for nasality perception those features whose reduction causes listeners to classify the originally nasal vowels as "oral".

In this process, we may come to find that no single feature is necessary nor sufficient for the perception of nasality. To address this, we will include some stimuli where our entire set of features has been added or reduced. In these cases, we will hope that the sum total of these modifications will be necessary and/or sufficient for nasal perception, and thus, that we will still gain valuable information about the perception of nasality.

Rather than setting up experiments for each type of stimulus modification, these stimuli will be tested simultaneously within a greater, unified experiment. These larger experiments, described in detail below, will involve the randomized presentation of feature reduction stimuli and feature addition stimuli (both groups and single features), alongside an equivalent number of control stimuli, all blocked together for convenience, but meant to be analyzed separately.

We must acknowledge the complexity of the experiment proposed. With five different manipulations of features (A1-P0, Duration, Formants, A3-P0, and all features modified ("AllMod")), each in oral-to-nasal feature addition and nasal-to-oral feature reduction versions, there are 10 different conditions, plus an equal number of control tokens, modified and then de-modified to provide similar artifacting, but with no feature change. Given this complexity, it will simply not be possible to test nasality in a large number of contexts while still collecting a meaningful amount of data about each condition.

In order to constrain the experiment, both for efficiency and for interpretability, we will limit the experiment to a just two vowels, and two speakers, within the English language. The choice of speaker and vowel will be described in depth in Section 6.3.1.

6.1.1 Experiment 3: Forced Choice Word Restoration

For English listeners, where nasality is not contrastive, "nasal" vs. "oral" is not a distinction which is consciously accessible to listeners. Thus, we must tread very carefully when evaluating speaker response to changed tokens.

One danger in testing this is accidentally designing an audio recognition task, in which listeners are simply saying that "That sounds like that other clip of a nasal vowel". Although this is a source of data, and an interesting perceptual question of itself, it is not the answer to the question being asked. Here, we are examining the linguistic phenomenon of nasality *as utilized by speakers in processing language*.

As such, rather than proposing a discrimination task or a priming task (both based on acoustical similarity), we will focus on the ability of coarticulatory nasality to provide information about surrounding segments (c.f Beddor et al. (2013),Scarborough et al. (2011), and others described in Section 2.1.3), and examine our stimuli in the context of a word restoration task.

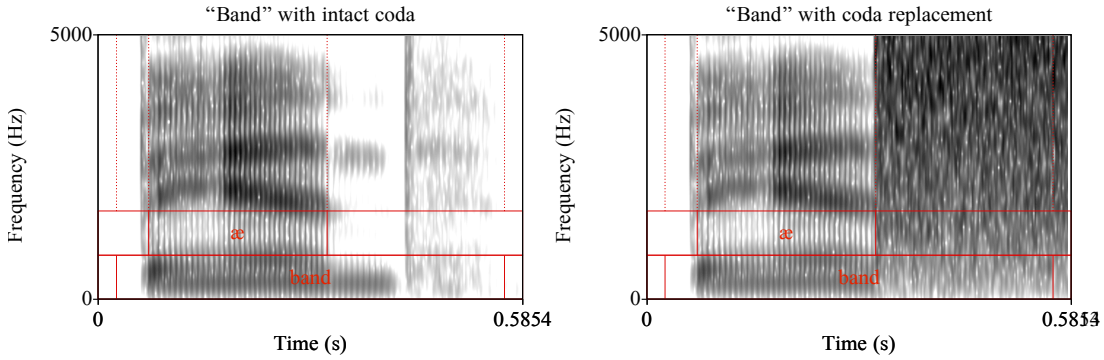


Figure 5: "Band" before and after noise replacement

In these experiments, listeners will hear whole words whose vowels have been modified according to the parameters of the experiment (with features added, reduced, or unmodified). In addition to the modification of the nasalized vowel, we will also remove either the onset or the coda of the word (depending on the phonological structure of the word), replacing it with wide-band gaussian ("white") noise. Listeners, upon hearing a stimulus, will be asked to identify which word they heard, choosing from a pair of candidates which share a place of articulation in the removed consonant, but where only one choice has a nasal consonant.

For example, if we were attempting to introduce confusion in the word "band", we would modify the vowel, then replace the coda with noise (leaving /bæ_/, where the underscore is noise). This is visualized in Figure 5. Listeners, when presented with this sound, would be given the choice between "bad" and "band", and asked to indicate, as quickly as possible, the word they heard. For NVC vowels, the onset would be replaced with set-duration noise, rather than the coda. Reaction time and response would be recorded.

This is a somewhat unusual, and quite possibly difficult task. However, this difficulty is desirable in the context of this study. The changes being made to the sounds will, in many cases, be quite subtle, and an easy task might not require careful enough evaluation for the details being modified to be noted or used in perception. Also, this task is not without precedent, as a similar paradigm has been shown to work well for eliciting nasality judgements in prior work. Both Lahiri and Marslen-Wilson (1991)), and Beddor et al. (2013) showed that American English-speaking listeners do use coarticulatory nasality for lexical disambiguation, even during the time course of the vowel.

So, even though English speakers cannot directly provide categorical judgements of vowel nasality, the use of a phoneme restoration task allows us to nevertheless straightforwardly gauge listeners' perception of "nasal" vs. "oral".

6.2 Hypotheses for Human Perception

Fundamentally, our hypothesis is that the features being tested will affect the perception of nasality. This can be evaluated by testing three predictions:

Hypothesis 1 – *When features are added or reduced, there will be an increase in RT (indicating increased difficulty) relative to the control stimuli.*

Hypothesis 2 – *When features are added or reduced, there will be increased oral-nasal confusion (shown as decreased accuracy) relative to the control stimuli.*

Hypothesis 3 – *Different features will show different changes in accuracy and reaction time, according to their differing degrees of perceptual usefulness.*

In addition, it is somewhat difficult to imagine a feature whose presence strongly signals nasality, but whose absence plays no role in identifying nasals, or vice versa, so we might predict that:

Hypothesis 4 – *Those features which are necessary for nasal perception (whose reduction in nasal vowels reduces perceived nasality) will also be sufficient for nasality perception (their addition to oral vowels will also create the percept of nasality).*

We might also imagine that adding or reducing many features will have a stronger effect than adding or reducing any one. Given this, we predict that:

Hypothesis 5 – *Reducing or adding all of the features will have a greater effect on classification and reaction time than any of the single feature manipulations.*

Finally, although we will be reducing the most promising features in our reduction stimuli, some features pointing towards "nasal" will likely remain. We might then assume that adding *any* nasal features to an oral vowel would prompt a stronger reaction and change in perception than reducing *some* nasal features in an already nasal vowel. So, we can predict that:

Hypothesis 6 – *Adding features to oral sounds will have a greater effect on the perception of nasality than reducing features in nasal sounds.*

It is possible that these additional, unmodified features may alone be sufficient for the perception of nasality, and thus, that modifications to the nasal vowels will have no effect on accuracy. However, given the strength of the tested features in machine learning, this does not seem particularly likely.

6.3 Methods: Stimulus Creation and Feature Modification

For this experiment, the source files will be chosen from the same recordings collected earlier for measurement. As mentioned previously, this experiment will have four different types of stimuli: Feature Reduction, Feature Addition, and Oral and Nasal Control.

Feature Reduction stimuli aim to evaluate which of the features of nasality are necessary in order for a nasal sound to be considered nasal by listeners. These will be created by starting with nasal vowels and, through signal processing, changing the A1-P0, A3-P0, Vowel Formants and Duration to the values found in oral vowels. In addition, we will also test groups of stimuli in which *all four* of the above nasal features are reduced to oral levels.

Feature Addition Stimuli, on the other hand, will change these four features in naturally oral vowel, giving them the values found in nasal vowels, in order to test which features, if any, are

sufficient for the percept of nasality. And, again, we will create some stimuli where all four features are modified to nasal levels.

Oral and Nasal Control Stimuli are our final stimulus type. Although one could simply compare performance for trials featuring the reduction and addition stimuli to performance identifying natural nasal and oral vowels, the comparison would not be direct, as we would be comparing modified and unmodified stimuli, and several of our modifications do introduce audible levels of acoustical artifacting. As such, reaction time differences could be either attributed to the feature changes being tested, or to unnaturalness or artifacts from to the modifications.

Although this could be overcome with an elaborate series of naturalness evaluations, an easier approach is to compare feature-modified stimuli with stimuli which are similarly unnatural, but where the changes have been made using the same signal processing techniques as above, *and then immediately reversed*. This should result in stimuli which have all the artifacts found in our addition and reduction stimuli, but where the nasal features are ultimately not meaningfully modified.

Feature modification will be conducted using the same scripts for feature addition, reduction, and control stimuli, with the settings varying only in terms of the final ΔFeature.

6.3.1 Methods: Choosing Vowels and Speakers for the Stimuli

The two vowels chosen for modification and testing here are two low vowels, /æ/ and /ɑ/. Vowel height is important, both in that A1-P0 is most robust in low vowels, and because nasality is thought to be strongest in low vowels (c.f. Delvaux et al. (2008a), Shosted et al. (2012),Delvaux et al. (2008a)). In addition, although the French /a/ will be somewhat different from both vowels (being more centralized than either), using /ɑ/ will allow more straightforward comparison with French than a completely unrelated vowel quality.

For /æ/, there are a total of 12 CVC-CVN and CVC-NVC real word minimal pairs which can be used for testing[29], and we have recorded 8 such pairs for /ɑ/. Thus, each condition (e.g. “modified formants”) will be tested 20 times, coupled with the same 20 word pairs for testing with control stimuli.

One source of difficulty in any stimulus manipulation experiment is choosing the speakers for modification. Because speakers vary greatly on many dimensions, speaker idiosyncrasies often interact with the desired manipulation, particularly when looking at individual words or small groups. Thus, speakers must be chosen on the basis of their specific speech characteristics.

Although each speaker shows significant ΔFeature effects for our tested features across the entire group of tested vowels, and each vowel shows a similar ΔFeature across our entire group of tested speakers, at the by-speaker by-vowel level, sample sizes are simply too small to run meaningful models (as each speaker recorded only 8 total tokens of each vowel in each of CVC, CVN, NVC, and

[29]Given the deeply linguistic nature of the lexical choice task, it does not seem prudent to use non-words in this experiment, as listeners are unlikely to have sufficient experience with the words to identify sub-optimal tokens. Thus, the study will be limited only to real English words.

NVN contexts, resulting in 32 total tokens). So, to characterize by-speaker-by-vowel differences, our only option is to compare means for these tokens.

Out of our desire to avoid picking speakers whose idiosyncrasies would complicate the analysis or task, we have chosen for the final experiment the two speakers whose by-speaker-by-vowel mean ΔFeature values from oral to nasal vowels best approximated the overall model predictions for ΔFeature, albeit with slight variations in degree of change.

The two speakers chosen, "Hazel" and "Molly", are fairly similar in linguistic background (18-19 year old females from Colorado), and both recorded full sets of stimuli. In addition, neither had other distinguishing characteristics which would otherwise disqualify their participation.

So, in the final perception experiment, we will be testing the effects of these ΔFeature values for both /ɑ/ and /æ/, in the speech of Hazel and Molly.

6.3.2 Methods: Final ΔFeature Values

Before generating the stimuli, we must specify the ΔFeature modifications made to each word. Before we go too much further, though, we must clarify a few points.

First, in modifying these stimuli, we are modifying ΔFeature directly, rather than targeting certain measurements for the features. Put differently, we will not be setting a feature "to" a certain value (e.g. "oral sounds made nasal will be given an A1-P0 of -6"), but instead, we will be simulating nasalization, taking the existing values for A1-P0, whatever they may be, and lowering them by 5.3 dB ("an oral sound which *had* an A1-P0 of 0.3 will now have a value of -5 db").

In addition to more transparently mapping to the correlations shown in the statistical analysis, this reflects the speaker- and vowel-independent nature of the modeling, and allows the same ΔFeature to be applied to all stimuli, no matter the initial values. Given that we are testing the hypothesis that a given change to these features will result in a change of perception, this is the most effective means of testing.

There is one exception to this: Formant bandwidths will be modified to target the oral and nasal means seen overall. This is because token-by-token variation in bandwidth, as well as variation in its measurement, led to some tokens, particularly the "shrinking bandwidth" nasal-to-oral tokens, becoming over-narrow and "chirpy". Given the relative similarity of our speakers in terms of formant characteristics, and the increased simplicity of the analysis, the benefits of this approach outweigh any drawbacks.

Secondly, we are using a single set of ΔFeature values across speakers and vowels. Although different speakers do produce nasality in different degrees, or that some vowels show greater or lesser ΔFeature, we are making only one set of modifications. The effects of this variation should be minimized by the speaker and vowel similarity discussed in Section 6.3.1 above. We should also state that we have no reason, based on the data above and in the literature, to suspect that the perception of nasality is only possible using speaker-specific per-vowel information about changes to nasal features.

However, the most important reasoning for applying a single set of ΔFeature values is the fact that

we are testing a general hypothesis about what perceptually constitutes nasality. Our machine learning algorithms (which had neither "speaker" nor "vowel" as input) were able to effectively classify nasality using generalized trends, and the features identified in the statistical analysis (and used here) proved well suited to classification, across speakers and vowels. So, we will test a single set of changes, with the hypothesis that reproducing the oral → nasal ΔFeature will trigger the perception of nasality, and analyze and by-vowel or by-speaker differences in performance post-hoc.

Finally, although we're testing CVC vs. NVC, and CVC vs. CVN, ΔFeature for each category will be based on the across-speaker model for CVC → NVN changes. Although this will result in slightly less natural stimuli, "completely natural" is simply not possible for stimuli where (only) four specific spectral characteristics have been changed. Instead, we want to give these changes to nasal features every possible chance to trigger the percept of nasality (or loss thereof). Thus, we will modify nasality according to the ΔFeature found in our most nasal context, CVC → NVN, across all speakers and vowels, taking the 'nasality' coefficient in the model to represent ΔFeature, and applying no change to those features which showed no statistically significant change.

Given all this, our final ΔFeature values will be as below, in Table 36. These numbers are taken directly from Table 8, rounded to the nearest Hz and ms. Note that both modifications to F2 as well as F3's Frequency did not reach significance in English, and thus, will be unmodified in test stimuli.

Table 36: Final Oral → Nasal ΔFeature values for English

Feature	**ΔFeature**	**Unit**
A1P0_HighPeak	-5.370	dB
A3P0	-5.335	dB
Duration	-36	ms
Freq_F1	36	Hz
Freq_F2	0	Hz
Freq_F3	0	Hz

For formant bandwidths, which are modified to absolute targets, the final values for oral and nasal vowels are as below. The "oral" values will be applied to nasal-to-oral modification stimuli, and vice versa. F2's bandwidth, which did not change significantly, is not modified.

Table 37: Final Oral and Nasal values for Formant Bandwidths

Feature	**"Oral" value**	**"Nasal" value**	**Unit**
Width_F1	171	296	Hz
Width_F2	–	–	Hz
Width_F3	498	615	Hz

If we went the other direction, taking a token of "band" and reversing the ΔFeature values in an effort to prevent the perception of nasality, the features might change as below:

Feature	Oral Value		Nasalized Value
A1P0_HighPeak	2.392	→	-2.978
A3P0	-14.9	→	-20.235
Duration	217	→	181
Freq_F1	601	→	637
Width_F1	171	→	296
Freq_F2	1770	→	1770
Width_F2	421	→	421
Freq_F3	2756	→	2756
Width_F3	498	→	615

If we were creating an "All Modifications", nasalized version of a particular token of "bad" with the given initial values for all features, the features would change as below:

Feature	Nasal Value		Oralized Value
A1P0_HighPeak	-2.27	→	3.1
A3P0	-19.6	→	-14.265
Duration	182	→	218
Freq_F1	632	→	596
Width_F1	296	→	171
Freq_F2	1762	→	1762
Width_F2	438	→	438
Freq_F3	2795	→	2795
Width_F3	615	→	498

Finally, it's worth noting that these are the target ΔFeature values input to the script. Although the modification scripts have been tuned and optimized for accuracy, due to the noise inherent in script-based signal processing, each stimulus will likely vary from the targeted deltas to some extent. This variation will be explored in more depth in Section 6.3.8.

6.3.3 Modifying Duration

The modification of duration is relatively straightforward, using Pitch-Synchronous Overlap-Add (PSOLA) modification.

1. The vowel was isolated, using information the textgrid, and its duration was measured.
2. The final vowel duration was calculated using the original duration and the ΔDuration value specified in Section 6.3.2.
3. PSOLA was used to shrink or expand the duration of the sound to match the desired target…
    - This is a pitch-aware process, and works by duplicating cycles, such that the spectral properties are not changed, even while the duration is.
    - A duration-modified TextGrid file was created at this time as well.

4. The modified vowel was spliced back into the original, unmodified word at the original position.
5. A "control" stimulus was created by simply duplicating the original sound file.
    - Compared to the other modifications performed here, PSOLA is *exceptionally* clean, and tests changing and then un-changing duration by PSOLA did not introduce any noticeable artifacts.

Thus, our duration control tokens will serve double-duty as the "natural" tokens against which other conditions can be compared.

6.3.4 Modifying Formant Frequency and Bandwidth

In speech, vowel formants are caused by the resonances in the mouth, controlled chiefly by the position of the tongue. This tongue positioning (and the positioning of other articulators like the lips or tongue root) controls the frequency, as well as the bandwidth, of formants. Given that the almost all formant effects stem from a single change in tongue position, it makes sense to consider the formant structure (bandwidth and frequency) as a relatively holistic phenomenon, where one is unlikely to see changes on just one dimension (e.g. changes to F1's bandwidth with *no* change in any other formant frequency or bandwidth).

With this in mind, as well as for practical reasons, in this study, all formant frequencies and bandwidths will be modified together as a single condition, creating a class of formant modified stimuli, with nasality-related changes to the bandwidth and frequency of F1 and in the Bandwidth of F3. Although this costs us the ability to drill down a formant-related effect (e.g. to claim that "F1's bandwidth *alone* is sufficient to cause a nasal percept"), it gives us the best chance of determining whether formants *in general* are a cue for nasality, and significantly reduces the complexity of the (already complex) perceptual experiment.

To create these formant-modified stimuli, Source-Filter vowel resynthesis was used, in the following manner:

1. Each word-containing sound file was resampled to 11,025 Hz, to aid in resynthesis
    - This process is performed on the whole word, rather than the vowel only, to minimize the impact of edge-related issues on the target vowel
2. An LPC (Linear Predictive Coding) analysis was done to identify existing formants in the word, capturing the "filter" through which voicing was run.
3. The sound was "Inverse filtered" using the LPC, where all "valleys" are reversed, yielding, ideally, a smooth and un-differentiated series of harmonics, a "source" file
    - This effectively represents the voicing leaving the larynx, before it has been "filtered" by the tongue and mouth
4. The LPC-based "filter" was then modified mathematically, using ΔFeature values for bandwidth and frequency specified in Section 6.3.2.
5. The modified "filter" was applied to the "source", creating the formant-modified stimuli.

6. The un-modified, original LPC was re-applied to the "source", creating control stimuli
    - These should have all of the artifacts of the resynthesis, without any meaningful formant changes.
7. The bottom 3500Hz of the resynthesized stimuli (control and modified) were then combined with the top 19,500Hz of the unmodified word, to minimize high frequency artifacting.
8. The vowel is extracted from the final, modified word, and spliced into the original, unmodified word at the relevant position.

Through this process, words are created which have the desired formant-related characteristics, while still retaining the speaker-specific formant structures and context of the original word.

