Acoustic Measurement in Phonetics: Current practices and future directions

by

Emily Jane Grabowski

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Linguistics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Keith Johnson, Chair
Professor Hannah Sande
Professor Gasper Begus
Professor Gopala Anumanchipalli

Spring 2024

Acoustic Measurement in Phonetics: Current practices and future directions

Abstract

Measurement in Acoustic Phonetics: Current practices and future directions

by

Emily Jane Grabowski

Doctor of Philosophy in Linguistics

University of California, Berkeley

Professor Keith Johnson, Chair

Acoustic analysis is a key component of scientific investigation in phonetics, and offers a flexible approach with significant freedom for researchers to decide what and how to measure. In some cases, methodological variation affects results to the point of qualitatively different conclusions (as in Roettger et al. 2019 and Coretta et al. 2023), which highlights a need for further study of the quality and consistency of acoustic analysis. Current approaches are typically focused on specific phonetic phenomena, and the field lacks a holistic understanding of the usage and development of phonetic methodology in the field.

This dissertation will focus on understanding and evaluating methods of acoustic measurement in phonetics. The first goal is to understand the current state of the field, the second is to quantify the relative performance of acoustic measurements, and the third is to outline key considerations for future development of acoustic measurements and point to promising future directions.

Chapter 1 describes a broad overview of the types of measurement that are possible in acoustic phonetics. This chapter is organized based on how the measurement is derived from speech and outlines key factors influencing the formulation, usage, and adoption of these methods. In particular, this chapter highlights the incredible variation possible when it comes to choosing acoustic measures. Chapter 2 investigates the practical usage of acoustic measurement strategies via a quantitative methodological study. This chapter illustrates a decrease in studies regarding consonants and an increase in those pertaining to vowels, and other results indicate a general lack of confidence in acoustic measurement of consonants. In addition, a relatively small number of acoustic measurement strategies make up most of the total strategies in the study, despite the wide range available.

Taken together, these chapters highlight a fundamental challenge in phonetic methodology - a pattern where a small subset of methods dominate the field despite a wide variety of options for acoustic measurement. This aligns with general intuitions in the field of best practices that are robust but cover a small set of phonetic phenomena of interest, and

supports the conclusion that further development of acoustic measurement is of benefit to the field.

The second section of the dissertation quantifies the performance of acoustic features in phonetics via two experiments. Chapter 3 evaluates how much incorporating context improves acoustic representation. The results of this experiment are two-fold: firstly context does seem necessary for capturing at least some phonetic categories, and secondly due to long-range context effects, selecting a model should be done with care and attention. Chapter 4 broadens this methodology to create a general benchmark for quantifying the performance of acoustic features. This benchmark and experimental design provides a framework for quantitatively evaluating the performance of acoustic representations, and is positioned to streamline acoustic parameter selection methods.

Chapter 5 concludes this dissertation by presenting an informed discussion of the future of acoustic features in phonetics. I begin by conceptualizing the key properties of a good phonetic feature: performance, interpretability, and accessibility, and discuss operationalization of these properties for evaluating features in the field. Finally, I use this conceptualization to identify one candidate for phonetic feature representation, a neural network with potential to have high performance, interpretability, and accessibility, as an example of one potential avenue of implementing the findings described in the previous sections to tangibly improve acoustic representations in phonetics.

In summary, this dissertation presents an argument for continued development of acoustic measurement in the field and provides a broad, concrete framework to assist in future development, testing, and selection of acoustic representations in phonetics.

For Tori

Table of Contents

Acknowledgements

First and foremost, I'd like to thank my committee, Keith, Hannah, Gasper, and Gopala, for their valuable insight and input into this dissertation, especially in the early formulation and scoping of this work. I am especially grateful to Keith for advising me throughout my career at Berkeley, from applying to fellowships in my first year to undertaking my final dissertation project. Several pivotal moments of my graduate career were only possible with his continued support and belief in my ability to tackle new challenges.

My journey has also been deeply influenced by my broader community at Berkeley, in particular the D-Lab, UCMAP Yongmudo Club, and GradPro check-in groups. Special thanks to Arjun, Sophie, Sierra, and Nirupika for creating a wonderful and supportive community. Thank you also to Cara and Pelagie for being my Zoom writing buddies, dissertation support, and so much more- we made it!

I am also grateful to friends who have joined me through the journey of graduate school. My cohort mates Maddy, Emily, and Wesley were there from the very beginning and inspired me every step of the way, Zach kept me grounded through the tumult of the pandemic, and Anna has been my fellow phonetician and co-advisee. Thank you also to Jennifer, for being there since the first day of college and for being a copyeditor, soundboard, occasional co-author, and all-around great friend. Last but not least, thank you to Simon, for obtaining canned coffee, post-writing gaming sessions, and reminding me to drink water.

Finally, I am grateful to my family for their staunch support throughout graduate school. While it was a long journey, you were there with me every step of the way, and I am grateful for your insight, advice, and listening ear.

Chapter 1: Methods of feature extraction in acoustic phonetics

1.1 Introduction: transforming speech into numbers

Phonetics is the study of speech, but in analysis it is more common to engage with a transformation of the original signal. These provide a visual, interpretable representation that aids in description and analysis. While common, acoustic representations are mathematically complex operations that carry their own assumptions, benefits, and limitations. For example, if a researcher records a sound and wishes to examine it in a general way, there are three main types of basic transformations that may be used:

The visual representation of recorded speech is the **waveform** (Figure 1.1, top) which shows the variation in sound pressure levels over time. This visualization is used to anchor where speech occurs in a signal, and for auditory coding. However, in order to identify subtle differences in properties of the sound, the waveform must be further transformed into the spectrogram and spectrum.

The **spectrogram** is a two-dimensional spectrotemporal series based on a transformation of the waveform. The x-axis refers to time, the y-axis to frequency, and the value in each cell corresponds to energy at that particular point. The spectrogram visually summarizes acoustic patterns in both the time and frequency domains, and is commonly used by analysts in exploratory and descriptive research.

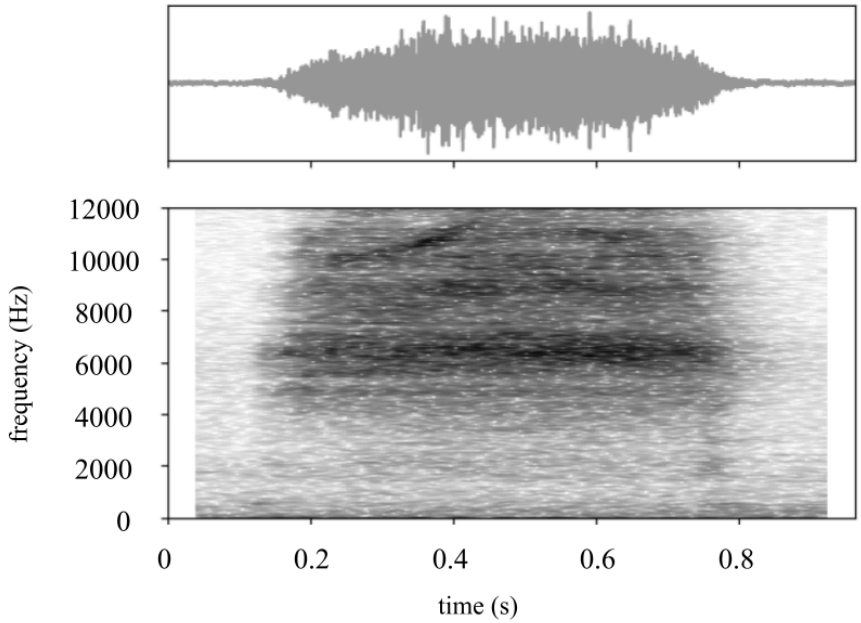

Figure 1.1 Example waveform (top) and spectrogram (bottom) for /s/

Finally, the **spectrum** shows energy across frequencies for a slice of time- e.g. a slice of a spectrogram. The width of the slice of time can be manipulated to generate representations that focus on different parts of the acoustic signal. Figure 2 gives an example of two spectra taken from the same spectrum, where the top uses a narrower window, and the bottom uses a wider window. Most acoustic correlates are calculated based on some form of spectrum transformation, since it provides a snapshot of the distribution of energy across the frequency domain.