6.3.5 Modifying A3-P0 (Spectral Tilt)

Although we have chosen A3-P0 as a useful feature of nasality, we have no evidence that nasality is particularly affecting the amplitude of the third formant. Instead, based on the decreasing amplitude of harmonics throughout the frequency spectrum (as described in Section 4.5.3), it appears that ΔA3-P0 is representing a change in spectral tilt expressed through the entire spectrum. So, although modifying A3 and P0 specifically would cause the desired ΔA3P0, it would be rather missing the point, and would render the output fragile and artificial.

So, rather than modifying specific peaks, we will modify the overall spectral tilt of the vowel, to a point specified by measurement of ΔA3P0. This is accomplished by following the below procedure:

1. The vowel was isolated using the TextGrid annotations, and the base A3-P0 was measured.
    - A3-P0 was measured by finding the amplitude of highest of the first two harmonics, and the amplitude of the highest peak +/- F_0 from F3_Freq *for each individual vowel.*
2. The vowel was filtered using Praat's "Filter (Formula)" function.
    - For increasing tilt, we use the formula:

      `self /(1+(x/100))^0.01}`

        - "The new amplitude is equal to the original amplitude (`self`) divided by (1 + ([frequency]/100))^0.01"
        - Put differently, as the frequency increases, the amplitude at that frequency is divided by increasingly large numbers (1.007 at 100 Hz, 1.024 at 1000 Hz), and thus, as frequency rises, amplitude drops.
    - For decreasing tilt, we use

      `self *(1+(x/100))^0.01}`

        - The only difference here is that the amplitude is *multiplied* by increasingly large numbers, thus, *increasing* amplitude as frequency increases.
3. A3-P0 was measured in the resulting sound, and Observed ΔA3P0 was calculated.

4. Steps 2 and 3 are repeated (via a `while` loop) until the observed ΔA3P0 reached the target ΔA3P0 value specified in Section 6.3.2.
    - This incremental approach is needed because one cannot reliably target a ΔA3P0 in dB in a single filtering pass, given the Pascal vs. dB conversion. It also mirrors the measurement process used to find the initial value, slightly reducing re-measurement error.
5. Once the target ΔA3P0 was reached, the vowel was spliced back into the original word context.
6. Control stimuli were generated by using the above process to reverse ΔA3P0, bringing the spectral tilt back to the originally-measured value.

6.3.6 Modifying A1-P0

The process for modifying A1-P0 is simple in concept. We locate the regions of the spectrum which correspond to A1 and P0, extract them, modify their amplitudes, and recombine them into a single word.

There are some complexities in implementation, though[30]. Thus, we modified each vowel as follows:

1. The vowel was isolated using the TextGrid annotations.
2. The vowel was split into 3 temporal 'bins', each [vowel_length/3] milliseconds long.
    - By treating each bin as an independent modification, we can modify using the H1, H2, and F1 values *at the start of each bin*, rather than using the same, static frequencies for these features throughout the duration of the vowel. Thus, this binning allows more accurate estimation and modification of H1, H2, and F1 even when they change throughout the word.
3. For each bin:
    a. The script collected the original A1-P0 (and its components), using the same code as the measurement script
    b. The script band-filtered the vowel by frequency to isolate the different components (the first formant, the nasal peak, and the rest of the word)
    c. The amplitudes of each isolated frequency band (e.g. increasing P0 and decreasing A1) were then modified by ΔA1-P0 dB, as specified in Section 6.3.2.
        - See the note below regarding the proportion of change to each.
    d. The bands were then recombined into a single modified bin
    e. The measurement process was repeated, to ensure that there were no troubles.

[30]The script which does this was developed over the course of several years by the author and Dr. Rebecca Scarborough.

- If ΔA1-P0 is not within an acceptable range (+/- 0.5 dB, here), the script was programmed to automatically return to (b) and try again with a slightly modified target, incorporating any overshoot or undershoot. This process is repeated until an accurate modification is made is made.

f. For control stimuli, the process was completed with a targeted ΔA1-P0 of 0 dB.

4. The bins were stitched back together, creating a final modified vowel
5. The vowel was spliced back into the original word context.

One additional note: When modifying A1-P0, one actually modifies two different spectral properties: the amplitude of F1, and the amplitude of P0. Across all speakers, when moving from oral to nasal in our CVC-NVN dataset, F1's Amplitude went down by 3.4, and P0's amplitude went up by 1.9, for a total ΔA1-P0 of 5.3. Put differently, the reduction of A1 accounts for 64% of the 5.3 dB of change in A1-P0 from CVC → NVN.

This script has been set up to model this percentage of feature change[31]: To lower A1-P0 by 5.3 dB, A1 will drop by 3.4, and P0 will rise by 1.9, and vice versa when raising A1-P0.

6.3.7 Creating the "All modifications" stimuli

The 'all modifications' ("All-Mod") stimuli were created using the same techniques described above, with two modifications.

First, the order of modification used was Duration → Formants → Spectral Tilt → A1-P0. This is largely for practical reasons (allowing re-use of measurements across modifications). The first three modifications happen in a single script, with A1-P0 modification occurring in a standalone script, due to its complexity.

Second, the all-mod control stimuli have undergone *all* modifications (with reversal) in this process. That is to say, the control stimulus is formant-modified and de-modified, then tilted and un-titled, then A1-P0 modified at 0 dB ΔA1-P0. This means that these should be the most heavily artifacted of the stimuli, by design, and can be used to evaluate the role of artifacting independently of ΔFeature.

6.3.8 Methods: Checking the Stimuli

The utility of a perception experiment is limited by the accuracy of its stimuli, and thus, considerable caution went into stimulus creation, both in the process and post-hoc.

To get an idea of the final accuracy of the ΔFeature modifications, following stimulus generation, the final (pre-noise) stimuli were re-measured using the same measurement script and process

[31]There is actually no reason to believe that the composition of ΔA1-P0 should matter to listeners, as the listener (and indeed, the author) *cannot* know what Amp_F1 or Amp_P0 *would have been* in an oral vowel produced in that particular token. Although ΔA1-P0 can be understood by comparison with other words, ΔAmp_F1 or ΔAmp_P0 exist only across two hypothetical versions of the same pronunciation. This model of the composition of ΔA1-P0 is presented only because *some* model is required, and this seemed the most principled approach.

described in Section 3.5, albeit with four measurement points per vowel (to more accurately capture any issues resulting from temporal binning).

Before we analyze the data, a note: We cannot expect these stimuli to be perfectly clean, nor for the actual ΔFeature values to match exactly the desired deltas. There are four principal sources of error in the modification process.

First, some of our changes interact. For instance, modification to A3-P0 (spectral tilt) will necessarily affect A1, and thus, result in some change to A1-P0 and to vowel formant bandwidths. Similarly, modifying formant bandwidths will often increase or decrease the height of P0. All of the effects described here are working in harmony (e.g. "more nasal" changes to A3-P0 will move A1-P0 in its "more nasal" direction), so this interaction will result in some change to features where none was intended (e.g A1-P0 changes when formants do), and will result in some overshoot when all features are combined (e.g. ΔA1-P0 will be higher than intended because of the sympathetic effect of ΔA3-P0)[32].

Second, there is error inherent in the modification. These modification processes, although tested and optimized extensively, are very complex, and are subject to the limitations of the programmer and of the scripting language[33]. In some cases, particularly A1-P0 and spectral tilt modification, modification is not deterministic. We cannot simply enter a value and complete a command, only to find the perfect ΔFeature, but instead, a series of small modifications must be made until an acceptable value is reached. This means that these processes must *specify* an error margin, and thus, that ΔFeature will be normally distributed around the exact goal. Thus, there is some noise inherent in the stimulus generation process, which will manifest as true inaccuracy in ΔFeature[34].

Third, there is measurement error in the stimulus generation process. Individual stimuli may be difficult to measure for any of the various characteristics (often pitch and formants), and aperiodicities may lead to some difficulties in generating spectra. Although the scripts have been hardened against these issues to the best of the author's ability (in many cases using the same code as in the measurement script), these issues can make it very difficult to accurately change features, as accurate change depends on accurate knowledge of the initial state. In these cases, the accuracy of ΔFeature is actually affected, and the stimuli miss the target.

Finally, there is *re*-measurement error. Although duration is independently measured, all of our other measures depend on the detection of particular landmarks in the spectral landscape. This means that two different measurement passes can disagree if, for instance, the re-measurement uses a different peak as "P0", or finds a different harmonic for "A1". Each of the modification processes described above consistently uses the same landmarks throughout the entire process (e.g. Freq_F1 or P0's harmonic is not allowed to change between steps). This means that the modification-script-internal measurements are consistent, and that ΔFeature is always calculated

[32]If degree of nasality beyond the natural NVN levels is important to listeners, the AllMod stimuli may show a strength of effect above and beyond what the compositional features imply

[33]This is not to disparage Praat in any way. Despite occasional difficulties, the fact that such complex modifications and measurements can be done automatically and with reasonable accuracy using only a scripting language is an incredible testament to the work of Drs. Boersma and Weenink, to whom the author owes a great debt

[34]Frustratingly, this inaccuracy is usually deterministic, that is, every run of a given stimulus will result in the same inaccurate outcome due to some idiosyncratic property of that token. Thus, re-running a stimulus to obtain a better modification is often not possible.

against the same peaks as for the original measurement. If the initial measurements are correct, then ΔFeature should be calculated and modified correctly based on the proper peaks, *even if a subsequent, independent measurement chooses different peaks*. Thus, there may be stimuli which are accurately modified, but which do not seem to be so based on a second, erroneous measurement.

With all that said, following measurement, the stimuli were divided into their conditions (A1-P0, A3-P0, Formants, Duration, and AllMod), and separated into experimental and control stimuli. Values for each condition (control and experimental) were compared to means for the completely unmodified tokens, and then this difference was compared to the intended ΔFeature (e.g. only A1-P0 and Allmod experimental stimulus should show variation in A1-P0, all the others should show 0 ΔA1-P0). Oral-to-Nasal and Nasal-to-Oral stimuli were evaluated separately, but in the same manner.

Through this process, we can calculate, across the entire stimulus set, the mean error relative to the intended ΔFeature for each feature[35]. For bandwidth, the final bandwidth was compared to the targets in Table 37. These means are given below in Table 38, alongside the intended ΔFeature to show the magnitude of the errors:

Table 38: Mean error in ΔFeature in the final stimulus set

Feature	ΔFeature	Average error	Unit
A1P0_HighPeak	-5.370	-0.306	dB
A3P0	-5.335	-0.434	dB
Duration	-36	0.88	ms
Freq_F1	36	17	Hz
Width_F1	–	-1.9	Hz
Width_F3	–	-105	Hz

We can see, then, that Duration modification was particularly accurate, and the formant bandwidth, particularly for F3, was somewhat difficult to manipulate accurately (although we should keep in mind that F3's average bandwidth is in the 400-500Hz range). In no case is the average error greater than the intended delta, and even with of all the possible sources of noise, these stimuli in the aggregate appear to be sufficiently accurate to allow us to proceed.

6.3.9 Methods: Disappearing onsets and codas

Finally, the stimuli were run through a script which replaced either the pre-vocalic or post-vocalic part of the finalized stimuli with 250ms of gaussian noise. Vowel boundaries were found using the human-checked TextGrid annotation files used in all previous steps, and the duration of the noise was fixed, such that any word, no matter the complexity of the replaced segment, will start or end with 250 ms of noise.

[35]This process was repeated several times, with tweaks to the modification process and new stimuli generated, to aid the author in improving different components of the modification process. The values reported here represent only the final stimuli presented to listeners.

To compensate for slight variations in the position of the vowel boundary (and to avoid giving away information about the pending segment), the script was programmed to start or end the noise 20 ms into the vowel (that is, onset noise ends 20 ms after the labeled boundary, and coda noise starts 20 ms before the end of the vowel). Because this is a fixed amount added to all stimuli, it will not mask the duration differences introduced above.

The set duration of noise, and the fact that it completely replaces the segment(s) in question should ensure that no information at all about the nature of the missing segment's identity survives the process, save the information about the vowel.

For longer words (such as in the bod-bonfire pairing), the entire final portion of the word is removed, leaving only /bɑ_/ for testing.

In addition, this script normalized the amplitude, setting all tokens to the same fixed, "70dB" overall amplitude level[36].

The .wav files output from this final script served, without further modification, as the stimuli for the perception experiment.

6.3.10 Notes on the Perception Stimuli

Due to the complexity of the A1-P0 modification task (which relies on accurate pitch, formant, and spectral measurements, each vulnerable to signal idiosyncrasies), approximately 1-3% of attempted A1-P0 modifications were not able to reach the required accuracy threshold. Anticipating this, each speaker recorded two tokens of each word, and in all cases but one, when the first token failed for a particular modification, the complete set of modifications for the second recording were substituted (this occured for only five words). Given that repetition proved to have no meaningful effect in the analysis, and that listeners will still hear only one of the two recordings of a given word, this should cause no problems.

One irregularity did arise: during recording, although her second production was produced as expected, Hazel produced the first recorded token of "Dan" as "Dane" (/dejn/), which would be unusable in a comparsion task. Because of this, no "fallback" token was available when the All-mod stimuli failed at A1-P0 modification for the control and experimental tokens. To balance the stimulus set for this particular word and modification type ("Dan" vs. "Dad" Nasal-to-oral All-Mod), Molly's "Dan" all-mod control and experimental tokens were presented twice during her block, resulting in 198 "Hazel" stimuli, and 202 "Molly" stimuli.

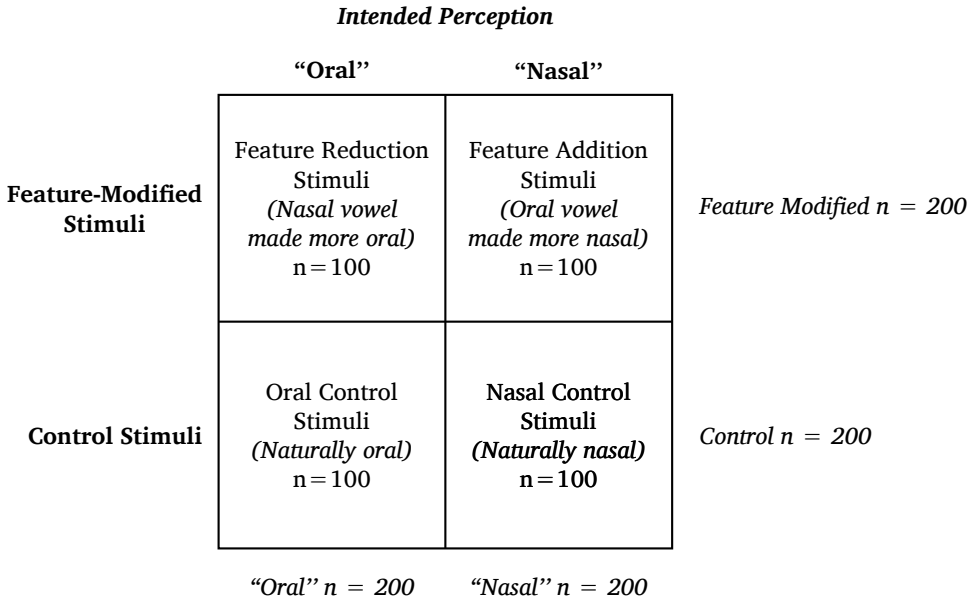
All other stimuli were processed without additional complications.

6.4 Methods: Experimental Design and Balance

For each of our five conditions (change A1-P0, change A3-P0, change duration, change formants, and change all features), we will create twenty of each of four types of stimuli: feature addition,

[36]This value represent's Praat's digital representation of the amplitude. Stimuli were played back through uncalibrated headphones at a fixed level, and thus, presentation amplitude is not known.

Figure 6: Experimental Design and Balance of Stimuli



feature reduction, oral control, and nasal control. Listeners will then hear each of these stimuli in the context of a lexical choice task with phoneme masking, and their relative performance in each condition will be compared to gain insight into the effects of each feature on nasality.

6.4.1 Experimental Balance

Because we do not want to have oral stimuli more common than nasal (or vice versa) which might bias perception in a forced choice task, all stimuli will be presented together in random order, balancing each other out as shown in Figure 6.

There is, of course, a good chance that some of the feature-modified stimuli will not be perceived as crossing the functional boundary between oral and nasal. However, this does not affect the balance significantly, as it's likely that the same features whose addition fails to make an oral vowel sound nasal will have no perceptual effect when removed from a nasal vowel, thus, in effect, balancing themselves.

In addition, by including the modified-but-unchanged control stimuli, which should be more straightforwardly oral or nasal, the listener is not set adrift on a sea of potentially ambiguous modified-feature stimuli, and has some equally probable clear choices, even among the manipulations.

So, by presenting all of the trials in a single randomized set, we can avoid using useless filler stimuli while still keeping the task engaging and unpredictable for participants. And, because this design includes all stimulus types in a single experiment, each listener completing the study will hear the same stimuli, and will complete the same experimental process.

6.4.2 Experimental Flow

Once seated in the sound-isolating booth and screened for hearing difficulties, listeners scrolled through a series of prompts explaining the experiment, which they scrolled through using two button-box buttons, green on the right, yellow on the left.

> You will hear a word. Using your index finger, indicate which of the words you think you heard. Please respond as quickly as possible. Press any button to continue.
>
> Even if you are not sure which word you heard or if some of them are particularly hard, please make a choice as quickly as possible.
>
> The words will be partially overlaid with noise, as if from a bad cell phone connection. Guess which word you heard.
>
> If you heard the word on the right, answer "green", if you heard the word on the left, answer "yellow"
>
> You will complete a small set of training examples first.

At this point, listeners heard 10 stimuli, modified as in the trials, but from a different speaker ("greta"). These stimuli represented the sort of artifacting present in the production stimuli, and included three of the less usual words ("bonfire", "nad", and "bod") for familiarization. After the training, listeners were given an opportunity to ask any questions, and then allowed to continue into the experiment at will.

Listeners were then presented with test stimuli, proceeding trial-by-trial as described in Section 6.4.3). To encourage comfort and attention, the stimuli were split into 8 blocks of 50.

As the stimuli came from two speakers, the speakers were blocked independently, to avoid ongoing speaker normalization issues. Participants first heard 200 words from one speaker, then heard 200 words from the other speaker in 4 additional blocks. In between speakers, listeners were given a rest, with the message:

> Switching speakers now. Please keep answering as quickly and accurately as possible. Press any button to continue.

Speaker ordering (e.g. "Molly's trials before Hazel" or vice versa) was determined randomly for each participant.

6.4.3 Trial Design

Each trial featured the playback of a single word coupled with the on-screen display of two words (the word and its "pair") in randomized order. Table 39 gives a small number of the 400 overall trials, showing the trials for one speaker for the formant modification condition.

We see here that there are 20 words used per condition, each providing 10 oral-made-nasal addition stimuli and 10 nasal-made-oral reduction stimuli for each condition, and that the same set of words and pairs repeats for the control stimuli. Then, the cycle repeats for stimuli with modified spectral tilt (and their control trials), and so forth, until each condition has been tested 10 times.