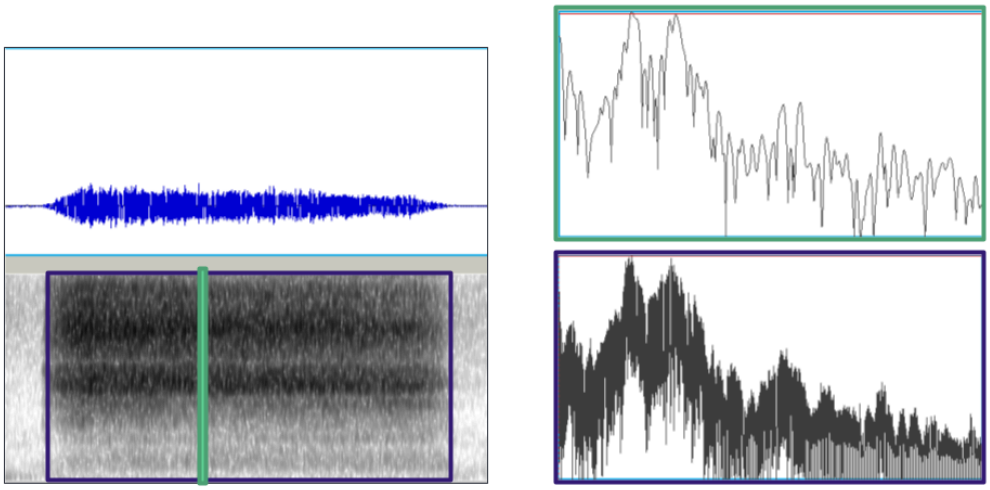

Figure 1.2 Two example spectra with different window lengths

These representations involve time series data, which poses a unique set of challenges for data analysis: they are high dimensional and each point is dependent on those around them. These properties pose an issue for many statistical methods and often include a level of detail that obscures the patterns of interest. For these reasons, a popular approach is to derive a small number of features that represent the specific acoustic properties of interest in the signal. These values are typically derived from the spectrum in some way at a single or small number of time points. There are four major categories of these derived representations:

1. **Qualitative analysis**: Description of patterns in the data based on visual or auditory impressions (e.g. describing the spectrogram, coding categories)
2. **Frequency measures**: Measurement of the frequency of key features in the spectrum (e.g. formant)
3. **Amplitude measures:** Measures based on the amplitude of certain parts of the spectrum (e.g. spectral noise measurements)
4. **Shape measures:** Numerical description of the overall properties of the spectrum (e.g. spectral moments - center of gravity)

Figure 1.3 gives examples of how these measures relate to the spectrum. A qualitative analysis would describe the spectrum, often in comparison with those of other sounds of interest. Amplitude measures might report the amplitude (y-axis value) of a significant event, while a frequency measure might report the frequency (x-axis value) of a similar event. Shape measures will calculate values comprehensive to the signal, such as the slope of the frequency. Typically, a

given analysis will usually report multiple acoustic parameters that belong to one or more of these analysis categories[1].

The remainder of this chapter will outline the foundation of acoustic features in phonetics and explore how these approaches set the stage for modern phonetic analysis. There are several

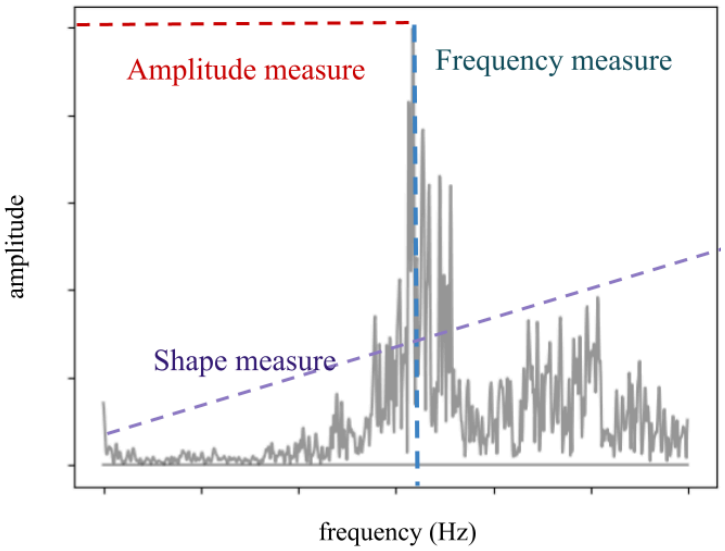

Figure 1.3 Three ways of parameterizing the spectrum

existing reviews that discuss measures in each major category of acoustic contrast (e.g. Harrington 2010; Johnson 2011). Rather than attempting to be an exhaustive review or tutorial, this chapter highlights the common usage, methodology, and validity of major numerical approaches to phonetic features.

1.2 Qualitative analysis

Qualitative analysis is an approach in phonetics where an analyst examines aspects of the visual representation, often using transformed speech to emphasize patterns in the data. While the focus is largely on quantitative measurement in the field, qualitative analysis still plays a, perhaps surprisingly, large role in acoustic analysis. Two primary strategies for qualitative analysis are visual interpretation of the data and auditory coding.

[1] So far, this work has abstracted away from the role of dynamic temporal changes in acoustic representations. Speech is known to be a dynamic process, and so representations that capture temporal variation might be essential to capturing phonetic contrasts. I will explore current practices in timing acoustic measurements in Chapter 2, and whether contextual information is necessary in Chapter 3.

1.2.1 Visual interpretation

Before computerized analysis, the spectrograph (Koenig et al. 1946) was used to generate physical spectrograms and the resulting image studied in order to better understand differences between different types of speech (such as in Fischer-Jorgensen 1954). Interpretation of a spectrogram includes inspecting the data and recognizing patterns such as the relative height of formants, the presence of frication, and other impressions of the acoustic content of the sound. These interpretations sometimes give rise to quantitative features derived from the spectrogram (as in Stevens 2000).

Perceptual and articulatory studies have found support for the impression that spectrograms are a useful representation of speech. It is possible for human raters to classify sounds with some accuracy from the spectrogram alone (Zue 1982). Articulatory and spectrographic data were combined to associate changes in articulation to changes in the spectrogram (Delattre and Freeman 1968).

However, interpretation of the spectrogram is not a replacement for speech perception. Reading spectrograms requires training and experience, and even then some phonetic detail is lost, more so for some contrasts than others. For example, it is easier to discriminate within the categories of stops and affricates than between the two categories (Lamel 1988). In order to have sufficient levels of phone recognition, it is also necessary to have context around the spectrogram of the target phone (Ingemann and Mermelstein 1975). Subsequently, the field trended towards statistical methods and quantification.