Table 39: Example trials for the experiment

Word	Pair	Contrast	IsNasal?	Missing	Vowel	Modification	IsControl?
dad	dan	CVN - CVC	no	coda	æ	nasalized_formants	no
dan	dad	CVN - CVC	yes	coda	æ	oral_formants	no
dad	nad	NVC - CVC	no	onset	æ	nasalized_formants	no
nad	dad	NVC - CVC	yes	onset	æ	oral_formants	no
bad	ban	CVN - CVC	no	coda	æ	nasalized_formants	no
ban	bad	CVN - CVC	yes	coda	æ	oral_formants	no
bad	mad	NVC - CVC	no	onset	æ	nasalized_formants	no
mad	bad	NVC - CVC	yes	onset	æ	oral_formants	no
dab	dam	CVN - CVC	no	coda	æ	nasalized_formants	no
dam	dab	CVN - CVC	yes	coda	æ	oral_formants	no
dab	nab	NVC - CVC	no	onset	æ	nasalized_formants	no
nab	dab	NVC - CVC	yes	onset	æ	oral_formants	no
bob	bomb	CVN - CVC	no	coda	ɑ	nasalized_formants	no
bomb	bob	CVN - CVC	yes	coda	ɑ	oral_formants	no
bob	mob	NVC - CVC	no	onset	ɑ	nasalized_formants	no
mob	bob	NVC - CVC	yes	onset	ɑ	oral_formants	no
bod	bonfire	CVN - CVC	no	coda	ɑ	nasalized_formants	no
bonfire	bod	CVN - CVC	yes	coda	ɑ	oral_formants	no
bod	mod	NVC - CVC	no	onset	ɑ	nasalized_formants	no
mod	bod	NVC - CVC	yes	onset	ɑ	oral_formants	no
dad	dan	CVN - CVC	no	coda	æ	nas_formants_con	yes
dan	dad	CVN - CVC	yes	coda	æ	oral_formants_con	yes
dad	nad	NVC - CVC	no	onset	æ	nas_formants_con	yes
…	…	…	…	…	…	…	…

Then, in the 5th-8th blocks, the entire list repeats, although again in randomized order, using words from the second speaker.

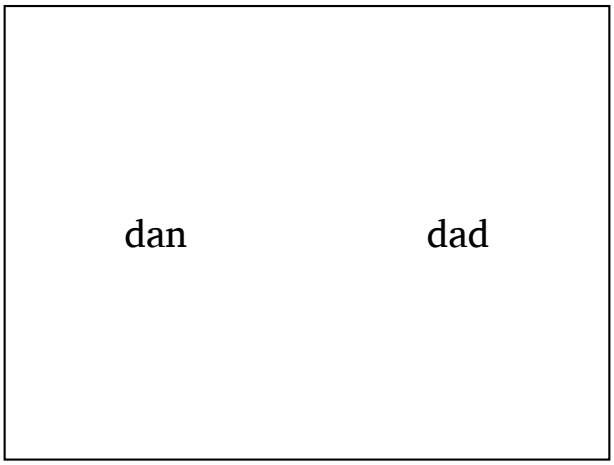
We see also that every condition is tested with every word. Thus, even if the word “bonfire” were substantially harder to identify than other words (for example), *all 10 conditions* would experience the same slowdown and loss of accuracy from the same token. This should help to balance out by-token variation, and will ensure that across-condition comparisons are as straightforward as possible.

The full list (with added filenames) forms the input to the PsychoPy experiment, and the script grabs trias from this file *completely at random*, playing the word, displaying the pair, and capturing reaction time and accuracy, as shown in Figure 7 below.

This trial randomization means that although all listeners heard the same trials, they heard them in different orders, and thus, effects of presentation order should be minimized.

Stimulus laterality was randomized as well, such that for any given participant and trial, the words could be displayed in either ordering (e.g. “Dad” on the left, “Dan” on the right, or vice versa).

Figure 7: In-experiment view of the first trial from Table 39



6.5 Methods: Running the Experiment

This experiment was conducted according to CU IRB Protocol 13-0668.

Participants were recruited from the CU Linguistics subject pool. Participants were given a consent form, and asked to provide (optionally) a gender, age, home state or country, and any additional languages spoken fluently.

In addition, in order to ensure consistency in perception data, a hearing screening was administered to each participant by the experimenter, using a Maico Hearing Instruments MA-19 portable audiometer[37]. Participants were tested using pure tones at 25 dB for 500Hz and at 20 dB at 1000, 2000, 4000, and 8000 Hz. Participants unable to hear one or more of the tones in either ear were disqualified from further participation. This disqualified a total of five potential participants.

All testing occurred at the University of Colorado at Boulder Linguistics Department's Phonetics lab, in a sound-attenuated booth. Stimuli were played back using Audio Technica ATH-M40FS headphones, at a fixed volume. The experiment was carried out using the PsychoPy Experimental Suite (as described in Peirce (2007)), using an ioLabs USB response box for collecting reaction times and responses.

Data from participants 26-37 was collected by Story Kiser, an undergraduate research assistant. Ms. Kiser was trained directly by the author in hearing screening and the experimental workflow and tools to ensure that the experimental process was as similar as possible across experimenters.

At the conclusion of the experiment, the participant was prompted for any questions, given a proof of participation form for the subject pool, and dismissed.

Participation took less than 15 minutes in total for each listener. Table 40 is a listing of experiment participants, annotated by gender (as given), age, place of origin, any other languages which the

[37] The audiometer had last been calibrated by MSR West on 10-29-14, approximately four months prior to the study.

participant claimed to speak fluently, and any other notes. All participants were native speakers of English, raised by native English speakers.

Table 40: Participants for Experiment 3

Identifier	Gender	Age	Place of Origin	Other Fluent Languages	Notes
p01	F	18	Colorado, USA	–	
p02	M	43	Wisconsin, USA	Thai, Lao, Khmer	
p03	F	26	Utah, USA	–	
p04	M	19	Massachusetts, USA	Polish	
p05	F	19	Colorado, USA	–	
p06	F	18	New Jersey, USA	–	
p07	F	19	Colorado, USA	–	
p08	M	18	Hong Kong	Mandarin, Cantonese	
p09	F	19	Colorado, USA	Korean	
p10	M	20	Colorado, USA	–	
p11	F	18	Massachusetts, USA	–	
p12	F	18	Colorado, USA	–	
p13	F	19	Virginia, USA	–	
p14	F	21	USA	Spanish	
p15	F	18	Colorado, USA	–	
p16	F	20	Colorado, USA	Korean	
p17	F	19	Colorado, USA	–	
p18	F	20	New Jersey, USA	–	
p19	F	19	Illinois, USA	French (College L2)	
p20	F	20	Colorado, USA	–	
p21	M	23	Washington, USA	–	
p22	F	20	Colorado, USA	–	
p23	F	18	Colorado, USA	–	
p24	F	20	Colorado, USA	–	
p25	F	18	Colorado, USA	–	
p26	F	20	Texas, USA	–	
p27	M	19	Texas, USA	–	
p28	F	21	Singapore	Mandarin	
p29	M	19	Colorado, USA	–	
p30	F	19	Indonesia	Indonesian	
p31	F	20	Colorado, USA	–	
p32	F	22	New Jersey, USA	–	
p33	F	19	Colorado, USA	Cambodian	
p34	F	18	New Jersey, USA	–	
p35	F	18	Colorado, USA	–	
p36	Agender	20	Colorado, USA	ASL	
p37	M	19	Colorado, USA	–	
p38	M	20	Illinois, USA	–	
p39	F	18	Colorado, USA	–	
p40	F	20	Texas, USA	–	
p41	F	24	Minnesota, USA	French	
p42	F	20	Texas, USA	–	

6.6 Methods: Analyzing Reaction Time and Accuracy

To address all of our research questions, we will use just two metrics: accuracy and reaction time.

Before we proceed, a note on "accuracy". All of our stimuli started as either oral words ("bad") or nasalized-vowel words ("ban"), and listeners are asked to identify whether they heard the oral word, or its nasalized-vowel minimal pair. When the listener "correctly" guesses the *original* identity of the word they overheard, this is counted as an "accurate" response, but the term "accuracy" is somewhat misleading.

The experimental tokens have all been purposefully modified to increase the odds of a "misperception", so, practically speaking, these "misperceptions" are of greatest importance, and are the "goal". If the listener picks "ban" when they hear the word "ban", the response is "accurate", but does not reflect the intent nor "desired" outcome in the experimental design.

Thus, instead of thinking "How often did the response match the nasality of the original stimulus?", we should reframe the question, and ask "Did the listener confuse oral and nasal sounds?", as this is the most useful perspective for evaluating these results. So, although we'll be testing and discussing accuracy throughout the remainder of this analysis, it will often be helpful to discuss lower accuracy values as "increased oral/nasal confusion", and to discuss this confusion explicitly.

This also motivates our choice to reaction time based analyses using data from *all* responses, not just from "correct" responses (as is the norm). Although it makes sense to exclude misclassifications in most RT studies, for the reasons outlined above, there *are* no misclassifications here, and the amount of time taken for an "inaccurate" response is just as interesting as the time taken for an "accurate" one[38].

6.6.1 Analysis: Software and Modeling

All statistical comparisons were again done using the R Statistical software suite.

For reaction time, we continue to use the `lmer` function of the `lme4` package to produce Linear Mixed-Effects Models.

For accuracy, which has just two options ("1" and "0"), we must use a slightly different analysis, called `glmer` or "Generalized Linear Mixed Effects Regression" (also part of the `lme4` package). In most regards, it functions identically to the previously discussed lmer models, taking both fixed and random effects, and permitting random slopes and intercepts to be used.

For all models, we will be using the "bobyqa" (Bound Optimization By Quadratic Approximation) optimization algorithm built in to `lme4`, as is widely recommended with complex models, and helps to avoid "failure to converge" issues. Although it is only necessary for the `glmer` models, it will be used for `lmer` models as well, for consistency.

[38] Across the entire dataset, listeners gave "accurate" responses 93 ms faster than "inaccurate" responses ($t = -8.81$), but there was no accuracy by control interaction ($t = 0.6$), indicating that although listeners took longer when confused, this was not an experimentally relevant or modification-related difference.

Finally, in all analyses, we will be separating feature reduction and addition stimuli, to improve our ability to separate "necessary" and "sufficient" features for nasality.

All 42 participants' data was included in the final analysis.

6.6.2 Analysis: Fixed and Random Effects

Although each analysis will be different, some effects will occur repeatedly, and deserve some specific mention. Take, for example, the glmer model below, which tests the effect of condition on accuracy for the experimental dataset:

```
glmer(accuracy ~ condition * control + (1|stimstructure) + (1|vowel) + (1|speaker
    ) + (1|correctword) + (1|participant),data=rem,binomial,control=glmerControl(
    optimizer"="bobyqa))
```

In addition to the parameter being investigated, 'accuracy', the model elements and parameters are as follows:

- `condition*control` - This term captures the effect of condition (which feature is being modified), control (experimental ("ex") vs. control ("con") stimuli), and their interaction (as we are only interested in changes in rt/accuracy by condition which *do not* manifest in the control stimuli.
- `(1|stimstructure)` - This random effect allows the model to account for any difference in the difficulty of the task in CVCs, CVNs, and NVCs in CVC/CVN pairs vs. CVC/NVC pairs.
- `(1|vowel)` - This random effect allows for variation by vowel in RT and accuracy[39]
- `(1|speaker)` - This random effect allows for speaker-specific differences in perceptibility.
- `(1|correctword)` - This random effect allows for word-specific differences in perceptibility (e.g. "People had a difficult time identifying "bonfire" ")
- `(1|participant)` - This random effect allows participants to differ in their RT or accuracy.
- `data=rem` - This specifies that the data used in this particular run is the reduction ("removal") data.
- `binomial` - This specifies to use a binomial variance function, as is required for a binary dependent variable.
- `control=glmerControl(optimizer="bobyqa")` - This has nothing to do with our "control" variable, but instead, controls the model, specifying that the BOBYQA optimizer should be used.

The models tested for accuracy use identical sets of effects, but test "rt" instead of "accuracy", using an `lmer` model instead of a `glmer` model, and omitting the "binomial" specifications, which is only relevant to `glmer`.

[39] Random Slopes by Condition by Vowel were tested, but did not provide significant reduction in error relative to random intercepts in a model comparison.

6.7 Results: Experiment 3 - Testing Human Perception of Nasality

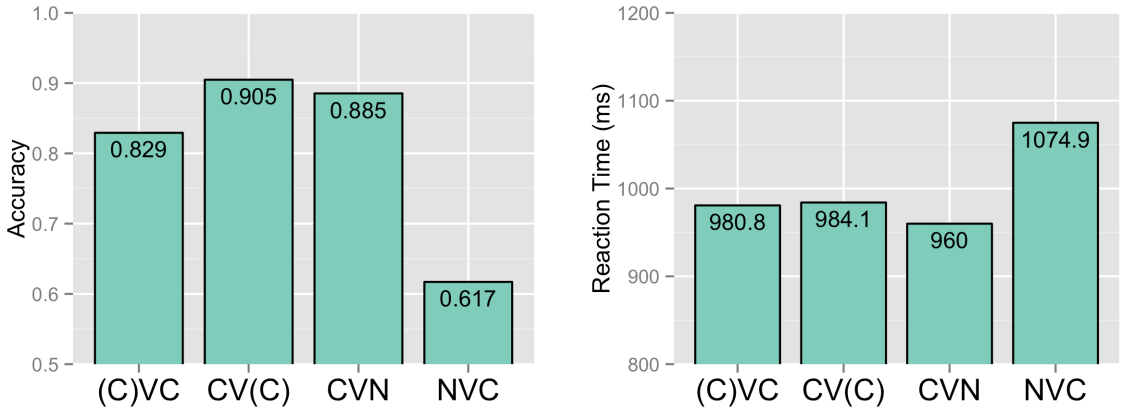
Below are the results from Experiment Three, presented for all 42 listeners. We will first discuss the baseline reaction time for unmodified stimuli to gauge the difficulty of the task, then describe the methods and models used to analyze the data. Then, finally, we will discuss the perceptual results of the addition and reduction stimuli, and the move on to discussion.

6.7.1 Results: Baseline Confusion and Reaction Time

First, we must find the baseline confusion and reaction time for this task.

To do this, we will use the Duration Control stimuli, which are completely unmodified. For these stimuli, we can find the mean confusion and reaction time, split across oral words ("bad" was played) and nasal words ("ban" was played). In addition, we can see whether the type of task (classification with onset masking vs. coda masking) makes a meaningful difference in terms of classifying natural stimuli, by examining means and reaction times for different phonological structures, as in Figure 8. In this figure, (C)VC indicates a CVC word in an onset masking context, and CV(C) indicates a CVC in a coda masking context.

Figure 8: Accuracy and Reaction Time for Unmodified Stimuli by Stimulus Structure



As we take both decreased accuracy and increased reaction time to indicate increased difficulty, we can see from these data that identifying nasality, *even completely unmodified*, was substantially harder in onset masking (NVC) stimuli than for coda masking (CVN). Put differently, with information about nasal coarticulation, coda recovery was far easier than onset recovery.

Although this does not strongly affect the experiment (as we care about increased difficulty only across the different conditions), it is worth noting, and justifies the inclusion of `stimstructure` as a random effect in all models.

6.7.2 Results: Interpreting LMER and GLMER Models

Before we examine our four models, we must note three complexities in interpretation which will apply throughout.

First, note that in the model's output, the coefficients and p-values given are always *relative to a reference level*. For instance, for the `control` fixed effect, "con" (control stimuli) is the reference level, and provides the default. So, any coefficients and interactions represent the change as we move to the "exp" (experimental) stimuli.

For condition, A1-P0 is the reference level (as we have no "natural" or "unmodified" condition with both control and experimental stimuli to compare against). This means that each other condition will show a significant effect *only if* it differs significantly from a model in which A1-P0 is modified. If A1-P0 were to be the sole significant condition, we would see all other conditions showing a significant effect of the difference (along with telling changes in the by-condition means). If A1-P0 shows no effect on perception, any condition showing meaningful effect will manifest this as a significant difference from A1-P0. This complicates interpretation slightly, but *post-hoc* comparisons of condition pairings allows more precise interpretations of across-condition differences.

We should also discuss the role of the condition*control interaction in this (and all subsequent) models. In the output of the model, we will see both main effects of condition and condition-experimental effects, showing the changes in the condition in experimental stimuli.

A significant *main* effect for Condition indicates that *all stimuli* with a certain modification performed differently, that is, both control and experimental stimuli with a certain modification were perceived uniformly differently. Such effects are interesting, but say more about the modifications made to stimuli than about the conditions themselves, as they occur regardless of whether there was a final ΔFeature or not.

We are primarily interested in those features which show a significant condition by control interaction, that is, those conditions which differ from the others *only in modified, experimental stimuli*. Conditions which show no interaction with `control` can be understood as irrelevant to perception.

6.7.3 Results: Oral-to-Nasal Cue Addition Stimuli

Now, with this in mind, we will examine the first two of our four main models.

In order to show that the manipulations performed here had *any* effect on perception, we must show that, when controlling for variability associated with vowel, structure, contrast, participant, speaker, and word, listeners responded differently to experimental stimuli in one condition versus another. To test this, the models below were run, testing accuracy and reaction time by condition in oral-to-nasal cue addition stimuli. Their output is displayed in Table 41.

```
glmer(accuracy ~ condition*control + (1|stimstructure) + (1|vowel) + (1|speaker)
   + (1|correctword) + (1|participant),data=add,binomial,control=glmerControl(
   optimizer"="bobyqa))
```

```
lmer(rt ~ condition*control + (1|stimstructure) + (1|vowel) + (1|speaker) + (1|
    correctword) + (1|participant),data=add,control=lmerControl(optimizer="bobyqa"
    ))
```

Table 41: Model Output for Addition Stimuli

	Dependent variable:	
	accuracy	rt
Intercept	2.184***	986.174***
	t = 4.864	t = 28.016
Condition-A3P0	0.024	3.105
	t = 0.157	t = 0.171
Condition-AllMod	0.153	−0.933
	t = 0.959	t = −0.051
Condition-Duration	−0.094	−3.754
	t = −0.614	t = −0.207
Condition-Formants	0.100	0.791
	t = 0.634	t = 0.044
Control-Exp	−0.235	−13.828
	t = −1.573	t = −0.762
Condition-A3P0:Exp	0.072	4.296
	t = 0.338	t = 0.167
Condition-AllMod:Exp	−1.431***	137.471***
	t = −7.005	t = 5.354
Condition-Duration:Exp	0.224	13.065
	t = 1.057	t = 0.509
Condition-Formants:Exp	−1.409***	70.639**
	t = −6.937	t = 2.751
Observations	8,200	8,200
Log Likelihood	−3,311.294	−60,117.720
Note:	*p<0.05; **p<0.01; ***p<0.001	

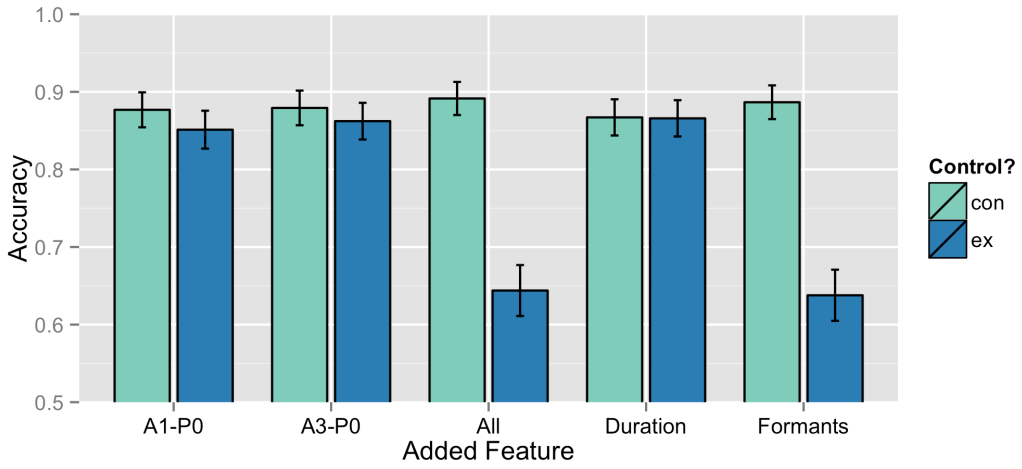
As discussed previously, the lack of main effects means that, as we might expect, the effect of `condition` and of `control` were limited to the experimental stimuli. Put differently, the *process* of modifying these features (regardless of final ΔFeature) did not universally affect accuracy or RT, only actual modifications to features.

More interesting are the condition*control interactions, which reveal the effects of our modifications in experimental stimuli.

First, we can examine the effect of condition on accuracy in the experimental stimuli. Figure 9 shows the mean accuracy for each group, with confidence intervals:

Examining Figure 9, we see rather saliently that while modifying A1-P0, A3-P0, and Duration

Figure 9: Accuracy by Condition for Addition Stimuli



had little effect on accuracy in control stimuli, modifying our formant cues, and all cues together, resulted in a large drop in accuracy, implying an increase in the number of oral stimuli listeners "misclassified" as nasal. This is supported by our model in Table 41, where Formant Modification and AllMod stimuli are the only two conditions to differ significantly from the A1-P0 baseline in the experimental stimuli.

Note, however, that even for the stimuli with the greatest oral/nasal confusion, we did not create uniformly nasal stimuli, and in the aggregate, listeners were still more likely to "accurately" call a modified oral sound "oral" than they were to re-classify as nasal[40]. Thus, although we can certainly claim to have affected nasal perception and that formant modification can give the percept of nasality, we cannot claim to have generated flawless and convincing nasal vowels through this modification alone.

This pattern is mirrored in Reaction Time. Means and confidence intervals are shown for reaction time in Figure 10.

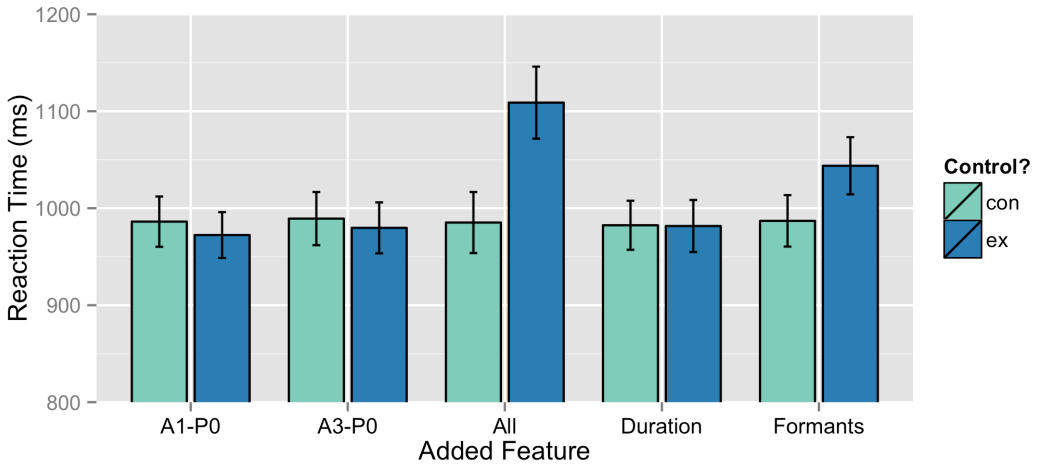
Again, we see that the perceptual differences caused by modification are inconsequential for A1-P0, A3-P0, and Duration, but are rather salient for Formants and AllMod stimuli: modifying formants, or all of the features, results in much slower reaction times for classification. And, again, the model reinforces this finding: *only* Formants and AllMod stimuli differ from the A1-P0 baseline in experimental stimuli.

The lack of main effects for Formant and AllMod indicates that these differences in accuracy and RT are *not* simply results of the modification process (artifacts of the artifacts, if you will). If it were the case that modifying formants simply made sounds more difficult to classify, and the nature of the modification had no bearing, we would expect *all* formant-modified stimuli to show decreased accuracy and increased RT, not just experimental stimuli.

Given these data, we might wonder whether the accuracy and RT differences in the AllMod stimuli

[40]There was considerable across-file variation, with some specific tokens showing 100% "accuracy", and some showing accuracy as low as 24%.

Figure 10: Reaction Time by Condition for Addition Stimuli



(which are themselves formant-modified) can be attributed entirely to the formant modification. To test this, we can run the same models on a subset of the data containing only Formant modified and AllMod stimuli. If the two conditions are identical in terms of their reaction time and accuracy effects (and thus, likely both showing the same formant-related effect), we will see *no* significant Condition*Control interaction. The output of such a model is shown in Table 42:

Table 42: Limited Model comparing Formants and AllMod (Addition)

	Dependent variable:	
	accuracy	rt
Intercept	2.314***	987.015***
	t = 4.579	t = 24.136
Condition-Formants	−0.052	1.724
	t = −0.325	t = 0.085
Control-Exp	−1.663***	123.643***
	t = −11.875	t = 6.090
Condition-Formants:Exp	0.022	−66.832*
	t = 0.114	t = −2.327
Observations	3,280	3,280
Log Likelihood	−1,524.952	−24,443.710
Note:	*p<0.05; **p<0.01; ***p<0.001	

When only AllMod and Formant stimuli are examined (and thus, AllMod becomes the reference level for the analysis), there are two changes.

First, we gain a significant main effect for Control-Exp, meaning that overall, experimental stimuli are both less accurate and slower than control stimuli. This is to be expected, given that we've cherry-picked the two conditions where Control and Experimental stimuli differed most.

For accuracy, we see no significant main nor condition*control effect of Formant modification, indicating that we cannot differentiate changes in accuracy from Formants from those in AllMod. Put differently, based on these data, the effect of modifying Formants on accuracy seems to be identical to the effect of modifying All stimuli, and thus, they are probably the same effect.

For reaction time, we do see a significant Formant:Exp effect, indicating that listeners reacted more slowly to the AllMod stimuli. Given the lack of a matching effect for accuracy, and the lack of any significant effect for the other conditions, the most reasonable interpretation is that the additional processing done to the modified "AllMod" stimuli resulted in some increased difficulty. However, there is no reason to believe that the difficulty was increased by modifications other than to formants.

So, based on these data, in feature addition stimuli, reaction time and accuracy are significantly affected *only* by the modification of vowel formants (alone or as part of a larger group of modifications). Perceptually speaking, modifying formant-related resulted in increased oral-nasal confusion, and in slower classification speeds, strongly implying that nasality-related formant patterns may be *sufficient* for the perception of nasality.

6.7.4 Results: Nasal-to-Oral Cue Reduction Stimuli

On the other side, we have the cue reduction stimuli, in which we modified nasal words to have oral vowels for nasal features, with aim to reduce the perceived nasality of the stimuli.

The models used to evaluate these data are identical in construction to those above in Section 6.7.3, but here are run only on responses to nasal-made-oral cue reduction stimuli. The output of these models is shown in Table 43.

```
glmer(accuracy ~ condition*control + (1|stimstructure) + (1|vowel) + (1|speaker)
    + (1|correctword) + (1|participant),data=rem,binomial,control=glmerControl(
    optimizer"="bobyqa))
lmer(rt ~ condition*control + (1|stimstructure) + (1|vowel) + (1|speaker) + (1|
    correctword) + (1|participant),data=rem,control=lmerControl(optimizer="bobyqa"
    ))
```

First, we will examine the accuracy data, summarized in Figure 11.

First and foremost, we see that *there are no significant condition * control interactions*. Thus, although these data are interesting, they will not help us to address our experimental hypotheses.

In fact, the only significant effects are *fixed* effects for A3-P0 and Formants. Recall that for our `lmer` and `glmer` analyses, coefficients are always stated relative to a reference level, and for Condition, the reference level is A1-P0's change in accuracy.

When we examine the means, we see that across both experimental and control stimuli, manipulating A1-P0 resulted in reduced accuracy, and this is reflected in the AllMod stimuli as well. Thus, these significant fixed effects for A3-P0 and Formants actually indicate that overall accuracy for both are significantly higher than the accuracy for A1-P0 stimuli, but that AllMod and Duration are not demonstrably different in overall accuracy.

Table 43: Model Output for Reduction Stimuli

	Dependent variable:	
	accuracy	rt
Intercept	1.297	1,062.575***
	t = 1.321	t = 12.383
Condition-A3P0	0.337**	−39.629*
	t = 2.690	t = −2.294
Condition-AllMod	−0.104	−11.491
	t = −0.859	t = −0.665
Condition-Duration	0.208	−37.195*
	t = 1.676	t = −2.153
Condition-Formants	0.296*	−54.496**
	t = 2.371	t = −3.154
Control-Exp	−0.00000	−40.432*
	t = −0.00003	t = −2.340
Condition-A3P0:Exp	−0.057	47.980*
	t = −0.324	t = 1.964
Condition-AllMod:Exp	0.052	68.021**
	t = 0.300	t = 2.784
Condition-Duration:Exp	0.154	49.344*
	t = 0.874	t = 2.019
Condition-Formants:Exp	−0.048	97.728***
	t = −0.275	t = 4.000
Observations	8,200	8,200
Log Likelihood	−3,870.367	−59,723.660
Note:	*p<0.05; **p<0.01; ***p<0.001	

Thus, the sole takeaway from the accuracy data in feature reduction stimuli appears to be that *the process* of A1-P0 modification reduces the perception of nasality in vowels, no matter the eventual ΔFeature.

This pattern is mirrored in the Reaction Time data, summarized in Figure 12.

Here, alongside a significant main effect of experimental vs. control, we again see fixed effects for A3-P0 and Formants, along with a marginal effect for Duration modification, all indicating *decreased* reaction times relative to A1-P0.

Again, we can interpret this to mean that some aspect of the modification of A1-P0, whether or not there was a final ΔA1-P0, made the task more difficult for listeners, and that all conditions which did *not* include A1-P0 modification were simply faster. Note as well that although A1-P0 does appear to show differences in means between experimental and control (and, indeed, in the opposite direction), the lack of consistent condition * control effects indicate that these differences

Figure 11: Accuracy by Condition for Reduction Stimuli

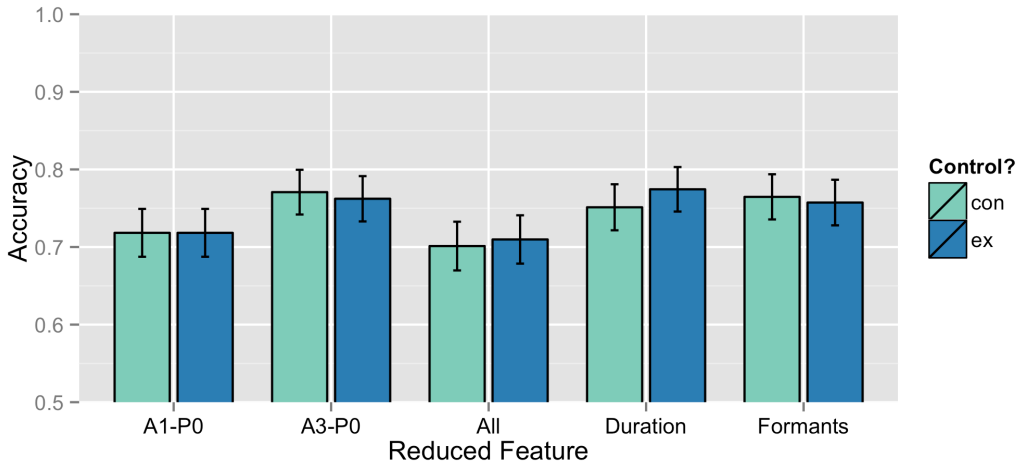
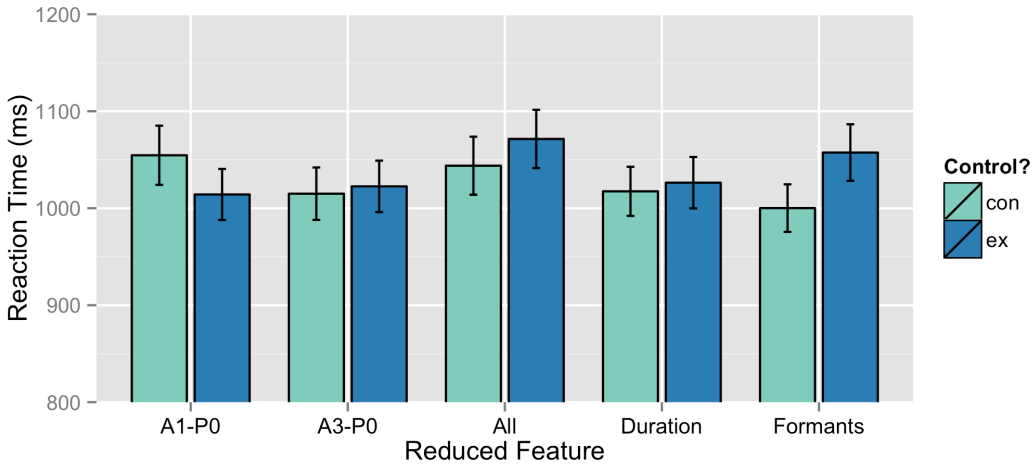


Figure 12: Reaction Time by Condition for Reduction Stimuli



do not reach the threshold for significance. We will discuss this finding in more depth in Section 6.8.7.

However, unlike for the accuracy data in feature reduction, *we do see two significant condition * control effects*, a rise in RT for AllMod, and a rise in RT in formant modified stimuli.

These interactions, representing the change in reaction time between experimental and control stimuli for each condition, mirror the findings in Section 6.7.3, and again seem to indicate that modifying formants to nasal values (either alone or as part of a larger set) is difficult, and slows down our categorization of vowels as “oral” or “nasalized”.

In an effort to again rule out “bad stimuli” effects, we can compare the main effect and condition * control effects for Formant. We see that although there is a main effect on reaction time for Formants, it shows that formant-modified stimuli show faster RTs than the somewhat unusual

A1-P0 data. Thus, the increased reaction times in experimental contexts seem again to indicate a change in perception due to ΔFeature, rather than due to the modification process itself.

Although the AllMod difference appears less strong, we can again perform a sub-group analysis, comparing AllMod and Formant stimuli, to attempt to establish whether the AllMod stimuli are different above and beyond the effects of Formant modification, shown in Table 44.

Table 44: Limited Model comparing Formants and AllMod (Reduction)

	Dependent variable:	
	accuracy	rt
Intercept	1.165	1,052.250***
	t = 1.238	t = 11.146
Condition-Formants	0.393**	−43.724*
	t = 3.196	t = −2.470
Control-Exp	0.051	27.589
	t = 0.422	t = 1.559
Condition-Formants:Exp	−0.098	29.707
	t = −0.566	t = 1.187
Observations	3,280	3,280
Log Likelihood	−1,606.446	−24,005.170
Note:	*p<0.05; **p<0.01; ***p<0.001	

We see from these comparisons that although Formant modified stimuli show different baseline accuracy from the AllMod stimuli (based on the significant *fixed* effects), there is no significant condition * control interaction, and thus, *the AllMod and Formant modified stimuli do not differ meaningfully in perception.*

6.7.5 Experiment 3 Results: Summary

Before we move into a more abstract discussion, we should summarize the findings of our analyses. As in the analyses, we will discuss addition and reduction stimuli separately.

For addition stimuli:

- *Modifying the Formant structure resulted in both increased oral-nasal confusion and slower responses*.
  - No other conditions showed a significant difference in accuracy or RT between experimental and control stimuli.
  - The fact that this Formant effect only shows up in experimental stimuli indicates that these differences are unlikely to be entirely artifacts of the signal processing.
  - The AllMod stimuli mirrors this pattern, but does not appear to represent a distinct effect.

For reduction stimuli:

- *Classification accuracy was not affected by our modifications for reduction stimuli*
  - No significant condition * control interactions were found
- Mirroring the addition stimuli, *modifying formant structure resulted in slower responses*.
  - Again, no other conditions showed a significant effect on perception in experimental stimuli.
  - Although there was a main effect, it was in the opposite ("faster") direction, contrasting with A1-P0, suggesting the experimental effect goes beyond "stimulus badness".
  - Again, there was no significant difference between the perceptual effects of the AllMod stimuli and the Formant Stimuli.
- *Any modification of A1-P0, control or experimental, resulted in increased listener difficulty*.
  - This was only a factor in feature reduction stimuli, and applied to all stimuli where A1-P0 was modified *at all*, experimental and control.
  - Although this does not affect the interpretation of the formant effects above, it does merit some discussion (see Section 6.8.7.

6.8 Experiment 3 Discussion: The Perceptual Cues to Nasality in English

Now, we will evaluate the hypotheses proposed in Section 6.2, discuss two of the interesting questions raised by the data, and finally, synthesize these results into a discussion of the perception of nasality in English.

6.8.1 Hypothesis 1: Modification will affect Reaction Time

Our first hypothesis states:

> *When features are added or reduced, there will be an increase in RT (indicating increased difficulty) relative to the control stimuli.*

For addition stimuli, we found that modifying formants resulted in an increase of reaction time of ~64 ms relative to control stimuli, and for reduction, formant modified stimuli were ~96 ms slower. The AllMod stimuli showed a statistically indistinguishable Reaction time effect, indicating that the perceptual differences are likely caused by their included formant modifications.

In both cases, we take this to mean that formant modifications did increase the difficulty of the task, and that regardless of their eventual choice, these modifications forced listeners to take a greater amount of time to make oral/nasal decisions.

Based on these data, we find that this hypothesis is **supported** in both addition and reduction stimuli when modifying vowel formants.

6.8.2 Hypothesis 2: Modification will affect Accuracy

Our second hypothesis mirrors the first, but discusses accuracy:

> *When features are added or reduced, there will be increased oral-nasal confusion (shown as decreased accuracy) relative to the control stimuli.*

In the addition stimuli, we see that both Formant and AllMod stimuli showed a reduction in accuracy of around 25%, which was found significant by a high margin.

However, in the reduction stimuli, no modifications showed significant experiment-control differences in confusion. Given that reaction time *was* consistent across both modification types, and mirrors the feature addition accuracy effect in direction and condition, it seems unlikely that "de-nasalizing" nasal vowels requires the manipulation of different features.

Instead, a more reasonable conclusion is that although reaction time may always be affected, it is more difficult to perceptually "de-nasalize" a nasal vowel. This is in line with Hypothesis 6, below.

So, based on these data, we find that this hypothesis is **supported**, but only for the Formant and AllMod conditions, and only in the addition stimuli.

6.8.3 Hypothesis 3: Features will differ in their effects

Our third hypothesis simply states that all features will not be equally perceptually important.

> *Different features will show different changes in accuracy and reaction time, according to their differing degrees of perceptual usefulness.*

Unfortunately, these data do not allow us to sort the features by perceptual usefulness, as only one feature modification caused significant experimental vs. control differences in accuracy and reaction time.

However, because of this strong difference among the conditions, this hypothesis is very clearly **supported**.

6.8.4 Hypothesis 4: Features will be symmetrically useful

Our fourth hypothesis simply asserts that features which are useful for perceiving nasality will be so in all cases.

> *Those features which are necessary for nasal perception (whose reduction in nasal vowels reduces perceived nasality) will also be sufficient for nasality perception (their addition to oral vowels will also create the percept of nasality).*

This is slightly harder to evaluate. Vowel formant changes show a clear effect on perception in feature addition stimuli, and thus, appear to be *sufficient* for nasal perception, at least for some listeners in some stimuli.

Formant modification shows a similar effect on reaction time for feature reduction stimuli, where changing formants to oral values decreases the speed of identification of nasal stimuli. However, no accuracy effect is found for *any* features in feature reduction stimuli, and thus, we cannot find that nasal formant characteristics are *necessary* for vowel perception.

So, although the symmetry in reaction time suggests that vowel formants may be useful for creating the perception of nasality and useful for taking it away, the asymmetry in accuracy prevents us from making a definitive claim.

6.8.5 Hypothesis 5: Modifying all Features will have the strongest effect

Our fifth hypothesis, based in the logic that more features will mean more perceptual evidence, and thus, stronger effects, stated that:

> *Reducing or adding all of the features will have a greater effect on classification and reaction time than any of the single feature manipulations.*

Given that in every case where the AllMod stimuli showed any significant perceptual effect, post-hoc testing revealed that the effect was statistically indistinguishable from that of the Formant Modification condition, it's rather straightforwardly **unsupported**.