One early quantitative approach to spectrogram data worked under the assumption that there was an abstract spectral template for each speech sound category- an idea that was present in speech recognition (Cole et al. 1990; Goldberg and Reddy 1977), perception (Matsumoto and Wakita 1996; Hillenbrand and Houde 2003), and in other fields such as music (Yoshii et al. 2005). A similar line of research uses Match Filters in order to identify spectral features relevant to vowels and consonants (Kashani et al. 2017). These models used machine learning to learn the patterns of differences between phonetic categories. However, it has been found that these do not work as well for more acoustically similar categories (Itakura 1975), and the field moved to focus on highly targeted acoustic features rather than general spectral templates.

1.2.2 Qualitative coding

Another type of qualitative analysis that is pervasive in the field involves classifying data based on the perception of one or more analysts. This is often based on some combination of the waveform, spectrogram, and auditory perception. Once the target sections are coded they can be quantified either by reporting the relative rate of each category in the data (e.g. Leemann et al. 2018), in the case of categorical properties, or by reporting the duration, such as duration of creaky voice in Davidson (2021).

However, perception of non-native contrasts is more difficult and, perhaps more concerning for phonetic research, non-native categories can be assimilated to those in the

analyst's native language (e.g. Best and Strange 1992). One approach to mitigate bias in the coding process is to identify criteria for coding based on acoustic landmarks in the spectrogram such as onset/offset of voicing, frication, or formants (e.g. Morris and Hejná 2020). This uses a combination of structured qualitative and quantitative criteria in order to mitigate interrater variation in qualitative coding.

An alternative to reporting criteria for qualitative coding is to use machine learning to reduce human influence. Forced aligners such as the Montreal Forced Aligner (McAuliffe 2017) can supplement qualitative coding in certain categorical cases, for example in automatically determining fricative identity (Rao and Shaw 2021), and targeted models have been developed in order to support the study of sociophonetic variation, for example automatically coding r-vocalization (Villarreal et al. 2020).

Qualitative analysis still has a large influence in acoustic description, and is often used in particular in cases where quantitative methods have not yet been adequately developed to describe a contrast. In a recent example (Davidson 2021), the author identifies 7 different realizations of glottalization in the target context, and interprets a spectrogram of each one, where key differences are described. These are accompanied by sub-categorical labels that describe acoustic features of the category. In addition, the relative duration of each of these segments is reported. This results in robust qualitative understanding of the meaningful variations in a case where the quantitative measures associated with these properties are not well-understood. However, the next step past this level of description is not clear in terms of operationalizing description into experimental or statistical analysis, which poses issues for extending these lines of research in subsequent work. A more general concern, as noted in Thomas (2011) in sociophonetics, is that auditory coding is frequently used in assessing consonantal variation, while vowels are usually measured quantitatively. This is an undesirable state, since it may bias research towards more easily measured phenomena and encourage underrepresentation of other categories. General trends highlight a need for more generally robust acoustic measures, so that quantitative approaches can be applied across different phonetic sounds.

1.3 Frequency measures

One of the most common types of acoustic parameters involves identifying key features of the spectral representation, often related to the frequency location of areas of high energy in the spectrum. The primary methods in this category are formants and fundamental frequency. The acoustic theory of speech production argues that speech is made up of two main components, the source and the filter (e.g. Fant 1970). Thus, one approach to acoustic analysis is to identify frequencies in the spectrum that relate to properties of either component, and long before modern methods of acoustic measurement, the frequencies associated with vowel formants and vocal fold tone were a topic of much research (reviewed in Metfessel 1929). This means that frequency

measures have the most extensive historical foundations of any of the different types of acoustic parameters.

1.3.1 Formants

Formants are resonant frequencies that are characteristic of a specific vocal tract configuration, and were a major topic of discussion in models of speech (as in Dunn 1950). After the introduction of the spectrograph, the resonant frequencies of formants were some of the most salient features in the spectrogram, and could be measured on printed spectrograms using lines associated with reference frequencies (e.g. Peterson and Barney 1952).

These days, formants can be extracted algorithmically, the most common being Linear Predictive Coding (LPC) implemented in Praat (Boersma 2011). This method uses a transform that estimates the filter function of the local tract, the coefficients of which can be interpreted as frequencies. The LPC algorithm is by far the most commonly used one for formant extraction, but there are documented issues with accuracy and improving formant measurement remains a major area of research (e.g. Harrison 2013). Alternative algorithms for formant extraction have also been proposed including Fastrack (Barreda 2021) and wavelet based analysis (Orellana and Ugarte 2021).

In addition, significant research has focused on understanding the relationship between these frequencies and the original speech. The perceptual relevance of formants was tested by synthesizing vowels from formant (and sometimes other) parameters, and testing how well the resulting signal was perceived. One study of particular interest compared formant-only vs whole-spectrum information vowels and found that formants-only gave almost equally good results as the whole spectrum (Hillenbrand et al. 2006). The same study found that the approach was less effective for consonants, citing that this is likely due to the more time-sensitive nature of consonants and therefore a less stable spectral indicator in frequency peaks.

Articulatory work has largely focused on investigating a small subset of relationships between the acoustic correlates and articulation. Formants have been related to vowel articulation via electromyography (Maeda and Honda 2004), ultrasound (Lee et al. 2015), and electromagnetic articulation (Blackwood Ximenes 2017), which indicate a relationship between formants and tongue position.

However, formants alone have not been sufficient for distinguishing between all vowel contrasts- for example formants alone are not sufficient for discriminating between +/- ATR vowels (Fulop et al. 1998). In addition, formants are known to vary across speakers, so significant work has focused on strategies for speaker normalization (reviewed in Johnson and Sjerps 2021).

While the first and second formants are by far most commonly used in vowel quality description, there are also some derived measures that give a more overall understanding of the spectrum shape. This can consist of a simple difference such as F2-F1 (e.g. Sharifzadeh et al. 2012), or a more complex calculation such as the formant centralization ratio (Vizza et al. 2017),

which is particularly common in voice research. Primarily, formant values for vowels are measured in terms of static properties, which are known to have some amount of variation (Kent and Vorperian 2018).

While formants are primarily associated with vowel analysis, they have also been measured in other types of sounds, particularly sonorants such as glides (Maddieson 2008; Ainswore and Paliwal 1984), approximants (Best and Strange 1992), liquids (Narayanan et al. 1999; Scobbie et al. 2013; Kochetov et al. 2018; Crosby and Dalola 2021) and nasals (Nakata 1959; Recasens 1983). In nasals, there are additional resonances that interact with the formants (Fujimura 1962) and further variation in the acoustic cues (Stevens 2000) which make both the measurement and interpretation of formants less straightforward for these sounds.

The idea of formant frequencies have also been briefly explored for fricatives (Jassem 1965, Soli 1981). However, transitions between consonant and vowel are more commonly used for consonant sounds to identify place of articulation (e.g. Malecot 1956; Halle et al. 1957; Cassidy and Harrington 1995; Tabain et al. 2020). This speaks to the relatively robust and nuanced understanding that exists of the relationship between formants and articulation in that they can be used to determine the place of articulation of their neighbors, and positions formants as a way to measure sounds that lack robust quantitative measures of their own.