6.8.6 Hypothesis 6 - Feature Addition will be more salient than feature removal

Finally, in our sixth hypothesis, we posited that addition would be more powerful overall:

> *Adding features to oral sounds will have a greater effect on the perception of nasality than reducing features in nasal sounds.*

The asymmetry predicted here does seem to manifest in our data. Modifying formants appears to have strong and significant effects on both accuracy and reaction time in feature addition stimuli, but the effects of formant modification are only visible in reaction time for feature reduction stimuli. This would appear to indicate that although formant modification can increase the *difficulty* the perception of existing nasal vowels, it does not increase *confusion*, and listeners are able to use some additional information to "correctly" classify the words.

The idea that addition is more perceptually meaningful than reduction works in terms of accuracy, but interestingly, we cannot claim that addition caused greater difficulty than reduction *on the basis of reaction time*. Although rigorous statistical comparison across these two models is not possible, simply looking at the RT differences attributed to Condition-Formants:Exp in both models, we see that while addition showed a ~64 ms increase in RT, cue reduction actually showed a greater increase, ~96 ms.

Assuming the model outputs are both accurate and comparable, this would indicate that listeners took longer (and thus, had more difficulty) in establishing the nasality of reduction stimuli, even if they ultimately ended up being less confused.

Regardless of the *reason* for the asymmetry, which we will discuss extensively in Section 6.8.8, our original hypothesis does not quite capture the finding. We could rephrase this hypothesis into a more accurate claim:

> *Adding nasal features to oral vowels is more likely to result in listener reclassification than reducing nasal features in nasal vowels.*

So, although such an asymmetry would help to explain the difference between reaction time and accuracy in the feature reduction stimuli in this experiment, we cannot straightforwardly accept the original claim. However, the modified claim, Hypothesis 6b, appears quite clearly **supported** by the data.

With all hypotheses addressed, we will now explore two questions raised by the data: the odd effects of A1-P0 modification, and the perceptual asymmetry of formant modification.

6.8.7 Discussion: Effects of A1-P0 Modification

Alongside the relatively straightforward effect of Formant manipulation on perception, the feature reduction stimuli displayed an interesting pattern for A1-P0.

In terms of accuracy, A1-P0 (as well as the AllMod stimuli also incorporating the modification) displayed a main effect showing an overall drop in accuracy, indicating that something about the modification process, *even if there was no final* $\Delta A1\text{-}P0$, resulted in a reduction in accuracy.

In reaction time, although the condition * control effects were not significant, we again saw increased reaction times for A1-P0 modified stimuli, control and experimental, indicating increased difficulty. This rather odd finding could be interpreted in several ways.

To be clear, this *must* be a consequence of the modification script. Because all of the words were tested in all of the conditions, it cannot be the case that all tokens associated with A1-P0 coincidentally were more difficult to perceive.

We must also note that this effect manifested *only* in the feature reduction stimuli, and that no main effects *nor* condition * control effects were found for A1-P0 in feature addition. The lack of equivalent inaccuracy and slowdown make it seem unlikely that the modification script is outputting universally "bad" stimuli.

Similarly, the fact that the control stimuli show this odd difficulty *only in feature reduction stimuli* implies that this isn't a matter of script settings or programmed ΔFeature (because the exact same code and settings created the control stimuli for both reduction and addition, in the same run of the script).

Thus, the only explanation is that some aspect of the A1-P0 modification script functions differently (and causes more perceptual difficulty) when manipulating already-nasal vowels. Put differently, there appears to be some interaction between the script and the acoustics of an existing nasal vowel which results in tokens which demonstrate the expected ΔA1-P0, but are more difficult for listeners to classify. Although no explanation or mechanism is immediately apparent, this clearly merits further investigation, but more immediately relevant is the effect of this rather subtle interaction on the experimental findings.

Certainly, the positive findings in terms of formants are not affected or diminished. Formant modification was conducted using a different technique, in a different script, and according to established best practices, and the strength and relative consistency of the effects speak for themselves. At worst, we might have to discount the AllMod effects altogether (which, given their statistical similarity to the formant effects, is no loss).

Even if the main, formant-modification finding is not affected, we should take a moment to consider the implications for our A1-P0 finding (or lack thereof). There are three functional possibilities for the effect of this anomaly on the data.

It could be the case that, yes, some aspects of our A1-P0 modification makes nasal vowels harder to discriminate, *but* the modification is still working largely as expected, and listeners are simply not attending to A1-P0 for perception. Nothing about a more difficult task necessarily precludes listener use of A1-P0 *if A1-P0 were perceptually useful*, and it's very possible that its manipulation *could* still have resulted in perceptual changes. In this case, if A1-P0 is simply not useful for perception, we would see the present result.

It could also be that whatever aspect of the script interacts with nasal vowels does so such that the reduction stimuli, although appearing to show the proper changes to A1-P0 by external measurement, are "broken" for human listeners, and not meaningfully testing A1-P0's perception for reduction. In this case, we would have *no* reliable information about the perceptual role of A1-P0 in existing nasal vowels, and our lack of significant findings may not represent the perceptual reality. *But*, in that case, oral vowels would still have been accurately tested, and we could still

interpret the oral-to-nasal addition stimuli (which appear to show no meaningful perceptual role of A1-P0). Given the symmetry between reduction and addition in reaction time, it's not out-of-line to interpret the lack of *any* significant perceptual differences for A1-P0 as implying a lack of perceptual utility.

Finally, it could be that the A1-P0 modification is so subtly broken that, although external measurement appears to show the proper ΔA1-P0 values, the results are somehow universally meaningless to human listeners. This would need to somehow explain how the addition stimuli pattern similarly to the other conditions while still not accurately testing human perception[41]. Although it's within the realm of possibility, it seems unlikely that such subtle-yet-comprehensive breakage would both evade discovery over several years of development and manifest so subtly and asymmetrically.

So, in summary, although the issue certainly does require some additional investigation, barring the somewhat unlikely third scenario, at least some of the A1-P0 data was perceptually valid. Even if we take a conservative approach and discard any findings for reduction stimuli, despite its near universal adoption as a measurement standard and its close tracking of nasal production, we still see *no* significant effects for A1-P0 in addition stimuli, suggesting little perceptual utility for A1-P0 in general. More importantly, though, even a pervasive subtle-yet-perceptually-disatrous A1-P0 issue does not affect our principal positive result.

6.8.8 Discussion: Asymmetry of Formant Effects

Before we summarize our result, we should discuss one other interesting finding. As we discussed in Section 6.8.6, although formants showed strong accuracy effects in oral-to-nasal feature addition stimuli (and RT effects in both), it had *no* effect on accuracy in the nasal-to-oral reduction stimuli. Put differently, the addition of cues to oral vowels seems to have a stronger perceptual effect than their reduction in nasal vowels.

There are two particularly promising explanations for this effect.

The first comes from the idea of "positive evidence". When we widen formant bandwidths in an oral vowel, it provides a single, salient, and positive cue which listeners can use to decide "this vowel is nasal". We have gone from "no evidence of nasality" to "some evidence of nasality", and thus, listeners are more willing to classify these oral vowels as "nasal".

On the other hand, when we reduce the formant bandwidths in an already nasal vowel, we are not providing positive evidence that a vowel is oral. Instead, we are simply removing a portion of the evidence that it is nasal. Given that the vast majority of the signal was not modified, there is the possibility that many weak cues to nasality were left behind. In this case, our modifications are only changing "strong evidence of nasality" to "less strong evidence of nasality".

Thus, although the atypical nature of the stimuli could easily cause a slow-down in perception (as they are missing what appears to otherwise be a very strong cue for nasality), given that there

[41] This is presented as a most-conservative possibility, and should be considered, but the author can picture no mechanism for such hidden and devastating stimulus corruption.

remains considerable evidence of nasality, their ongoing classification as "nasal" is not unreasonable.

The main problem with this line of thinking is that, based on the statistical and machine-learning analyses, the AllMod stimuli have reduced nasality using *all four* of the top-performing potential cues. While it's straightforward to imagine that changing one predictor might not suffice to shift perception, the lack of accuracy effect even for the AllMod stimuli leaves us arguing that the perception of nasality in nasal vowels depends primarily on features of nasality which showed relatively poor statistical and predictive links to nasality, or features completely unattested in the literature and unrelated to the 29 features investigated here.

An alternative hypothesis for the cause of this asymmetry is that the acoustic consequences of nasality itself are not the sole distinguishing factor of nasal vowels. Put differently, when determining whether a vowel is oral or nasal, listeners may be listening for oral articulation changes, in addition to whatever changes to the signal are caused by opening the velopharyngeal port.

The possibility of such oral cues to nasality are suggested by the literature examining the oral articulations of nasal vowels, such as Carignan (2014), Carignan et al. (2015), Carignan et al. (2011), Delvaux et al. (2002b), Shosted et al. (2012), and Krakow et al. (1988). These papers all describe systematic and non-arbitrary differences in vowel quality between oral and nasal vowels which are potentially perceptible to listeners, but which are *not direct consequences of nasality itself*. These vowel quality differences may serve to enhance (or create) part of the contrast between oral and nasal vowels, improving the perceptibility of nasal vowels for listeners, even if they don't actually change or enhance the acoustical consequences of nasal airflow.

This can help to explain our findings of perceptual asymmetry because the formant modifications we made were purposefully *vowel agnostic*. The goal of this work was not to identify the perceptual cues to *specific nasal vowels*, but to identify the *acoustic consequences of nasality*. We are principally interested in formant changes which occurred due to nasality *across all vowels*, as was the case with bandwidth and F1's 36 Hz frequency change, as these changes can be attributed to nasality itself, rather than to oral shifts in articulation.

However, this means that for our specific /ɑ/ and /æ/ vowels, the modifications we made reduced the direct consequences of nasality, but without undoing the vowel-specific articulation changes which provide secondary evidence of nasality[42]. In the same way that, although it might certainly give us pause, a swan painted orange would still be identifiably a swan (as there is more to swan-hood than whiteness), nasal acoustics may not be the sole characteristic of nasal vowels. Although listeners might react more slowly to a nasal vowel which is missing the crucial nasality-related formant features, the vowel itself may still be, in terms of oral articulation, a "nasal vowel".

This "whole vowel" approach, where oral articulations play a key perceptual role in nasal perception, provides a second and perhaps more compelling explanation as to how our feature removal stimuli can cause little confusion, even when missing a seemingly powerful cue to nasality.

[42]Such articulatory changes are well known in English, particularly the diphthongization or raising of /æ/ in pre-nasal contexts.

6.8.9 Discussion: On the Perception of Nasality in English

With this data, we finally find ourselves able to answer the question at the core of this paper: What acoustical cues are listeners *actually using* to distinguish oral and nasal vowels?

Our experiment tested four features (A1-P0, A3-P0, Duration, and Formant Frequency/Bandwidth), in the speech of two different speakers and across two different vowels (/ɑ/ and /æ/), both adding these features to oral vowels, and reducing them to oral levels in nasal vowels. Although the analysis was occasionally complex, the final result is clear: the only feature whose modification had *any* effect on listener perception of nasality was formant modification.

This formant modification finding was robust, showing strong experimental vs. control differences in both accuracy and reaction time for addition stimuli, and showing significant reaction time differences in feature reduction stimuli. More importantly, this was the only feature to show significance *anywhere*, and even the all-features-modified stimuli showed no statistically meaningful difference in effect, implying that not only were formants an important cue for nasality, but they were *the* important cue for nasality. So, ultimately, for English /ɑ/ and /æ/ vowels, and there is no question that the primary cue to the presence of nasality in vowels comes from some aspect of the vowel's formant structure. However, we'd like a clearer understanding of what *element* of the formant structure is being used here.

Both statistical analysis and machine learning predicted that, for English, F1's Frequency and the Bandwidth of both F1 and F3 would be meaningful nasal cues. We modified all three features together in the human experiment, partly for efficiency and partly in recognition that formant shifts reflect the configuration of the tongue and vocal tract in general, and that formant properties often shift *en masse*. The perceptually meaningful formant change, then, consisted of one relative change (to F1's Frequency), and then two absolute changes, to F1 and F3's bandwidth, both shown below in Table 45:

Table 45: Formant Modifications in Experiment 3

Feature	"Oral" value	"Nasal" value	Unit
Freq_F1	-36	+36	Hz
Width_F1	171	296	Hz
Width_F3	498	615	Hz

Although we cannot prove, on the basis of this experiment, which of these three modifications was most important for nasal perception, we do have some clues.

In the machine learning study, F1's bandwidth generated the most accurate single-feature model in SVM classification, and in an all-inclusive RandomForest model, Width_F1 was the most "important" feature by far. In contrast, neither F1's Frequency nor F3's Bandwidth were particularly important to Machine learning (ranking 6th and 7th place in importance, respectively), indicating that they carry less predictive information than F1's bandwidth.

In addition, in our statistical analysis, F1's bandwidth showed a relatively strong oral $\rightarrow$ nasal ΔFeature relative to the oral standard deviation (0.479), particularly when compared with Freq_F1 (0.137) and Width_F3 (0.139). This suggests that the bandwidth change in F1 is likely to be

more salient, in light of the normal variability between speakers and tokens, than changes to F1's frequency or F3's bandwidth.

The idea of F1's bandwidth as a cue to nasality is not without precedent in the literature. Those few studies which have examined the perception of nasality have all described some vowel formant bandwidth involvement in nasal perception (c.f. Delvaux (2009), Macmillin et al. (1999), Kingston and Macmillin (1995), Hawkins and Stevens (1985b) and Stevens et al. (1987b)). One study which particularly parallels the present work is Hawkins and Stevens (1985b), a perception experiment looking at perception of nasality in English, Gujarati, Bengali, and Hindi. In their conclusions, the authors characterize the perceptual nature of nasality by saying:

> *"Another way of describing the stimuli [identified as nasal] is in terms of the degree of low-frequency prominence in the spectrum. […] This reduced prominence is achieved by creating an additional spectral peak near F1 or by splitting or broadening the F1 peak"*

Although we saw no evidence that the nasal peak (P0) itself played a role (given the poor showing of A1-P0), "broadening the F1 peak" is precisely the modification done here, with similar perceptual results. Thus, though the methodology, modifications, listeners and features tested were quite different, the results of this experiment appear to mirror this finding, lending further credence to the idea that the effect shown is due to bandwidth more than anything else.

Finally, there is some intuitive, acoustic sense to bandwidth as a primary cue to VP port aperture. Given that formant bandwidths are thought to be primarily reflective of thermal, friction and tissue energy loss (c.f. Stevens (1998), Huang et al. (2001)), it makes sense that formant bandwidths would increase in nasal vowels, given that opening the VP port adds substantially more cavity volume and surface area in which these losses can occur.

Thus, although we cannot definitively state that our listeners were attending primarily to F1's bandwidth, on the basis of all the other data, it seems a very reasonable hypothesis.

Of course, this experiment is not the final word on nasal vowel perception, even in English. The present study evaluates the perception of nasality in only two vowels, and although formant bandwidths will not be as affected by formant frequency as, say, A1-P0, there is no doubt that by-vowel and by-speaker variation will play a major role in the perception of nasality.

In addition, the participants in this study were all normal-hearing college students, listening to citation-speech tokens from two young female speakers. Although this provided a clean sample to maximize the chances of finding effects, it is not representative of the full range of speakers, speech contexts, and listeners. There is ample room for additional work looking at the perception of nasality in connected speech and in diverse speakers, and by persons with atypical hearing.

So, alongside some evidence of a role for oral articulation in perception of nasal vowels, we do rather conclusively find that the primary cue to vowel nasality in this experiment lies in the formant structure of the vowel, most likely in the bandwidth of the first formant.

Before we come to our final conclusions, though, we should run this experiment with just one more "listener".

7 Comparing Human and Machine Perception

Throughout this work, we have been using machine learning to model, predict, and form hypotheses about human perception. With actual human perceptual data in hand, we can now directly compare the human classifications to those provided by machine learning models, with the goal of better understanding both.

To do this, we will run one final machine learning experiment, wherein we use our best performing algorithm, Support Vector Machines, to classify the modified and unmodified stimuli from Experiment 3, and evaluate the "perceptual" effects of the various modifications on the classifier's accuracy. The results of this study should help us understand and support our findings from experiment 3, and should help to reveal some of the strengths (and weaknesses) of using machine learning in perceptual tasks.

7.1 Experiment 4: Structure and Goals

Experiment 4 is composed of two steps. First, we will test whether SVMs trained on a variety of datasets actually display any increased confusion in control vs. experimental (feature-modified) stimuli, as our human listeners did. This will tell us whether our feature modifications have had any effect on non-human classification, and will offer an initial point of comparison to human perception.

Then, if a difference between experimental and control stimuli is found, we can proceed to examine the by-condition differences in classification confusion for SVMs, which will allow us to see whether the patterns of confusion in a classifier mirror those of the listeners in Experiment 3.

All comparisons will be conducted using accuracy, and human/machine comparisons will be made across optimally comparable datasets. Reaction time cannot be compared, as support vector machines do not provide any similar and interpretable difficulty metrics.

7.1.1 Experiment 4: Hypotheses

Unlike in the previous experiment, there is only one hypothesis to test:

> **Hypothesis 7 –** *When given the stimuli from experiment 3, Support Vector Machines will show the same patterns of variation in by-condition accuracy as the human listeners.*

7.2 Experiment 4: Methodology

Before we proceed, we must discuss the algorithms, methods, and datasets used in this final experiment.

First, features were extracted from the testing stimuli (pre-noise-addition) using the same measurement script as has been used throughout the study, and relevant information about their

condition, speaker, was appended[43]. Z-score normalized values for each feature were calculated for each timepoint.

Then, the same data used for statistical analysis and classification in Sections 4 and 5 was imported, and, again, Z-scored.

To build the model, the `e1071` R package was again used. The SVM methods used were nearly identical to those used in Section 5, although the feature set was slightly reduced, to eliminate some redundancy and increase evaluation speed. Each classification discussed here used the same parameters, although several different datasets were used for training. After extensive tuning, using a value of "1" for the cost parameter (C) was determined to provide the best accuracy in this task. Below is the command used to generate the English-trained SVM model:

```
ensvm <- svm(nasality ~ Amp_F1 + Amp_F2 + Amp_F3 + Amp_P0 + Amp_P1 + Freq_F1 +
    Freq_F2 + Freq_F3 + A1P0_Highpeak + A1P1 + Duration + Width_F1 + Width_F2 +
    Width_F3 + P0_Prominence + A3P0, kernel"="radial, cost=1,data=EnAll)
```

Although we will ultimately be testing our models on the modified, experimental stimuli, our goal is to create a model which will simulate a human's perception of nasality in general, we will train the model using unmodified tokens from actual human speakers. Three different datasets were used for training, resulting in three different models.

The **EnNoNVN** dataset includes CVCs, CVNs, and NVCs from the English dataset used throughout this project. It provides exemplars of the two types of nasality included in the stimuli, as well as CVCs. This is the most specific dataset, although this specificity does mean a lower number of training tokens.

The **EnAll** dataset includes CVCs, CVNs, NVCs and NVNs for English, and provides exemplars of all of the possible types of nasality in English, and is the largest monolingual dataset possible. Comparison with the previous model will address whether these models are better when trained on analogous data, or when given more data, even if only somewhat related.

Finally, the **EnFrAll** dataset includes CVCs, CVNs, NVCs and NVNs for English, as well as oral and nasal vowels for French, and corresponds to the *entire* corpus collected for this project. It contains the most training data, and if one can build a language-general acoustical model for nasality, one trained on this dataset should come as close as is possible in the present work.