1.3.2 Spectral peak

Formants are based on the idea of resonant frequencies in the vocal tract. However sounds like fricatives are made with turbulent airflow, meaning that resonances do not hold the same phonetic meaning as they do in vowels and sonorants. In place of formants, in early phonetic literature, another measurement, the frequency of spectral peak, was described as a 'formant-like' analogue for fricative consonants (Strevens 1960, Gordon et al. 2002). This value can be straightforwardly determined from the spectrum by identifying the frequency at which there is peak energy in the spectrum, but can be influenced by many non-target factors, resulting in a noisy signal. Jongman et al. (2000), in a study of fricatives, find that as the place of articulation goes further back in the oral cavity, the location of the peak frequency also decreases. Similar measurements in this same vein have also been used, for example spectral roll-off, or the frequency value of the spectrum under which some percent of the total energy is located (used in Frid and Lavener 2010), or a modified spectral peak measurement (presented in Shadle et al. 2023). Many of these derived methods are determined based on first a qualitative understanding of the spectral differences between the segments of interest, then operationalizing these differences through quantitative measurement, another example of the influence of qualitative analysis on the conceptualization of quantitative features in phonetics.

1.3.3 Fundamental frequency

Fundamental frequency (f0) is an acoustic correlate of pitch of a sound. F0 is typically calculated using autocorrelation (Boersma 1993) to detect periodicity in the source. Analogously in f0 measurement, alternatives to the common autocorrelation method of f0 analysis have been proposed (Chu and Alwan 2012; Hirst 2007). In general, both f0 and formants can be quickly and automatically measured with a reasonable degree of accuracy using any one of a number of tracking algorithms, which makes it possible to make measurements much more easily for a larger amount of data than previous decades.

F0 is most widely used for suprasegmental contrasts, particularly tone and stress. However, f0 has also been used as a secondary correlate of other contrasts where pitch may be expected to interact with other contrasts such as voice quality (Gerfen and Baker 2005) and voicing (Maddieson 1984). In voicing, the f0 of the neighboring vowel is used to predict the voicing of the consonant. The use of both formants and f0 to measure contrasts in neighboring segments is also an indication of the relative confidence and ease of which it is possible to measure these values when compared to consonant-internal measures.

In contrast to the previous measures, which are related to properties of the filter, fundamental frequency (f0) is related to the rate of vibration of the source (i.e. vocal cords). Pitch has been related to vibration of the vocal folds through electroglottography, a technique that measures the movement of electric current across the tissues of the vocal cords (Bough et al. 1996). This measure is fairly transparent, well-understood, and easy to measure. However, it is limited to a very narrow part of the total acoustic information, and like formants, its utility as a direct measure is limited to a small set of phonetic sounds.

1.4 Amplitude measures

The previous section focused on the frequency values of significant landmarks in the spectrum. The other major approach to acoustic features is to describe the amplitude of the spectrum, which can correspond to landmarks, regions, or the spectrum in its entirety.

1.4.1 Harmonic-based methods

Harmonic measures first detect the location of harmonics in the signal, then calculate the amplitude differences between certain harmonics. The most well-known of these measures is H1-H2, the difference in amplitude between the first and second harmonics. H1-H2 is part of a larger class of harmonic measures, which measures the difference in amplitude between any two harmonics in the spectrum, for example harmonics near formants and nasal formants. The development of these measures was largely driven by questions around the acoustic correlates of phonation. With non-modal phonation such as creaky or breathy voice, there is often a clear perceptual difference in the quality of the sound independent of vowel quality, and there have

been several attempts to derive features that relate to this qualitative difference. An early study found that the amplitude of the first harmonic was identified as a potential correlate of breathiness in Gujarati (Fischer-Jorgensen 1967). A related class of measures quantifies the relative amplitude of all harmonic components vs. other components. The most common of these is the Harmonic-to-Noise ratio, which is the ratio between amplitudes of harmonics and noise in the signal.

Harmonic measures can be manually measured from a spectrum by identifying the location of the two points of interest and measuring the difference in amplitude between them. However, with the development of VoiceSauce (Shue et al. 2011), a software to measure these metrics, automation is now possible. VoiceSauce does this by finding the fundamental frequency, estimating the location of the harmonics based on that value, and then doing a local search of the spectrum around those points in order to find the precise locations of peaks. A key contribution of VoiceSauce is correction for formant values which previously would have a strong effect on the measurements and made it labor-intensive to compare across vowel qualities. Noise measures are calculated by comparing the amplitude of the harmonics to the noise floor in the signal within a specific frequency band (Boersma 1993).

Given the number of harmonics in a spectrum, there are many different combinations and permutations of these two general parameters. This has led to significant increase in the number of metrics that are reported for a specific contrast, and the challenge is to subset the vast number of parameters available down to the most meaningful. A study on vowel nasality (Styler 2017) investigates 22 different methods, and a meta-analysis on phonation types identifies over eighty (Latoszek et al. 2018). For that reason, subsequent work has focused on narrowing down the acoustic correlates using perceptual and articulatory evidence.

Generally, harmonic and noise measures are related to properties of the source signal. Perceptual evidence has found that, cross-linguistically, H1-H2 is the most robust correlate of phonation type (Esposito 2010; Kreiman et al. 2007 ;Klatt and Klatt 1990; Garelleck 2021) and is linked to the articulatory Closed Quotient in electroglottographic measurements (Kreiman et al. 2012). There is additional evidence for the anatomical link of measures like H2-H4 (Khan 2012; Garelleck et al. 2013), as well as the measures harmonic-to-noise and subharmonic-to-harmonic ratio (These and further measures are reviewed in Garellek 2022).

HNR is a measure of noise in the signal and is typically related to breathiness or other frication in the sound (Brotherton and Block 2020). Harmonic and noise measures have been primarily developed with regard to voice quality in vowels, but have also been applied to voice quality in other speech sounds, such as sonorant breathiness (Berkson 2022). They have also been used to study vowel nasalization (Styler 2017). In obstruents, which typically lack harmonic structure, these measures may be made on the neighboring vowels, which is common in research in fortis/lenis or tense/lax consonant contrasts (e.g. Cho et al. 2002). HNR has been used in comparing fricatives and approximants, since fricatives can be expected to have a higher noise component (e.g. Brotherton and Block 2020)

1.4.2 General amplitude measures

The measures described in the previous section rely on the presence of harmonics in the speech signal, which restricts them to being used for sonorant sounds that are expected to have harmonics. Another group of amplitude measures can be applied to any spectrum, rather than relying on the detection of harmonics.

The global metric of spectrum amplitude is intensity. This can be measured directly on the segment of interest, but is more often measured in terms of the relative amplitude - the difference between, typically, the target segment and the neighboring vowel (e.g. Martínez-Celdrán and Regueira 2008). Intensity has been found to be useful in determining the place of articulation of stops (Ohde and Stevens 1983) and fricatives (Jongman et al. 2000; Shadle and Mair 1996; Ho 2021, Nartey 1982). It can also be used as a manner of articulation measure- for example in the difference between liquids and nasals (Cheng and Jongman 2019) and for stop lenition (Hualde et al. 2019).

Intensity can be related to some extent to airflow, for example intensity associated with fricative place of articulation can be placed on an intensity scale related to the rate of airflow varying between different places of constriction (Strevens 1960). It is useful because it is not limited to a specific phonetic category and can be used to compare across manners of articulation, but it is not a very precise measure, which introduces some limitations.