Once a model was generated, trained on a given dataset, it can then be applied to another dataset using the `predict` function. So, to classify the stimuli ("stimnew") using the English model created above, one would use:

```
enclassstim <- predict(ensvm, stimnew[,-1])
```

This results in a set of predictions, from which accuracy can ultimately be calculated.

It is worth noting briefly that the number of classifications performed here on the test stimuli is higher than for the humans, as there were three timepoints measured per word. This is intentional, meant both to reduce the effects of a single mis-measurement in a given stimulus, as well as to

[43]The data used here are the same as were generated for checking the modifications in Section 6.3.8. All of the same provisos about re-measuring modified stimuli apply here as well.

increase the size of the test set, reducing the influence of any individual misclassification on the overall score.

7.3 Experiment 4 Results: Human Perception vs. Machine Perception

Again, there are two fundamental questions being asked in this experiment: Did our modifications affect the classification by the SVM models, and if so, how did classification accuracy differ across the different conditions.

7.3.1 Experiment 4 Results: Control vs. Experimental Stimuli

To evaluate whether the modifications affected machine classification at all, we must simply compare scores in the control vs. experimental stimuli. If the machine learning algorithms performed more poorly (showed decreased accuracy and increased confusion) in the experimental stimuli, our modifications have succeeded in influencing machine "perception".

When using SVMs, each run of the model is deterministic; that is, running the same model on the same data will always return the same results. So, whereas for the humans we have one classification per word * 42 participants, each SVM model provides just one classification per point, and additional "listeners" can't be generated without modifying accuracy-critical parameters. This means that we simply do not have sufficient data for the individual machine learning models to perform more rigorous statistical modeling, and we will simply take the classifier accuracy figures at face value, and compare them to the means for human listeners.

Figure 13 and Table 46 show the relative accuracy of our three models on control and experimental stimuli, as well as the mean accuracy for human listeners on the same data.

Figure 13: Confusion by Stimulus Type for Humans vs. SVMs

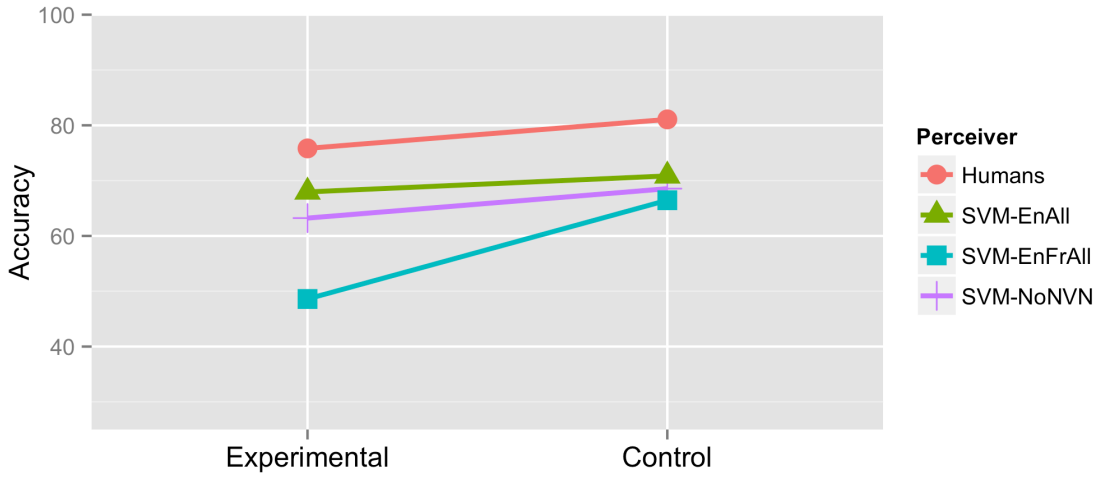


Table 46: Accuracy Classifying Control and Experimental Stimuli as Oral vs. Nasal

	Overall	Control	Experimental
SVM-EnNoNVN	65.924	68.553	63.226
SVM-EnAll	69.427	70.860	67.957
SVM-EnFrAll	57.643	66.457	48.602
Human Listeners	78.451	81.073	75.829

We can see that the human listeners outperformed the SVMs in both experimental and control contexts, and that all three models showed reduced accuracy in classifying experimental stimuli.

We also see that overall, the EnAll-trained model, trained on the entire English dataset, performed best, followed by the EnNoNVN-trained model, followed distantly by the English/French combined model. This again points towards language specificity, at least in the computational perception of nasality.

Finally, we see that both English-based models classified well above chance for all stimuli, and that the EnFrAll-trained model actually dipped below chance for experimental stimuli. This indicates that knowledge of French nasality patterns actually left the SVMs *more* vulnerable to being mislead by the stimuli.

So, it appears that although our SVMs *can* classify these stimuli (all are at or near chance), our humans bested them at the task, although SVMs suffered a similar drop in accuracy when classifying modified stimuli.

7.3.2 Experiment 4 Results: Accuracy by Condition

Given that the experimental modifications do appear to have an effect on the SVMs, we can now more closely examine the nature of this effect across each of the different modification conditions.

First, we'll examine the control stimuli, which should display no meaningful across-condition variability, as shown in Figure 14.

We see here that, as expected, there is little variation across the control stimuli, with all three models performing similarly, and again, with lesser accuracy than the humans. However, this itself is interesting, as it implies that the majority of the across-model difference is found in the experimental stimuli.

Across-condition variation in the experimental stimuli is shown in Figure 15, and numerically in Table 47.

A few patterns are immediately apparent from these data.

First, we see that again, the humans have the best classification accuracy (although it's only around 3% better than EnAll for formants).

Figure 14: Confusion by Condition for Humans vs. SVMs (Control Stimuli)

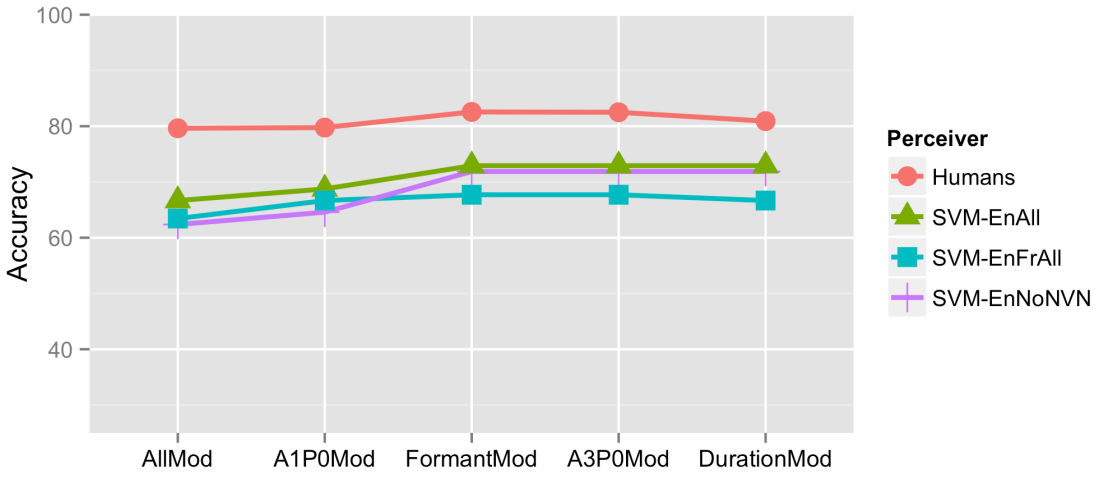


Figure 15: Confusion by Condition for Humans vs. SVMs (Experimental Stimuli)

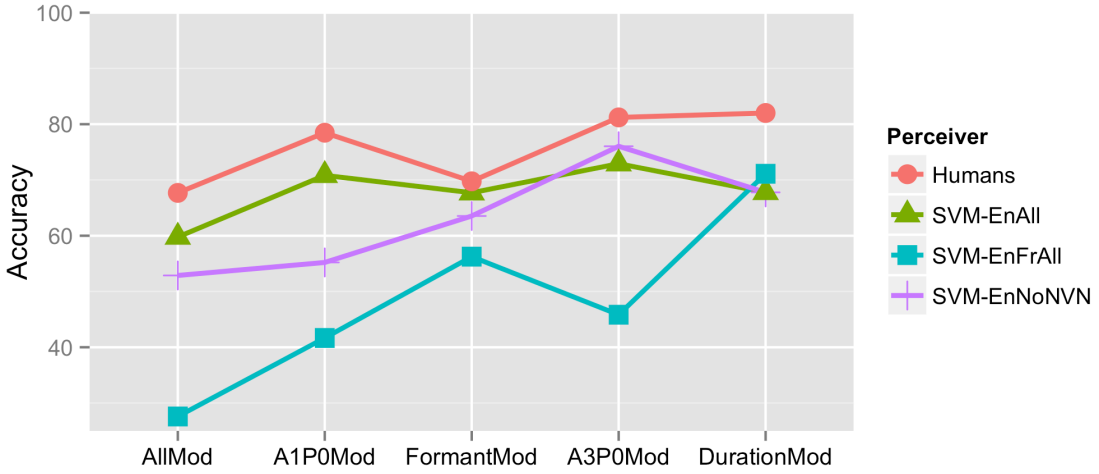


Table 47: Accuracy Classifying Experimental Stimuli as Oral vs. Nasal by Condition

	AllMod	A1P0Mod	FormantMod	A3P0Mod	DurationMod
SVM-EnNoNVN	52.874	55.208	63.542	76.042	67.778
SVM-EnAll	59.770	70.833	67.708	72.917	67.778
SVM-EnFrAll	27.586	41.667	56.250	45.833	71.111
Human Listeners	67.683	78.476	69.756	81.220	82.012

Far more interesting is the fact that the overall pattern of the human listeners is strikingly similar to the EnAll SVM. AllMod is the least accurate, followed by FormantMod, and the relative confusion

of A1-P0 and AllMod appear fairly comparable.

One specific point of interest, given the complexities with A1-P0 described in 6.8.7, is the distinctly human-like reduced accuracy of A1-P0 and AllMod in both the control *and* the experimental stimuli for all three machine learning models. This would seem to suggest that whatever caused the increased difficulty in classifying A1-P0 modified stimuli (A1P0Mod and AllMod) for human listeners also affected the SVMs, and serves as a sort of replication of an already odd phenomenon.

Duration modification tells an interesting story here. In Section 5, we found that Duration was a *very* highly ranked and useful feature for nasality in English, far more so than its correlations (and the literature) suggested. Thus, when duration is modified, it appears to hurt accuracy in the English-based models rather more than for the humans.

The exceptional performance of duration for the EnFrAll is explained by the combination of the two datasets: Because duration varies with nasality *in opposite directions* for English and French, a model which gives equal credence to the pattern in both languages will end up assigning the feature relatively little predictive weight. Thus, when presented with misleading duration information, it will be less affected than a model which makes strong predictions for duration.

However, the same lack of single-language consistency that benefitted the EnFrAll model for Duration caused problems for the other features. A3-P0 modified stimuli proved particularly problematic for the model, likely because French's rather strong spectral tilt effects for nasality trained the model to weight tilt more heavily than the English-only models. A1-P0 modified stimuli showed poor performance as well, as the consistency of A1-P0 across the two languages, the model likely weighted it rather strongly, resulting in poor performance on modified data. Given all of these sources of difficulty, the AllMod stimuli performed abysmally, well below chance[44].

So, in summary, although none of the models could match human perception, our best performing model, as well as the one which most closely paralleled human perception, was the English model trained on all data. Within that model, as with for the human listeners, modification of all tested features was the most "confusing", followed by the modification of formants.

7.4 Experiment 4 Discussion: Human vs. Machine Perception

Although we must be cautious in interpreting these results too strongly, given the relatively few stimuli and algorithms tested, we can at the very least evaluate our final hypothesis and discuss some useful parallels between the machine data and our human listeners.

7.4.1 Hypothesis 7: Human and Machine Perception

In Hypothesis 7, we predicted strong perceptual parallels between SVMs and humans:

[44]Interestingly, 27.6% corresponds to 72.4% accuracy *in the opposite direction*, indicating that when trained with French, the classifier is in fact rather good at identifying oral-nasal differences in these stimuli, albeit with the improper label. This suggests that although it does not properly label the pattern, some French training data does help to find *a* pattern.

Hypothesis 7 – *When given the stimuli from experiment 3, Support Vector Machines will show the same patterns of variation in by-condition accuracy as the human listeners.*

Although this is somewhat difficult to evaluate, when we rank the relative accuracy of each condition from highest ("1") to lowest ("3"), we do see some similarities, particularly when comparing humans to the EnAll SVM:

Table 48: Accuracy rankings by Condition for Human and SVM perception of experimental stimuli

	DurationMod	A3P0Mod	A1P0Mod	FormantMod	AllMod
Human Listeners	1	2	3	4	5
SVM-EnAll	3	1	2	4	5
SVM-EnNoNVN	2	1	4	3	5
SVM-EnFrAll	1	3	4	2	5

As we've known since our first machine learning experiments, our English-based models have a very strong affinity for duration, which is not shared with our listeners. If we consider the features *without* duration, the EnAll SVM model ranks the remaining features *identically* to our listeners:

Table 49: Accuracy rankings by Condition for Human and SVM perception of experimental stimuli (without duration)

	A3P0Mod	A1P0Mod	FormantMod	AllMod
Human Listeners	1	2	3	4
SVM-EnAll	1	2	3	4
SVM-EnNoNVN	1	3	2	4
SVM-EnFrAll	2	3	1	4

In addition, some idiosyncrasies of the data are shared across the human and machine perception. Although it is far from conclusive, the fact that the SVMs *all* picked up on the unexpectedly poor accuracy across control *and* experimental A1-P0Mod stimuli is rather compelling.

So, although the sparse nature of the data does not permit rigorous statistical testing, based both on the accuracy rankings, the similarities in degree of accuracy reduction across conditions do appear to support this hypothesis, and do strongly suggest that SVM perception is analogous to human perception in this experimental task.

7.4.2 Discussion: Learning from Machine Learning

There are four important lessons suggested by the results of this experiment.

First, these findings provide a rather peculiar sort of replication of the results in Experiment 3. Once again, formants appear to be an important cue for nasality, and when nasal characteristics of formants are removed from the signal (with or without other cues), accuracy in classification

suffers. Although some of the machine-learning specific idiosyncrasies showed through, particularly for duration, the data is more similar than different, and suggest that the patterns discovered previously, even the unusual ones, are at least plausible.

Second, this finding allows us to feel somewhat more confident in our choice of features for human testing. However cautiously the decisions were made, we ended up testing only four of the 29 features evaluated in the course of this paper *in actual human perception*. Although conventional statistical analysis helped to inform this choice, the results of the machine learning studies in Section 5 provided the final evidence for human feature selection. Had the output of our machine learning models shown *no* similarity to human perceptions, a large part of our feature selection methodology becomes rather tenuous. The fact that the machine learning models *did* seem to reflect human perception, albeit with some noise, allows us to feel more confident that our feature selection was reasonable, and we did test the most promising out of our original 29 features.

Third, we see that the specificity of training data must be chosen carefully when classifying nasality. The most accurate (and most human-like) model resulted from training not with the most specifically similar dataset (EnNoNVN), nor with the most extensive dataset (EnFrAll), but with the structure-general but language-specific EnAll dataset. This further bolsters the argument that the perception of nasality is language specific, and again suggests that we cannot model "nasality" as a language-independent acoustical phenomenon (at least, not based on the data from these two languages).

Finally, the results of this experiment showed that a properly-trained model did, at least in this case, roughly approximate the judgements of human listeners, and that those features which did not significantly affect machine perception were not terribly important to humans, either. Although we must be on the lookout for idiosyncrasies (like the strong role of duration in the data), the relative alignment of SVMs and living breathing humans bodes well for our hypotheses about perceptual cues to nasality in French, and suggests that a human perception experiment using a similar methodology and feature set may find some success there as well.

More generally, the present work represents a victory, however small, for the idea that machine learning and classification can provide useful information about the nature and process of speech perception in human listeners. Although the agreement between the two was by no means complete, by using machine learning, we were able to predict the utility of different features for human perception, and to efficiently reduce our feature set to a more manageable size.

This suggests that machine learning in general may prove to be a viable tool for the study of human speech perception.

8 General Discussion

So, after collecting and measuring 4,778 vowels, performing statistical analysis and machine learning studies on 29 acoustical features, and conducting a perception experiment with responses from 42 actual humans, we can now evaluate, in a more integrated way, both the findings of the present research, and their consequences for those of us who work with nasality.

We will start by summarizing what we have learned about the acoustics of vowel nasality in French and English, examining these acoustical findings in general terms by considering both the statistical and machine learning findings.

Then we will focus on the perception of nasality, first summarizing the results of our English experiment, and then using these results (coupled with our machine learning information) to form hypotheses about the nature of nasality in French.

Finally, with these results sorted, we can discuss what they mean in a larger context. We will first discuss the implications of this work for those who aim to measure nasality acoustically in natural language, interpreting the acoustical and perceptual results to make recommendations for future work. Then, we will discuss the findings in terms of speech recognition and machine classification. Finally, we'll discuss the implications of this study for those who work with the phonetics and phonology of nasality in general.

8.1 On the Acoustics of Nasality

Although this study is not intended as an exhaustive study of nasal acoustics, using relatively few speakers, words, and languages, the data presented here do provide some useful information for the wider canon of nasality literature.

8.1.1 On the Acoustics of Nasality in English

In our evaluation of 29 features (shown in Table 5, comparing vowels in oral and nasal coarticulatory contexts across 10 vowel qualities and 12 speakers, we found significant statistical relationships with nasality for 19 of the features (shown in Table 6). However, not all of these correlations were particularly *meaningful* for research, for machine classification, or for perception, whether due to small oral $\rightarrow$ nasal ΔFeature values, inconsistent token-by-token performance, or limited utility outside of certain vowel contexts.

When considering both the aggregate, whole-dataset relationship with nasality, as well as the granular, token-by-token predictive power shown in machine learning, we can isolate the four most strongly correlated and predictive acoustical consequences of nasality.

A1-P0, a comparison of the Amplitude of F1 with a low-frequency nasal peak, proved very useful, meriting its prominent position in the nasality literature. The ΔA1-P0 in oral-to-nasal contexts appears to be caused more by **a reduction in F1's amplitude** than by a rise in P0's (although P0's Prominence relative to the surrounding harmonics does show some promise as a feature of nasal vowels).

In addition to a strong showing in the statistical studies, **F1's bandwidth** proved particularly useful for classification, showing the strongest classification power in both single-feature models and as part of a larger composite model.

Spectral Tilt, here measured by A3-P0 (the amplitude of F3 minus the amplitude of the nasal peak), showed a strong and robust relationship with nasality in English, although it proved to have less predictive power than A1-P0 and F1's bandwidth.

Finally, in these data, **Vowel Duration** was a *very* strong predictor of vowel nasality in our machine learning studies, with nasal vowels showing shorter durations than corresponding oral vowels.

So, we find that in English, the strongest correlates of nasality are in the F1 spectral region, with spectral tilt and vowel duration as useful secondary points of measurement.

8.1.2 On the Acoustics of Nasality in French

One particularly interesting finding of the present work is that we actually *need* two different language-specific sections to discuss the acoustics of nasality. Even though English and French speakers are using the same anatomical structures to create similar phonetic phenomena, the acoustical consequences of nasality are indeed different.

Like English, French has strong relationships between nasality and A1-P0, F1's bandwidth, spectral tilt and duration. However, based on our data, consisting of oral-nasal vowel pairs collected from 8 speakers, we see that the *degree* of oral-to-nasal shift is different in French, as are their utility for classification.

Again, in French, 19 of the 29 features showed significance (although only 12 of those 19 were significant in both French and English). But, again, only a subset proved particularly promising.