A more local approach is frequency band analysis, which compares the amplitude of different frequency regions of the sound. This approach is less common in current practice, but has been applied to a variety of speech sounds including fricatives (Hughes and Halle 1956; Nartey 1982; Koenig et al. 2013), nasals (Takeuchi et al. 1975; Kurowski and Blumstein 1987) and vowels (Tahiry et al. 2016), where, rather than using peaks as landmarks, the amplitude across bands of the frequency region are used. Other amplitude approaches use the ratio between amplitudes of components of the sounds, analogously to harmonic measures: for example the ratio of low frequency energy to total energy (Gradoville 2011), or the relative amplitude of two spectral peaks (Stevens 1985). These approaches are able to capture more detailed information about characteristic patterns in the spectrum.

1.5 Shape measures

Another way of approaching acoustic measures is to directly describe the shape of the spectrum rather than measuring properties of significant landmarks. These types of measurements tend to focus on describing global properties of the spectrum rather than local features.

Although not in widespread use today, one of the earliest global measures was spectral slope, where researchers would fit a regression line to the spectrum in order to identify overall shape (Maniwa et al. 2009; Shadle and Mair 1996; Ho 2021). Spectral tilt has also been shown to be predictive of perceptual differences (Katseff 2008). One of the difficulties in using spectral slope is that it quickly became clear that a single slope measurement is not sufficient for

describing a spectrum, which typically has one or two important peaks. This can be modified by looking at piecewise measures of slope within specific frequency ranges (as in Klatt 1982), which have also been reported as useful in differentiating between fricative places of articulation (Shadle and Mair 1996). However the piecewise approach results in metrics that are tailored to a specific scenario and are therefore less generalizable. While local measures can capture nuanced patterns more effectively, global measures that are more easily applied across phenomena seem to be more commonly used in a range of phenomena.

1.5.1 Spectral moments

Currently, the most common family of global shape measures is spectral moments, which was popularized for the use in obstruent place identification (Forrest et al. 1988). This approach treats the average spectrum of a segment as a probability distribution and calculates average statistics under those assumptions. The first four moments and their typical interpretation are:

1. Center of Gravity (CoG): Weighted mean
2. Standard Deviation (SD): Overall spread
3. Skewness (Skew): positive or negative skew
4. Kurtosis (Kurt): flatness/peakedness

While the original use of spectral moments was exploratory, there has been significant work after the fact investigating how this method interacts with acoustic and articulatory properties of the data. Much of this work has highlighted conditions under which spectral measurements have not been sufficient in discriminating between relevant phonetic categories in the data (e.g. Jassem 1995). In the original paper applying spectral moments to fricatives, while discrimination was high for stops and sibilants, the authors recognized that the spectral moments analysis was not successful in discriminating between the non-coronal fricatives (Forrest et al. 1988). Similarly, in Swedish moments did not differentiating between non-coronal fricatives (Wikse Barrow et al. 2022), an overlap in moments for fricatives was reported in Polish (Zygis and Hamann 2003), and between dental and labiodental obstruents in Hong Kong English (Ho 2021). Several perception studies have also reported results where perceptible differences aren't adequately reflected in spectral moment measurements (Schadle and Mair 1996; Jannedy and Weirich 2017; Zygis and Hamman 2003). Finally, several studies have critiqued the reliability of interpretation of the spectral moments at all, especially of kurtosis and skewness (Koenig et al. 2013; Fulop et al. 2004). More abstract methodological research has investigated the derivation and interpretation of spectral moments as a whole. Both skewness and kurtosis can correlate with CoG, which calls into question how much unique information each spectral moment contributes (Hargus et al. 2021). In addition, spectral moment measurements are quite sensitive to the frequency range of the spectrum (Schadle and Mair 1996).

Despite this active area of research, spectral moments, particularly CoG, are the most commonly used shape measurements and appear to dominate modern acoustic analysis of fricatives. These measures can be made on any spectrum, since they don't rely on any particular

acoustic presence such as harmonics or voicing in order to be measured. In addition to fricative noise, moments have been used in any case of obstruents, notably, in the burst release of (ejective) stops (Jaworski and Baran 2021) and affricates (Li and Gu 2015), as well as clicks (Fulop et al. 2003). CoG and Standard Deviation have been used for nasals (Tabain et al. 2016) in order to identify place of articulation to partial success, in particular in differentiating the labial /m/ from other nasals. Spectral moments have also been used in the acoustic analysis of vowels (Tahiry et al 2016) and liquids (Carter 2003; Gobl and Ni Chasaide 1999). A slight variation on spectral moments reports centroid frequency within a certain frequency range, which presents an intermediate representation of the spectrum (Li et al. 2007). However, moments are used primarily in a descriptive sense, and there is little to establish a link between these methods and the relevant articulation or perception of the sound.

1.5.2 Discrete Cosine Transform

The other major approach to quantify spectral shape is to numerically decompose the spectrum into components, most commonly using the Discrete Cosine Transform (DCT). Analogously to how a Fourier transform is used to decompose the waveform into a spectrum, the DCT decomposes the spectrum into coefficients applied to different cosine functions. This has been used in phonetics for parameterizing time-series of formants (Watson and Harrington 1999) and pitch trajectories (Wu et al. 2008; Yu et al. 2022). It is also possible, although less common, to apply this approach to frequency-series data, for example in fricatives (Jannedy and Weirich 2017), and in stops (McCarthy 2019).

DCT coefficients are in some cases better at discriminating between fricative series, and perform equally well as spectral moments otherwise (Giles 2019). While the reconstructed signal from the DCT can be treated as a smoothed spectrum, the individual coefficients are difficult to interpret. The first few components, which are the ones most frequently reported, correspond to low frequency components of the transform. However, these first few components may not adequately capture more local features of the spectrum that distinguish between more similar types of spectral.

Shape measurements are attractive because they do not rely on particular acoustic properties, and are often derived fairly transparently from the spectrum. However, with this broad usage, they lose the interpretive power associated with the other measures described here, and are largely limited to descriptive and exploratory analysis. In addition, while these measurements work for acoustically distinct sounds, they appear to be less effective in distinguishing between similar sounds. Variations on the measures that look at properties like centroid frequency or spectral slope within a specific frequency range suggest that the desired level of detail within the spectrum is greater than that offered by purely global shape measures.

1.6 Choosing a phonetic measure

This review has highlighted the current state of acoustic measurement in phonetics: there are a massive number of possible acoustic measurements, but a small set of measures are common practice.

For some contrasts, there are robust acoustic cues that have become the common practice in the field (e.g. formants and vowel quality). These measures have significant meta-methodological research invested into them and a clear connection to the articulatory and perceptual evidence for the contrast. For many contrasts, however, the appropriate measurement is not as well established. In these cases a variety of strategies can be used in order to determine what features might be appropriate. This section illustrates an example of approaching the acoustic analysis of a phenomenon that does not have a well established set of acoustic correlates.