In French, **Spectral Tilt** was *the* most important acoustical feature of nasality, by large margins. When measured using A3-P0, spectral tilt showed the strongest oral-to-nasal ΔFeature of any of the measured feature (relative to variation in oral vowels). More impressively still, it bested every other feature in terms of accuracy in single-feature machine learning models, as well as in importance for multi-feature models, beating out even A1-P0. We should also note that in our data, the *degree* of spectral tilt was much higher than in English, and higher than a simple increased degree of nasality would suggest.

In French, **A1-P0** was still strongly linked with nasality. Placing 2nd in both strength of ΔFeature and in terms of classification power, A1-P0 is without a doubt linked to nasality. Given that French lacks the high vowels in which A1-P0 struggles, the increased strength of the feature relative to English is not surprising.

F1's Bandwidth again played a strong role in classification (ranked as the 3rd most important feature), and showed sizable ΔFeature values. Thus, we cannot discount F1's bandwidth as a potential perceptual feature in French.

Finally, French showed an opposite pattern in **Duration**, with phonemically nasal vowels being *longer* than their oral counterparts (versus in English, where nasalized vowels are shorter). French vowels also showed more *consistent* duration changes relative to variability in oral vowels. This difference again proved particularly useful for automated classification.

So, although many of the same features proved statistically related to nasality in both English and French, and the same features can be effectively used to predict nasality, they appear to be *differently* useful. Each showed different ΔFeature values in the two languages, and as a result, machine learning models needed to be trained on language-specific data, for accurate classification. Some potential reasons for these differences are discussed in Sections 4.8 and 5.9, but given the difference in the phonological status of nasality between English and French, it's not unreasonable to expect differences in how speakers (and perhaps listeners) implement nasality.

8.2 On the Perception of Nasality

The statistical studies described here provided an excellent foundation for our understanding of the perception of nasality. From them, we learned which features were unlikely to be perceptually useful cues, namely, those which showed *no* statistical or predictive link with nasality. The machine learning studies built on this understanding, allowing us to see, in a token-by-token "perception" task, which of the features actually provided useful information for distinguishing oral and nasal vowels, and suggested promising features for evaluation.

However, the fundamental question of this work is based not in math or statistics, but in human language: What acoustical feature or features in speech do humans hear as "vowel nasality"?

8.2.1 The Perception of Nasality in English

Here, we will again summarize the findings more clearly described in Section 6.8. Our human perception experiment examined the perception of /ɑ/ and /æ/ vowels, as produced by two English speakers in CVC, CVN, and NVC words. To isolate the cue or cues responsible for the perception of nasality, we tested all four of the important features discussed above (A1-P0, spectral tilt, duration, and vowel formants), manipulating each feature by adding it to oral vowels, or reducing it in nasal vowels, along with a subset of vowels where all features were modified.

Across 42 different listeners, the *sole* feature modification which had any statistically significant effect on the perception of vowel nasality was the modification of **formant frequency and bandwidth** to match oral or nasal means.

Although the three formant properties tested (F1's Frequency and Bandwidth, as well as F3's bandwidth) were modified simultaneously in the experimental stimuli, based the phonetic nature of the modifications, evidence from machine learning, as well as prior experimental work (Hawkins and Stevens (1985b)), it appears that the changes to F1's bandwidth most likely caused the changes in listener perception.

Also of note is that formant modifications only affected listener oral-nasal confusion *in oral vowels*. Although reaction time was increased when formant cues were reduced in nasal vowels, listeners

still consistently classified the resulting stimuli as nasal, suggesting an interesting asymmetry in cue usefulness for vowel perception, and a potential perceptual role for the non-nasal aspects of nasal vowel articulations.

8.2.2 Predictions for Nasality Perception in French

The information gained here does allow some hypotheses about French vowel perception to emerge.

First, given the strength of the English finding, it would be somewhat surprising if formant bandwidth *didn't* play some role in the perception of nasal vowels in French. There's little reason to believe that the already-demonstrated French oral-to-nasal bandwidth changes will be completely ignored by listeners, even if it is supplemented by other cues. And, given that a similar bandwidth-like finding is reported for Gujarati (where nasality is also phonemic) in Hawkins and Stevens (1985a), we have even less reason to suspect that formant bandwidth's utility is limited to English.

However, given the similarity of the English perceptual data to the machine learning models, we would be wise to anticipate spectral tilt playing *some* role in vowel perception. The strength of A3-P0 as a predictor for nasality in both the SVM and RandomForest models cannot be overlooked, and the degree of ΔA3-P0 in French could very easily permit spectral tilt to be a more viable perceptual cue than in English.

In addition, we should remember that spectral tilt and formant bandwidths are *not* incompatible cues. By necessity, increased spectral tilt will reduce formant amplitudes, particularly in the higher frequencies, which will manifest as increased bandwidth (which was already visible in our production data). It is very possible that the two phenomena reinforce each other, perhaps to the point where it becomes difficult to de-convolve them and isolate a single perceptual cue.

Given the similarity of degree of the French and English A1-P0 effects, and the fact that manipulating A1-P0 for English speakers does not appear to have affected perception at all, it seems unlikely that French speakers will use A1-P0 as a cue in perception. However, there were some complexities in our modification of A1-P0, so A1-P0 does likely merit re-testing in French with improved methods, even if just to confirm the original result.

Duration did show major oral-to-nasal changes in French, and did perform quite well in machine learning. But given that duration played no measurable role in English perception of nasality, despite the strength of effect in the machine learning studies, it's not unreasonable to suspect that machine learning algorithms simply have a weakness for duration, and overestimate the usefulness of the feature. Coupled with duration's vulnerability to common sources of variability such as speech rate, sentence planning issues, and prosodic modification of duration, although it may be worth testing, placing major bets of duration as a useful cue in French is unwise.

Finally, we should be on the lookout for the effects of oral articulations on nasal vowels, particularly given the known and measured differences in the oral articulations of French vowels (c.f. Carignan (2014), Carignan et al. (2015), Carignan et al. (2011), and Delvaux et al. (2002b)). It is likely that we would see a similar asymmetry if these oral articulation changes are not accounted

for, and given the present finding, it may be reasonable to explicitly test the perceptual role of these oral articulation differences.

The role of oral articulations in nasal vowels in French is particularly interesting, given the phonemic nature of nasality in French. One possibility, as suggested above and in the literature, is that we would see even stronger roles for oral articulation in the perception of nasal vowels. With the increased importance of nasality in French, we might expect greater perceptual reliance on the contrast-enhancement of oral articulations in the French data. This reliance could even be strong enough that the formant structure of oral vowels becomes useful to classify "oral" as a category, and thus, analogous to the nasal vowels in English, the modification of nasality results in no change in classification.

However, this same heightened phonological importance of nasality in French could also lead to an increase in the importance of cues to nasality. Unlike English listeners, who always have an initial or trailing nasal consonant, listeners of French need to regularly distinguish between oral and nasal vowels *in isolation from other nasality*. We can imagine a world in which, although nasal and oral vowels differ in articulation, we might want stronger secondary evidence of nasality or orality, which is not as vulnerable to token-by-token changes in oral articulation. French listeners, for whom this is an everyday task, may develop more perceptual nuance for nasality, and take advantage of cues to the acoustics of nasality which were not used by our English speakers. So, although we expect a strong role for oral articulations in French nasality perception, it's not unreasonable to suspect a stronger role for the acoustical features of nasality itself as well.

Of course, the only way to test these hypotheses is by running a similar experiment with French listeners and words. In many ways, the task is easier: French speakers are aware of nasality as a phonological phenomenon, and stimuli can be presented in a simple forced choice lexical identification task ("beau" or "bon"?). We might also assume that French speakers have stronger intuitions and more definitive judgements of nasality than English speakers, given the importance of nasality in the language. In addition, given that the nasal vowel space in European French has only three contrastive qualities, perception can be tested in all applicable contexts. Knowledge of the nature of the differences between English and French nasality perception, will provide considerable benefit, both to our understanding of French, and of nasality in general.

However, even without this information, the results of the present study have considerable implications for the measurement, classification, and study of nasality in general.

8.3 On the Acoustical Measurement of Nasality

In many ways, the question of "How do we best measure nasality?" is what motivated the questions which eventually became this paper. Dissatisfaction with existing approaches led to the search for better ones, which, in turn, led to questioning how *any* entity, man or machine, could detect nasality without resorting to magic.

Here, with the final acoustical and perceptual data for our 29 features in mind, we will discuss the problem of measuring nasality more explicitly, discussing features which demonstrate utility for the measurement of nasality, features which should be used with caution, and general difficulties

of nasality measurement.

8.3.1 Useful features for the Measurement of Nasality

A1-P0, codified by Marilyn Chen (Chen (1997)), is the standard for measuring nasality in the literature. Although it is not without its quirks and inconsistencies, particularly when F1 and P0 are in close proximity, the results here agree that it is among the best measures of nasality in current use.

It is a relative measure, thus avoiding the pitfalls of absolute amplitude measurements (such as F1's amplitude alone). It's based on easy to find landmarks (F1, and the highest of the first two harmonics), and with considerable caution, can be measured automatically.

It also captures a large swath of what seems to make nasal vowels "nasal". The measure directly captures the reduction in F1's amplitude due to the lowest nasal zero, which is either the cause of, or a consequence of, F1's increase in bandwidth. In addition, A1-P0 captures the nasal resonance, P0, and captures, albeit less distinctly,some of the spectral tilt associated with nasality (as tilt will also lower F1's amplitude relative to the first harmonics). In fact, of our four major features, duration is the only one which is not in some way touched by A1-P0.

It's also worth pointing out that, although several of the features tested here were specifically designed to capture similar phenomena using different baselines or points of comparison, plain A1-P0 consistently showed the strongest correlations with nasality, as well as the best predictive power. All of this is to say that, even if it isn't perceptually useful, A1-P0 appears to be a very good tool for measuring nasality. But it is not without downsides.

First, one must be careful with vowel choices. As we found in Section 4.5.7, A1-P0 performs very differently in high vowels than in low vowels, and is mathematically impossible to measure when F1 overlaps the first few harmonics. In addition, the "vowel compensation" formula proposed in Chen (1997) was not universally helpful, although it did outperform the "vanilla" A1-P0 measurement in places. So, as good as it is, A1-P0 is not universally good, and care must be taken with the words and vowels to which it is applied.

Perhaps more dangerous for phonetic research is the cross-speaker variability of A1-P0. As described in Section 4.7.2, A1-P0 varies across speakers not just in terms of baseline (a speaker's "nasal" versus "oral" values), but in terms of ΔA1-P0, such that the same change in expected nasality might produce a much larger measured change in A1-P0 for one speaker than for another. Based on these data, it appears that A1-P0 should be treated in the same way as raw or Z-Scored vowel formant measures: *within-speaker* comparison is completely reasonable, and overall trends can be discussed (e.g "nasality increases in hard words"), but direct *across-speaker* comparisons of values are simply not reliable[45].

P0's Prominence, the height of the higher of the first two harmonics relative to the harmonics to either side, did show some link with nasality, and has the advantage of being fairly robust to spectral tilt effects (particularly if you suspect that the tilt is coming from nasality-external factors

[45]Establishing some sort of algorithm for the normalization of A1-P0, perhaps based on across-speaker comparisons with a known-comparable measure such as airflow %nasalance, would be a particularly useful innovation.

like voicing quality). However, it proved a poor predictor of nasality, and practically speaking, by finding P0's prominence, one has already accomplished the most difficult part of measuring A1-P0. So, although it may have some merit as a backup or secondary measure, it should not be the first choice.

Finally, **Formant bandwidth**, although not used directly for measuring nasality in the literature, shows considerable promise. Formant bandwidth shows strong whole-dataset correlations and predictive power, it makes acoustical sense (given the increased thermal and surface loss of energy with increased cavity volume), and it appears useful for human perception. Thus, it seems a good candidate for nasality measurement.

Of course, vowel formant bandwidths can vary based on non-nasal articulatory changes as well, However, in these data, the oral-to-nasal ΔBandwidth relative to the oral variation is quite strong, similar to the strength of A1-P0, indicating that in the general domain, formant bandwidths could also be used for nasality measurement. The use of F1's bandwidth as a measure deserves further investigation, particularly with regards to accuracy, cross-speaker differences, and interference from other phenomena, but it could prove particularly useful for studies where nasality is being measured to make listener-centric or perceptual claims (such as in Scarborough (2013), Scarborough (2012), or Beddor (2009)). Although A1-P0 and Bandwidth both appear to track nasality, when investigating listener-directed issues, it makes more sense to work directly with what we now know to be the listener-relevant acoustical cue, at least in English. So, although some further investigation into interactions and cross-speaker variability is merited, formant bandwidth appears to be a very reasonable direct metric for nasality.

8.3.2 Questionable features for nasality measurement

Of course, not all features evaluated showed strong statistical or predictive links with nasality. Rather than discussing each, we'll focus on a few features which proved less useful for measurement than anticipated, and discuss their pitfalls.

First, a note on **Spectral Tilt**. Despite the strength of its statistical relationship to nasality, and its predictive power, particularly in French, spectral tilt is strongly affected by many non-nasal factors in speech, the most common being voicing type (c.f Laver (1980), Gordon and Ladefoged (2001)) and vocal pathology (c.f. Murphy et al. (2008))[46].

Although it may be useful as a secondary feature of nasality for measurement, particularly in languages where it shows a stronger ΔFeature with nasality, even with the strong finding in French, we cannot in good faith call A3-P0 a specific "measure of nasality".

Similarly, although it showed strong effects for nasality, **Duration** is not a particularly good measure of nasality. Putting aside the effects of speech rate and vowel type, and the fact that all comparisons would have to be relative to other vowels, measurement of duration necessarily involves the segmentation and labeling of vowels. This is not an easily automated process, and particularly in coarticulatory contexts, requires considerable practice and skill to do consistently.

[46]So much so that A3-H1, corresponding almost directly to A3-P0, is a measure often used in the voice quality literature.

Thus, although differences in vowel duration should be attended to in the creation of stimuli where meaningful, duration should not be used for measurement.

Finally, **A1-P1**, which compares F1 with a higher frequency nasal peak around 950 Hz, proved useful only for high vowels, and even then, was only marginally more effective than A1-P0. More importantly, in all non-high vowels, including the entirety of the French dataset, A1-P1 was *not significantly correlated with nasality at all.*

Although A1-P1's poor performance could be a consequence of the measurement process, if we reliably "missed" the P1 harmonic, this itself is a condemnation of the measure. Even for human researchers, the "P1" peak is difficult to find for some speakers, even in the high vowels where the measure is most useful. So, although it is possible that there are some speakers in some languages for whom the hand measurement of A1-P1 in high vowels only could be useful, based on the present work, the utility of A1-P1 does not justify the difficulty involved in its measurement. For high vowels, one should use A1-P0 (when F1 != P0) or formant bandwidth instead.

8.3.3 General Notes on Nasality Measurement

Three final notes on the acoustical measurement of nasality.

First, based on these data, acoustical nasality measurement is simply not ready for clinical or diagnostic use. Although a straightforward measurement of nasality from sound alone would be a massive boon to the speech pathology community, such a measure does not (yet) exist. Even evaluating 29 different features with a hand-tuned model trained on several thousand tokens, the best oral vs. nasal machine classification accuracy rate possible for English was 84%. When dealing with children (whose speech is something different entirely) or pathological speech, we have no reason to expect any improvement. Although this is not unreasonable for speech recognition, this is far less than the accuracy needed for reliable diagnosis of pathology. So, unfortunately, it seems that the state of the art is still as described in Vogel et al. (2009) and Buder (2005): although acoustical detection of pathological hypernasality could be possible in some cases, our current measures are not reliable nor trustworthy for judging nasality in a vacuum, particularly when more accurate means of measurement like nasometry or pneumotachography exist.

Second, across-speaker comparison of *degree* of nasality based on acoustical measures does not appear reliable, given the current state of the art. All of the viable measurements of nasality show considerable across-speaker variability both in baseline and oral-to-nasal range, and until additional work can be done to characterize the nature of these differences and develop a feature-sensitive normalization method, any such comparisons lack a solid foundation.

Finally, based on the differences between English and French and the nature of the measures themselves, *nasality measurement should be tailored to the task, dataset and language being investigated.* There is no such thing as a one-size-fits-all nasality measurement. Each study should begin with a frank discussion about the needs of the project, phenomenon to be characterized (production or perception), and the constraints of the language (are high vowels a factor?). Then, the data should be examined in detail along with the measurement of multiple features, to see which features are most characteristic of nasality *in these particular speakers*.

As with any phonetic study, the acoustical study of nasality requires caution, control, and careful attention to variability. But, based on these findings, the approaches in wide use are quite reasonable, and accurate measurement is certainly possible.

8.4 On Machine Classification of Nasality

Rather than re-framing all of the above measurement and classifications finding for machine-learning, we will let the data stand on their own, and instead make three smaller points.

First, based on this study, the best features for automated classification of nasality are not terribly complex. Both formant bandwidth and spectral tilt, our top performers for English and French respectively, can be extracted straightforwardly, and at many points in a given vowel. Common signal processing algorithms (such as LPCs, DCTCs or Cepstral Analysis) which are already used to capture vowel features will detect both features reliably. Moreover, those features tested which relied on finding particular harmonics proved relatively less useful for classification. So, it appears that relatively little would be gained from attempting to isolate "nasal peaks" during feature extraction.

Second, models of nasality will need to be trained for each dataset and language. As we found in both Experiments 2 and 4, a model trained to detect nasality in English will perform very poorly on French, and vice versa. Although perhaps a "universal" pattern could emerge with sufficient data from enough different languages, based on the findings here, the best results will come from training on a broad, single-language dataset.

Finally, we see that the machine classification of nasality is a tractable problem, even without any context-awareness, and based on a very limited feature set. Given that our machine learning models showed 84% accuracy in English and 94% accuracy in French, clearly, there is sufficient information in the signal to classify nasal vowels with some accuracy without reference to context, frequency, or other classification shortcuts.

In short, given the relative simplicity of the useful features involved, as well their strong predictive power, so long as models can be trained on a wide and representative dataset, the automated classification and identification of vowel nasality should not present major difficulties.

8.5 On the Linguistic Study of Nasality

Finally, a brief discussion of the meaning of these results to linguistic research on nasality in general.

One could easily picture a world in which, perceptually and acoustically speaking, vowel nasality is mostly about nasality. In such a world, perception of nasality would be carried by nasality-specific peaks or zeroes. Individual resonances caused by the coupling of the nasal cavity would lead to nasality-specific peaks, like P0, P1, or P2, along with detectable zeroes in the signal. The perception of nasality would be listening *past* the vowel, in an effort to find these nasality-specific cues which, independently of the vowel, trigger nasal perception.

However, based on the results of these experiments, this is *not* the case. No such "bellwether" features exist for vowel nasality, which instantly and independently scream "nasal" for listeners, nor for machine models. None of the "nasality specific" peaks (P0, P1, or P2) proved useful for classification, and indeed, the sole relevant features, the vowel's formant structure, also carries vowel quality, consonant place, and many other contrasts in speech.

Certainly, the widening of formant bandwidths relative to oral vowels is likely due in large part to the additional surface area and volume associated with the vocal tract following oral-nasal coupling. However, these widened bandwidths are primarily interpretable in comparison to oral vowels, and are strongly affected by the oral articulations associated with nasal vowels. Thus, we cannot examine a vowel's "nasality" in a vacuum.

This hints nicely towards the idea that the perception of vowel nasality is at least partly grounded in oral vowel articulations. We've discussed this already in terms of the perceptual asymmetries in Experiment 4 (see Section 6.8.8), but it's useful as a wider concept as well. Recent studies, such as Carignan et al. (2015), Carignan (2014), Carignan et al. (2011) and Shosted et al. (2012), all highlight the acoustical consequences of these oral articulation shifts, framing them explicitly as an attempt to "enhance the contrastiveness" of nasal vowels.

For instance, Carignan et al. (2015) focuses on the effects of "vowel nasality" on F1 and F2, finding that the oral articulations produced during French nasal vowels help to enhance the formant changes associated with nasal vowels. In this work, using Electromagnetic Articulography and acoustics, the author is able to attribute nasality-related changes in F1 and F2 to the simultaneous and concerted effort of several different articulatory phenomena (VP port aperture, lip rounding, pharyngeal changes, and tongue movement). This makes a strong argument that nasal vowels, at least in French, involves distinct and non-arbitrary oral articulations from oral vowels, and more importantly, that these articulations all serve to enhance nasality-related F1 and F2 changes.

Although the answer does not lessen Carignan's argument, there remains somewhat of a chicken-and-the-egg problem: are French speakers enhancing a pre-existing formant contrast associated with nasality *per se*, or are French speakers *creating* a formant contrast associated with nasal vowels, to enhance perceptibility of nasal vowels?

In our data, we found little evidence of formant changes *due to nasality itself* (across vowels and contexts), which seems to indicate that simply opening the VP port does not *necessarily* result in a meaningful formant change. Similarly, with the possible exception of F1's bandwidth, no features which are direct consequences of velum lowering proved perceptually useful for English speakers, again casting doubt on the idea of "inherent nasal cues". However, we also have found a strong by-vowel set of formant changes *associated with* nasality. Based on our perceptual asymmetry in the English data, these changes appear to be sufficient for the perception of nasality, even when the acoustical cues associated with nasality itself are removed.

This points towards a view of nasality in which nasality does have some acoustical consequences and cues (like changes to F1's bandwidth in English), but where speakers enhance the perceptibility of nasality in their speech by producing nasal(ized) vowel specific oral articulations, which result in nasality-specific effects on vowel formants. Put differently, the oral articulations of nasal vowels are not necessarily enhancing *an inherent aspect of nasality*, but enhance the perceptual

distinctiveness of specific nasal vowels *relative to oral vowels*[47].

So, based on the weak finding for inherently "nasal" cues to nasality, on the perceptual asymmetry between reduction and addition stimuli, and on the preexisting literature on differing patterns of oral articulation for nasal vowels, we should consider nasal and nasalized vowels as perceptually and articulatorily separate entities from their oral counterparts.

Such a perspective is helpful in a larger linguistic context as well. First, it offers a straightforward phonetic explanation for the patterns of vowel quality shift between nasal vowels and their "oral counterparts" seen around the world. If the oral nature of nasal vowels is a useful perceptual cue, nasal vowel quality shifts away from the oral vowel can be viewed as necessary and meaningful, rather than as noise or as necessary consequences of oral-nasal coupling. Similarly, the idea of nasal vowels having distinct oral identities makes more understandable the diachronic tendency of nasal vowel systems to develop and shift independently of oral vowel systems. Put differently, although thinking about nasal and nasalized vowels as independent of their oral counterparts vowels is perceptually useful in light of these results, it also may be helpful to broader phonetic and phonological theory.

Yet, the same asymmetry in these findings that suggests a role of oral articulation prevents us from dismissing "nasality" as an acoustical concept altogether, at least in our English data. If vowel nasality were purely a matter of oral vowel quality, we would expect none of our modifications to have had any effect on classification at all. A vowel which is produced with an "oral" articulation would forever remain "oral", and modifying only one aspect of the vowel would introduce little or no confusion. Instead, we see that while there appears to be more to nasal vowels than nasality, listeners do attend to *some* acoustical consequences of nasality, and the addition of a nasal cue to an oral *can*, although not with 100% agreement, cause listeners to reclassify the vowel as "nasal".

So, although we now understand that vowel formants, specifically formant bandwidths, are useful and meaningful cues to the perception of nasality, the story is slightly more complex.

Taken as a whole, the findings of the present work all seem to indicate that in English, vowel nasality is about both the vowel *and* the nasality. Nasal or nasalized vowels are not just oral vowels plus the acoustics of an open VP port, but are unique vowels in their own right, with specific oral articulations alongside nasal acoustics. More simply, there is more to [ã] than /a/ with a lowered velum.

[47]There is strong analogy to other situations where a secondary cue is "recruited" to enhance an unrelated contrast, e.g. vowel duration for English VOT, or F_0 as an indicator of a weakening consonant contrast

9 Conclusion

In this study, we set out with one goal: to find perceptual cues for vowel nasality in English.

To this end, we collected two corpora of elicited speech data, one in English and one in French, then measured and analyzed 29 different acoustical features in both languages. By using statistical analyses (linear mixed-effects regressions) and machine learning analyses (RandomForest and Support Vector Machines), we evaluated the statistical and predictive links between each feature and nasality. Through this process, although their degrees of change and utility differed between the English and French data, we were able to narrow the field to just four promising features: A1-P0, Duration, Vowel Formants, and Spectral Tilt.

These four features were then tested in both /ɑ/ and /æ/ vowels, enhancing them in oral words in an attempt to "create" nasality, and reducing them in coarticulatorily nasalized vowels, to remove the perception of nasality. These stimuli were then presented to 42 Native English speakers in a lexical choice task with phoneme masking, and then analyzed on the basis of reaction time.

Of the four features tested, only manipulation of vowel formants showed *any* statistically significant effect on listener perception of vowel nasality. When F1's frequency was changed by 36 Hz (increasing in nasal words), and F1 and F3's bandwidths were adjusted to oral or nasal norms, listeners showed a significant increase in reaction time across the board. In addition, formant modification showed a significant increase in oral vs. nasal confusion, but *only* when manipulating oral tokens. The classification accuracy of nasal vowels was not affected by any manipulation of nasal correlates, suggesting that English speakers may also be attending to other, non-nasality-caused differences in nasalized vowels.

Finally, the same machine-learning classifiers used to narrow the field were given the experimental stimuli. With the exception of duration, which was consistently over-used by classifiers, the SVM classifiers and the human listeners showed similar patterns of confusion. This reinforces the human perceptual result and allows for more concrete predictions about the perception of nasality in French, which was not tested with humans in this project.

These findings point towards the idea that the perception of vowel nasality is not about identifying nasal-resonance-specific elements of the signal, but instead, involves comparing the entirety of the vowel signal to one's perceptual model of a nasalized vowel. In this process, although perceptual cues like vowel formant bandwidth can provide positive evidence of nasality, they are not the *sole* evidence for nasality. Additional cues, particularly nasal-vowel-specific oral articulations, can also play a very strong role in listener classifications.

This study is not the final word in research on nasality perception in English, and merits reproduction and improvement, particularly using a greater variety of speakers, listeners, and vowels. However, based even on these limited data, we can answer our initial question.

Of the features examined, it appears that vowel formant structure, particularly F1's bandwidth, is the primary cue to the perception of nasality in English. However, we also see evidence pointing to a strong perceptual role for the oral articulations of nasal vowels, indicating that nasal vowels are unique and complex phenomena, and that there's more to vowel nasality than nasal airflow.

References

Arai, T. (2006). Cue parsing between nasality and breathiness in speech perception. *Acoustical Science and Technology*, 27(5):298–301.

Atal, B. S. (1985). Computer speech processing. chapter Linear Predictive Coding of Speech, pages 81–124. Prentice Hall International (UK) Ltd., Hertfordshire, UK, UK.

Baayen, R. H. (2008). *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge University Press.

Basset, P., Amelot, A., Vaissière, J., and Roubeau, B. (2001). Nasal airflow in French spontaneous speech. *Journal of the International Phonetic Association*, 31(1):87–99.

Bates, D., Maechler, M., Bolker, B., and Walker, S. (2014). *lme4: Linear mixed-effects models using Eigen and S4*. R package version 1.1-6.

Beddor, P. S. (2009). A coarticulatory path to sound change. *Language*, 85(4):785–821.

Beddor, P. S. and Krakow, R. (1999). Perception of coarticulatory nasalization by speakers of English and Thai: Evidence for partial compensation. *The Journal of the Acoustical Society of America*, 106:2868.

Beddor, P. S., McGowan, K. B., Boland, J. E., Coetzee, A. W., and Brasher, A. (2013). The time course of perception of coarticulation. *The Journal of the Acoustical Society of America*, 133(4):2350–2366.

Boersma, P. and Weenink, D. (2012). Praat: doing phonetics by computer [computer program].

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Buder, E. H. (2005). The Acoustics of Nasality: Steps towards a bridge to source literature. *Speech Science and Orofacial Disorders*, (July 2005):4–8.

Bybee, J. (2003). *Phonology and language use*, volume 94. Cambridge University Press.

Carignan, C. (2014). An acoustic and articulatory examination of the oral in nasal: The oral articulations of French nasal vowels are not arbitrary. *Journal of Phonetics*, 46(0):23–33.

Carignan, C., Shosted, R., Shih, C., and Rong, P. (2011). Compensatory articulation in American English nasalized vowels. *Journal of Phonetics*, 39(4):668 – 682.

Carignan, C., Shosted, R. K., Fu, M., Liang, Z.-P., and Sutton, B. P. (2015). A real-time MRI investigation of the role of lingual and pharyngeal articulation in the production of the nasal vowel system of French. *Journal of Phonetics*, 50(0):34 – 51.

Chang, C.-C. and Lin, C.-J. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

Chen, M. Y. (1995). Acoustic Parameters of Nasalized Vowels in Hearing-Impaired and Normal-Hearing Speakers. *The Journal of the Acoustical Society of America*, 98(5):2443–2453.

Chen, M. Y. (1997). Acoustic correlates of English and French nasalized vowels. *The Journal of the Acoustical Society of America*, 102(4):2350–2370.

Cohn, A. C. (1990). Phonetic and phonological rules of nasalization. *UCLA Working Papers in Linguistics*, (76).

Delvaux, V. (2006). Production des voyelles nasales en Français Québécois. In *XXVIe Journées d'études sur la parole*, pages 383–386.

Delvaux, V. (2009). Perception du contraste de nasalité vocalique en Français. *Journal of French Language Studies*, 19(1):25–59.

Delvaux, V., Demolin, D., Harmegnies, B., and Soquet, A. (2008a). The aerodynamics of nasalization in French. *Journal of Phonetics*, 36(4):578–606.

Delvaux, V., Demolin, D., Harmegnies, B., and Soquet, A. (2008b). The aerodynamics of nasalization in French. *Journal of Phonetics*, 36(4).

Delvaux, V. and Huet, K. (2006). Perception de la nasalité en Français de Belgique: catégorisation dirigée et catégorisation libre. *Revue PArole.*, (39).

Delvaux, V., Huet, K., Piccaluga, M., and Harmegnies, B. (2012). Inter-gestural timing in French nasal vowels: A comparative study of (liëge, tournai) northern French vs. (marseille, toulouse) southern French.

Delvaux, V., Metens, T., and Soquet, A. (2002a). French nasal vowels: acoustic and articulatory properties. In *Interspeech*.

Delvaux, V., Metens, T., and Soquet, A. (2002b). Propriétés acoustiques et articulatoires des voyelles nasales du Français. In *XXIVèmes Journées d'Étude sur la Parole, Nancy, 24-27 juin 2002*, number 24, pages 357–360.

Demolin, D., Delvaux, V., Metens, T., and Soquet, A. (2003). Determination of velum opening for French nasal vowels by magnetic resonance imaging. *Journal of Voice : Official Journal of the Voice Foundation*, 17(4):454–67.

Dickson, D. R. (1962). An Acoustic Study of Nasality. *Journal of Speech and Hearing Research*, 5(2):103.

Fagyal, Z., Kibbee, D., and Jenkins, F. (2006). *French: A linguistic introduction*. Cambridge University Press.

Fougeron, C. and Smith, C. L. (1999). French. In *Handbook of the International Phonetic Association*, pages 78–81. Cambridge University Press, Cambridge.

Gordon, M. and Ladefoged, P. (2001). Phonation types: a cross-linguistic overview. *Journal of Phonetics*.

Hajek, J. and Maeda, S. (2000). The effect of vowel height and duration on the development of distinctive nasalization. *Papers in Laboratory Phonology V: Acquisition and the Lexicon*, 5:52.

Haspelmath, M. (2005). *The World Atlas of Language Structures*. Oxford University Press, Oxford.

Hawkins, S. and Stevens, K. (1985a). Acoustic and perceptual correlates of the non-nasal-nasal distinction for vowels. *The Journal of the Acoustical Society of America*.

Hawkins, S. and Stevens, K. N. (1985b). Acoustic and perceptual correlates of the non-nasal–nasal distinction for vowels. *The Journal of the Acoustical Society of America*, 77(4):1560–1575.

Hlavac, M. (2014). *stargazer: LaTeX/HTML code and ASCII text for well-formatted regression and summary statistics tables*. Harvard University, Cambridge, USA. R package version 5.1.

Hsu, C.-W., Chang, C.-C., Lin, C.-J., et al. (2003). A practical guide to support vector classification.

Huang, X., Acero, A., Hon, H.-W., et al. (2001). *Spoken language processing*, volume 18. Prentice Hall Englewood Cliffs.

International Phonetic Association, T. (1999). *Handbook of the International Phonetic Association: a guide to the use of the International Phonetic Alphabet*. Cambridge University Press, Cambridge, U.K.; New York, NY.

Kingston, J. (2007). The phonetics-phonology interface. *The Cambridge handbook of phonology*, pages 435–456.

Kingston, J. and Macmillin, N. A. (1995). Integrality of nasalization and F1 in vowels in isolation and before oral and nasal consonants: a detection-theoretic application of the Garner paradigm. *The Journal of the Acoustical Society of America*, 97(2):1261–85.

Krakow, R., Beddor, P. S., Goldstein, L. M., and Fowler, C. A. (1988). Coarticulatory influences on the perceived height of nasal vowels. *The Journal of the Acoustical Society of America*, 83(3):1146–58.

Ladefoged, P. and Johnson, K. (2011). *A Course in Phonetics*. Wadsworth/Cengage Learning, Boston, MA.

Lahiri, A. and Marslen-Wilson, W. (1991). The mental representation of lexical form: A phonological approach to the recognition lexicon. *Cognition*, 38(3):245–294.

Larousse (2014). Larousse dictionnaire de Français en ligne.

Laver, J. (1980). *The phonetic description of voice quality*. Cambridge University Press, Cambridge, England.

Liaw, A. and Wiener, M. (2002). Classification and regression by randomForest. *R News*, 2(3):18–22.

Macmillin, N. A., Kingston, J., Thorburn, R., Dickey, L. W., and Bartels, C. (1999). Integrality of nasalization and F1. ii. basic sensitivity and phonetic labeling measure distinct sensory and decision-rule interactions. *The Journal of the Acoustical Society of America*, 106(5):2913–32.

Maddieson, I. (1980). Wpp, no. 50: UPSID (UCLA Phonological Segment Inventory Database).

Mermelstein, P. (1977). On detecting nasals in continuous speech. *The Journal of the Acoustical Society of America*, 61(2):581–587.

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., and Leisch, F. (2014). *e1071: Misc Functions of the Department of Statistics (e1071), TU Wien*. R package version 1.6-4.

Milne, P. (2014a). Introducing SPLAligner: Comparing automatic and manual identification of the variable pronunciations of word-final consonant clusters in French.

Milne, P. (2014b). *The variable pronunciations of word-final consonant clusters in a force aligned corpus of spoken French*. PhD thesis, University of Ottowa.

Murphy, P. J., McGuigan, K. G., Walsh, M., and Colreavy, M. (2008). Investigation of a glottal related harmonics-to-noise ratio and spectral tilt as indicators of glottal noise in synthesized and human voice signals. *The Journal of the Acoustical Society of America*, 123(3):1642–1652.

New, B., Pallier, C., Ferrand, L., and Matos, R. (2001). Une base de données lexicales du Français contemporain sur internet: Lexique™//a lexical database for contemporary French: LEXIQUE™. *L'année Psychologique*, 101(3):447–462.

Oh, E. (2008). Coarticulation in non-native speakers of English and French: An acoustic study. *Journal of Phonetics*, 36(2):361–384.

Ohala, J. J. and Ohala, M. (1995). Speech perception and lexical representation: the role of vowel nasalization in Hindi and English. *Phonology and Phonetic Evidence. Papers in Laboratory Phonology*, 4:41–60.

Paradis, C. and Prunet, J.-F. (2000). Nasal vowels as two segments: Evidence from borrowings. *Language*, 76(2):pp. 324–357.

Peirce, J. W. (2007). Psychopy–Psychophysics software in python. *Journal of Neuroscience Methods*, 162(1-2):8–13.

Price, S. and Stewart, J. (2013). Nasal harmony fading in Guaraní. In *Canadian Linguistics Association Conference 2013*, page 143. Canadian Linguistics Association, University of Victoria, Victoria, BC, Canada.

Pruthi, T. and Espy-Wilson, C. (2004). Acoustic parameters for automatic detection of nasal manner. *Speech Communication*, 43(3):225–239.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Roche, A. P., Rochet, B. L., Sovis, E. A., and Mielke, D. L. (1998). Characteristics of nasalance in speakers of Western Canadian English and French. *Journal of Speech Language Pathology and Audiology*, 22(2):94–103.

Rositzke, H. (1939). Vowel-Length in General American Speech. *Language*, Vol. 15, No. 2:99–109.

Scarborough, R. (2004). *Coarticulation and the Structure of the Lexicon*. PhD thesis, University of California, Los Angeles.

Scarborough, R. (2012). Lexical similarity and speech production: Neighborhoods for nonwords. *Lingua*, 122(2):164–176.

Scarborough, R. (2013). Neighborhood-conditioned patterns in phonetic detail: Relating coarticulation and hyperarticulation. *Journal of Phonetics*, 41(6):491–508.

Scarborough, R., Styler, W., and Zellou, G. (2011). Nasal Coarticulation in Lexical Perception: The Role of Neighborhood-Conditioned Variation. In *ICPhs XVII*, pages 1–4, Hong Kong.

Scarborough, R. and Zellou, G. (2013). Clarity in communication: "clear" speech authenticity

and lexical neighborhood density effects in speech production and perception. *The Journal of the Acoustical Society of America*, 134(5):3793–3807.

Schwartz, M. F. (1968). The acoustics of normal and nasal vowel production. *Cleft Palate J*, 5:125–140.

Shosted, R., Carignan, C., and Rong, P. (2012). Managing the distinctiveness of phonemic nasal vowels: Articulatory evidence from Hindi. *The Journal of the Acoustical Society of America*, 131(1):455–465.

Simpson, A. P. (2012). The first and second harmonics should not be used to measure breathiness in male and female voices. *Journal of Phonetics*, 40(3):477–490.

Stevens, K. N. (1998). *Acoustic Phonetics*. MIT Press, Cambridge, Mass.

Stevens, K. N., Andrede, A., and Viana, M. C. (1987a). Perception of vowel nasalization in VC contexts: A cross-language study. *The Journal of the Acoustical Society of America*, 82(S1):S119.

Stevens, K. N., Fant, G., and Hawkins, S. (1987b). Some acoustical and perceptual correlates of nasal vowels. *Honor of Ilse Lehiste. Dordrecht, Holland, Foris Publications*.

Vogel, A. P., Ibrahim, H. M., Reilly, S., and Kilpatrick, N. (2009). A comparative study of two acoustic measures of hypernasality. *Journal of Speech, Language and Hearing Research*, 52(6):1640–1651.

Winter, B. (2013). Linear models and linear mixed effects models in R with linguistic applications. *arXiv*, page 1308.5499.

Yuan, J. and Liberman, M. (2008). Speaker identification on the SCOTUS corpus. In *In Proceedings of Acoustics 2008*.

Zellou, G. (2012). *Similarity and Enhancement: Nasality from Moroccan Arabic Pharyngeals and Nasals*. PhD thesis, University of Colorado at Boulder.