The example is a study that is an investigation of nasal place of articulation contrasts in three languages of Australia (Tabain et al. 2016). Nasal contrasts do not have a generally accepted parameterization, and so the study used the following strategies to determine what features to use:

1. **Existing measures:** First, the study takes stock of the existing literature on nasal acoustics and finds that nasal formants and bandwidths will likely be useful in determining place of articulation differences. In addition, they note points of variation in the existing literature as to the specific predictions regarding these values.
2. **Extension of measures:** Given that the existing methods have been so far determined to be insufficient for distinguishing between the contrast of interest, it is also possible to repurpose existing measures to apply them to the phenomenon in question. To this end, the authors include two spectral moments, CoG and SD, as two measures to characterize the general shape of the nasal sounds, extrapolating from their utility in other consonant contrasts, specifically stops and laterals.
3. **Novel measures:** From preliminary analysis, the authors elect to add another measure: average bandwidth of the nasal formants.
4. **Qualitative analysis:** Finally, the authors present the average spectrum of each of the nasals in the set and qualitatively describe them including the locations of features such as antiresonances and spectral minima.

This study included a relatively large number of acoustic features, including adapting existing measures applied in other contexts in an attempt to describe the sound. Despite this extensive work, qualitative and impressionistic description of the different sounds was still required to reach a satisfactory analysis. This illustrates both the need for a more comprehensive understanding of phonetic features, especially in these types of cases, as well as the tendency for

qualitative analysis to fill in in cases where acoustic analysis is not yet robust enough for this type of analysis.

This study used four different strategies for describing the nasal place of articulation. A fifth strategy, which wasn't used in the paper discussed above, is to use feature selection. Feature selection entails choosing a large set of phonetic features, then using a systematic approach (often based in machine learning) to select a smaller subset of features from this one. Feature selection has been used in fricative analysis (Jongman et al. 2000), rhotic variation (Villareal et al. 2020), and plosives (Li et al. 2012) Another approach to feature selection considers the contribution of each feature to the variance in the signal. Commonly, this is done by using Principal Components Analysis or PCA (as in Abdi and Williams 2010) to identify the primary axes of variation in the data and then calculating the relative contribution of each feature to it, which has been applied to fricatives (Ulrich et al. 2021) and sociolinguistic sub-categorical variation (Adank et al. 2004).

1.7 Conclusion

The current landscape of acoustic parameters in phonetics consists of a patchwork of acoustic measures that have developed organically over several decades of research. Crucially, there appears to be persistent limitations on the kinds of contrasts that can be addressed acoustically in linguistics based on the current best practices in feature selection. In addition, the use of common approaches is not directly related to the robustness of the methodology. For example, spectral moments is a measure of spectral shape that is in wide use today, despite significant criticisms of both the theoretical and empirical effectiveness of the method. Part of this is the lack of suitable alternatives with traction and accessibility, part of this is inertia in the field - since following the same approaches as previously successful studies is a major motivation for using a specific methodology, the use of a particular method becomes a self-reinforcing process.

However, the current state of phonetic parameters appears to constrain possible avenues of quantitative research. Although the measurement of an acoustic parameter is not strictly reliant on a given phonetic contrasts, acoustic measures are typically developed in association with a specific phonetic contrast. It follows that the more common acoustic contrasts are also associated with the most robust measurement methodology, which is seen in the number of validation and methodological studies associated with measures like formants and spectral moments.

In some cases, indirect measurements using these established methods are preferred over direct measurements of the segments of interest for a given study. For example, one study that was concerned with the dental/retroflex contrast across stops, nasals and liquids noted that place of articulation is manner-specific, and constrained their study to using formant transitions as the shared correlate to place of articulation in the absence of the availability of direct measurements (Kochetov et al. 2018). However, in some cases it is not possible to even indirectly leverage these more common contrasts and acoustic parameters. In this case, the result is typically to use

qualitative strategies to describe the contrasts. For example, one study of the Polish rhotic used spectrograms and coding to describe variation in the phonetic realizations of the phoneme (Jaworski and Gillian 2011). These facts motivate the need for further development of acoustic features in order to be able to quantify all key contrasts of interest to phonetics. In particular, given the asymmetry in how established best practices are across phonetic categories, it seems advantageous to take a broad view in this process to account for and ameliorate this imbalance going forward.

Chapter 2: Current practices in acoustic measurement in phonetics

2.1 Introduction

Chapter 1 highlighted that the range of acoustic parameters used in practice are much narrower than the total possible variation. However, in addition to the qualitative patterns established there, it is also useful to have numerical data to confirm how these patterns translate into practical usage.

Quantitative methodological surveys can address this gap, and are common practice in adjacent fields, including articulation (Kochetov 2020; Rebernik et al. 2021), speech (Bhatt et al 2020), prosody (Yi 2011) and forensic phonetics (Gold and French 2011). A structured investigation of the range of variation in the use of acoustic measurements in phonetics can be expected to highlight areas where the field would most benefit from further methodological investigation, and as a point of reference when determining best practices in the field.

The only example of a methodological survey in phonetic measurement consisted of three collections of phonetic descriptions (Journal of Phonetics, Journal of the International Phonetic Association, and Sounds of the Worlds Languages) to identify how frequently certain phonetic structures were reported (Whalen et al. 2022), including some acoustic measures. Since the goal of including quantitative measures in description is in part to facilitate cross-linguistic and typological comparison, this study provides some indication of attitudes towards robust acoustic measures. The results from the study are replicated below in Table 2.1.

Consonant	%	Vowel	%	Suprasegmental	%
voice onset time	25	formants	35	stress	20
closure duration	16	dispersion	24	length	9
voicing	16	additional features	24	tone/pitch accent	23
formant transitions	6	duration	25	intonation	13
fricative spectra	5	intensity	3	interactions	18
fricative duration	5	interactions	19	other	4
burst	8	other	15		
pre-aspiration	3				
sonorants	16				
other	13				

Table 2.1: Aggregated coverage of certain categories in phonetic descriptions (from Whalen et al 2022)

For vowels, formants are by far the most commonly reported, followed by dispersion measures which are also derived from formants. For consonants, the primary methods are related to duration: voice onset time, closure duration, and voicing. In terms of spectral correlates, measuring transitions into adjacent vowels (6%) is more frequent than reporting measures related to the fricative spectra (5%). Particular acoustic measures relating to sonorants and suprasegmentals have not been measured in this survey, which implies that these do not have well established acoustic measures suitable for typological comparison.

This chapter gives quantitative insight into common practice in acoustic measurement via a methodological survey of phonetic studies to further investigate the patterns identified in Chapter 1. The results of this survey can be expected to be useful in identifying areas of priority in methodological research in the field of acoustic phonetics.

2.2 Methodology

2.2.1 Data collection

There is a large range of contrasts and methods within acoustic phonetics, which requires a larger sample size. In the survey described above (Whalen et al. 2022) of 1536 JPhon papers published over 30 years, 110 fit the criteria to be included in the study. Thus, finding a large enough sample of papers to provide insight into best practices was an important consideration for this survey.

The International Congress of Phonetic Sciences is a major conference on phonetics that occurs every four years. While it is primarily associated with linguistic phonetics, it also has contributions from overlapping fields such as speech science, natural language processing, and forensic phonetics. It also has a relatively large number of studies within the collection (600-800 each session of the conference).

The sample for this particular study contains the proceedings for the years 1999 and 2019. The most recent proceedings available at the time of data collection was 2019, and 1999 contains papers published before the introduction of the most recent set of phonetic software tools. There were 640 papers in 1999 and 792 in 2019, for a total of 1432 papers. Of these, 395 (155 in 1999, 240 in 2019) were determined to be relevant based on the title and the abstract as pertaining to a consonantal, vocalic, or suprasegmental contrast or variation. Papers that pertained to a unit of study larger than a segment (e.g. sentence prosody) were excluded from the sample, since acoustic measurement works quite differently when dealing with multi-segmental units.

2.2.2 Data processing

The papers in the sample were collected using a Python script to semi-automate data collection and processing. A Python script aided in the paper coding process using a human-in-the-loop approach to prompt a set of survey fields and validate inputs. Each relevant paper was coded for contrast under study, measure used, and sampling strategy. After coding, these measures were organized into the following broad categories for contrast type and subtype, parameter type and subtype, and temporal measure.

Phonetic contrasts in the sample includes suprasegmentals, vowels and consonants. Consonants were divided into two broad categories: obstruents and sonorants. Sonorants tend to be more vowel-like in measurement strategies because they have similar properties such as voicing and harmonic frequencies. Obstruents such as fricatives and stops are produced differently, which necessitates different measurement strategies. These categories are further divided into subtypes based on what groups can be expected to share acoustic characteristics (Table 2.2).

In addition to contrast, the acoustic measures used in the sample were also grouped into common types (Table 2.3). This includes two frequency measure types: formants and pitch, one

Category	Subtype	Category	Subtype
Obstruent	manner	Suprasegmental	stress
	length		length
	voice quality		voice quality
	place		tone
	devoicing	Vowel	devoicing
	voicing		other
	insertion/deletion		vowel quality
Sonorant	manner		insertion/deletion
	other	Other consonant	manner
	place		other
	insertion/deletion		place
	vowel quality		
	length		
	voice quality		

Table 2.2: Contrast groups and subtypes in the survey

Measure	Description
formant	measures based on formants or combinations thereof
spectral moments	spectral center of gravity, standard deviation, kurtosis, skewness
pitch	measures of fundamental frequency
voice quality	H1-H2, HNR, other harmonic-based measures
other spectral	spectral peak, spectral tilt, and other spectral measures

Table 2.3: Categories of measurement in the survey

amplitude measure group: voice quality, and one shape measure: spectral moments. These are the four categories that appear to dominate the discourse around phonetic methodology and features.

The fifth category encapsulates alternative measures, and an increase in this category suggests increased innovation in terms of developing and applying new measures to the data.

Finally, the data were coded for where the measure was taken within the segment (Table 2.4). Change in how the measure is defined can be expected to indicate another layer of methodological variation. In addition, even in cases where the acoustic parameterization is well established, variation in these sampling strategies can be an indication of how well established the norms are in the field.

Measure	Description
point	measured at one or more points in the segment
target	measured at some articulatory target
window	average over some window of the target, often the average or the middle 50%
trajectory	reported as a time series
other	idiosyncratic criterion, such as taken at the maximum or minimum point of the target
not specified	authors do not record sampling strategy

Table 2.4: Measure groups and subgroups in the survey

2.3 Results and Discussion

The proportion of total conference papers in the sample that were included in the survey are reported in Table 2.5. In 1991, 22% of the total papers in the proceedings were relevant, while in 2019 the proportion increased to 32%. The increase in papers in the sample could be due to several different factors, including increased attention to questions that were conducive to quantitative analysis, along with a general increase in access to tools that made taking and reporting acoustic measurements easier. In particular, in the last 20 years, there has been an increase in tools for automated and semi automated extraction of phonetic parameters, which have made acoustic analysis more accessible to researchers. This increase also indicates that

quantitative measurements are valued approaches in the field, and highlights the importance of further development of the methods used in this field.

2.3.1 Trends in frequency of analysis by contrast

Although there is a general increase in the usage of acoustic methodology between the two time points, this may vary by the types of categories under study. Some areas of acoustic analysis have more robust measurement practices associated with them than others, and this variation may impact usage in public work. In this survey, the overall trend in the rate of papers covering each type of contrast (Table 2.6) is used to give an indication of general attitudes towards the usefulness of acoustic measures in those cases. Most subtypes see a slight decrease in the proportion of studies in the overall sample dedicated to them. The main exceptions to these are vowel and suprasegmental categories, including vowel quality, tone, and voice quality. These are also three categories that have seen significant theoretical discussion and methodological development.

Category	Subtype	1999	2019	Change
Obstruent	manner	10.3	6.7	-3.6
	length	1.9	0.8	-1.1
	voice quality	0.6	0	-0.6
	place	9.7	9.2	-0.5
	devoicing	1.9	1.7	-0.2
	voicing	14.8	14.6	-0.2
	insertion/deletion	0.6	0.8	0.2
Sonorant	manner	6.5	2.5	-4
	other	2.6	0.8	-1.8
	place	4.5	3.8	-0.7
	insertion/deletion	1.3	0.8	-0.5
	vowel quality	1.9	1.7	-0.2
	length	0.6	0.8	0.2
	voice quality	0	0.4	0.4
Suprasegmental	stress	4.5	2.5	-2
	length	7.1	6.2	-0.9
	voice quality	5.8	7.5	1.7
	tone	5.8	8.3	2.5
Vowel	devoicing	0.6	0.4	-0.2
	other	0.6	2.1	1.5
	vowel quality	22.6	32.5	9.9
Other consonant	insertion/deletion	1.3	0	-1.3
	moa	1.3	0.4	-0.9
	other	1.3	0.4	-0.9
	poa	1.3	0.4	-0.9

Table 2.6: Change in analysis of each type and subtype, normalized by total papers in the sample. Green indicates a relative increase in studies, red indicates a general decrease. Highlighted cells indicate categories that are investigated further in the results.

The categories that see the largest decrease between study years are manner of articulation studies are both consonant categories- sonorants and obstruents. There are several factors that may be driving this trend, such as shifts in experimental and phonetic questions and overall frequency of contrasts. However, the shift to fewer consonant analyses may also indicate a general lack of confidence in the acoustic features typically applied to some categories, in particular consonants. Consonants have higher acoustic variability, and even today there is still uncertainty associated with the best way to characterize the contrasts as was discussed in Chapter 1. In contrast, there are many resources available for automatic or near-automatic analysis of vowels, and the major feature associated with vowels, the formant, has been the subject of significant methodological discussion. This imbalance may be helping to encourage a trend towards more vowel analysis and less consonant analysis.

Table 2.7 gives the proportion of papers in each category that contains at least one acoustic measurement, using a subset of eight categories from Table 2.6. There is no change in the number of acoustic measurements for vowels, which also have a high rate of acoustic analysis in both years- this suggests that the methodology for vowels has remained stable over the years. The only category with an increasing rate of acoustic analysis is voice quality. This

Category	Subtype	1999	2019	Change
Obstruent	manner	56.2	31.2	-25
	place	66.7	63.6	-3.1
	voicing	87	71.4	-15.6
Sonorant	manner	40	33.3	-6.7
	place	57.1	55.6	-1.5
Suprasegmental	tone	100	85	-15
	voice quality	66.7	88.9	22.2
Vowel	vowel quality	88.6	88.5	-0.1

Table 2.7: Change in percent of analyses with at least one acoustic measurement in them for each category of interest

coincides with the establishment of amplitude measures, such as H1-H2, as useful for measurement of voice quality. All other categories have a decrease in rate of acoustic measurement, with the biggest decrease in manner of articulation analysis for obstruents. While the mentality of the field has shifted to be largely in favor of quantitative analysis, in practice there has been an overall decrease in the amount of acoustic analysis in these categories, suggesting that current common practices have been unsatisfactory in terms of capturing

variation in these cases. Instead, alternative methods of describing the acoustic categories may have been employed.

2.3.2 Trends in alternative measures

The prevailing trend in phonetics has been to measure differences by comparing acoustic features of the target segment. However, when acoustic features fail, there are several possible alternative strategies.

1. **Neighboring measures**: Acoustic measurement, typically of a neighboring vowel (for which acoustic features are quite well established).
2. **Qualitative measures:** A higher rate of qualitative measures can be taken as an indicator that acoustic measures are unsatisfactory for describing the contrast under study.
3. **Articulatory measures:** While there are many reasons why this may be an independent goal, articulatory methods also require more resources in order to produce a study, so typically they are used more in cases where the current acoustic methodology is somehow insufficient. Therefore, in this case, in combination with the other alternatives above, a higher rate of articulatory measures is expected to indicate less confidence in our overall understanding of the sound.

Category	Subtype	1999	2019	Change
Obstruent	manner	6.2	18.8	12.6
	place	13.3	13.6	0.3
	voicing	26.1	37.1	11
Sonorant	manner	10	16.7	6.7
	place	14.3	33.3	19
Suprasegmental	tone	0	10	10
	voice quality	-	-	-
Vowel	vowel quality	-	-	-

Table 2.8: Change in percent of papers in each subtype that use neighboring segments for analysis

Table 2.8 shows the rate of analyses using acoustic measures of neighboring segments in order to investigate a phonetic phenomenon. Among consonant categories, there is an increase in measurements of neighboring segments across the manner of articulation. In addition, there is an increase in the use of neighboring segments in analyzing sonorants both for place and manner of

articulation. This suggests that the use of the neighboring vowel for analyzing consonants is still a popular approach. The trend upward over time suggests that the increased availability of formant extraction techniques may have affected not only phonetic analysis of vowels and their popularity, but also influenced that of consonants. It also indicates that any shift in acoustic analysis of consonants has not replaced the use of adjacent formant analysis.

Category	Subtype	1999	2019	Change
Obstruent	manner	43.8	31.2	-12.6
	place	20	27.3	7.3
	voicing	13	22.9	9.9
Sonorant	manner	50	66.7	16.7
	place	14.3	33.3	19
Suprasegmental	tone	11.1	20	8.9
	voice quality	44.4	22.2	-22.2
Vowel	vowel quality	8.6	3.8	-4.8

Table 2.9: Change in percent of papers in each subtype that use qualitative analysis

Qualitative analysis (Table 2.9) trends upwards for most the categories in this sample, and the most marked increase is in sonorant analyses, which see higher rates in both subcategories. The only consonant category to see a decrease is obstruent manner of articulation studies, which still have a relatively high overall rate of qualitative description. This could be due to a trend towards quantitative analysis especially for lenition, one of the main types of study in this category.

There is a large decrease in qualitative analysis for voice quality, and a slight decrease for vowel quality. The overall rate of qualitative analysis for vowel quality is very low across all categories, indicating that even in the earlier sample there is not much need for qualitative analysis of vowel quality. The highest rate of qualitative analysis is analyses related to sonorant manner of articulation, where two-thirds of studies qualitative analysis.

An increase in qualitative analysis across consonant categories indicates that there is generally less confidence in those measures, particularly for sonorants, when compared to vowels. The sharp decrease in qualitative approaches to voice quality coincides with the establishment of acoustic measures of voice quality, and the availability of software for analyzing voice quality.

Category	Subtype	1999	2019	Change
Obstruent	manner	6.4	25.8	19.4
	place	9.4	10.3	0.9
	voicing	3.3	3.8	0.5
Sonorant	manner	16.7	0	-16.7
	place	10	17.6	7.6
Suprasegmental	tone	0	9.8	9.8
	voice quality	6.2	3.6	-2.6
Vowel	vowel quality	3.1	4.6	1.5

Table 2.10: Change in percent of papers in each subtype that use articulatory analysis

Finally, there is a higher rate of articulatory measurement (Table 2.10) in consonants relative to vowels. There are many reasons for using articulatory studies, but the hardware, software, and technical skill necessary to conduct these studies mean that they are more often reserved for cases where acoustic analysis alone is insufficient. The biggest increase in articulatory studies is in analyzing manner of articulation for obstruents, while the biggest decrease is in studying manner of articulation for sonorants.

In general, trends in all three alternative measurement strategies indicate that there is less confidence in the acoustic measurement of consonants than vowels- consonants are more likely to be qualitatively analyzed, have articulatory measurements, and use acoustic measurements of the neighboring vowel. While direct causality cannot be established by this survey, places where a decrease in alternative methods is observed can be related to the emergence of specific accepted acoustic methodologies for the given category in question.

The low values of all three measures across vowels indicate that the measurement of vowel quality has been fairly stable across the period of the study, while the decrease in alternatives to direct acoustic measurement of voice quality indicates that there is likely an increase in confidence in acoustic measurements in these cases. Both of these facts coincide with a robust general intuition of best practices in vowel acoustic measurement.

These findings, taken together with the finding in the previous section: that the field is trending towards more vowel analysis and less consonant analysis, paints a concerning image. If there is more study of vowels, which are more well understood, and less study of consonants, then the methodology for vowels will outpace that of consonants, reinforcing this imbalance in confidence in acoustic measures.

This is already evident in the existence of many discussions of formant extraction techniques and normalization, while there is much less literature in the field when it comes to methodological development for consonants. In particular, there is broader acoustic variation in consonants, which presents a challenge for traditional approaches of justifying each measure based on the acoustics and theory of a particular contrast. This a) increases the amount of work needed to validate and develop metrics for a particular contrast and b) makes it difficult to compare between categories that have different established acoustic measures.

2.3.3 Types of parameters used in phonetic contrasts

Now that the overall trends in acoustic analysis have been established for the major categories of interest, the next step is to examine the types of phonetic parameters that are used in said acoustic analyses. Table 2.11 gives the proportion of each category of feature used for five consonant-related categories in the sample.

Obstruent	Measure	1999	2019	Change	Sonorant	Measure	1999	2019	Change
Manner	formant	55.6	12.5	-43.1	Manner	formant	37.5	33.3	-4.2
	other spectral	25.9	37.5	11.6		other spectral	37.5	0	-37.5
	pitch	3.7	12.5	8.8		pitch	-	-	-
	spectral moments	3.7	25	21.3		spectral moments	12.5	33.3	20.8
	voice quality	11.1	12.5	1.4		voice quality	12.5	33.3	20.8
Place	formant	28.6	14.3	-14.3	Place	formant	56.2	66.7	10.5
	other spectral	28.6	8.2	-20.4		other spectral	18.8	11.1	-7.7
	spectral moments	42.9	77.6	34.7		spectral moments	25	0	-25
	voice quality	-	-	-		voice quality	0	22.2	22.2
Voicing	formant	13.6	8.7	-4.9					
	other spectral	9.1	21.7	12.6					
	pitch	27.3	34.8	7.5					
	spectral moments	0	21.7	21.7					
	voice quality	50	13	-37					

Table 2.11 Change in the frequency of each measure type across consonantal categories

Usage of spectral moments increases across all consonant comparison types, with the exception of sonorant place of articulation studies. The largest increase is in obstruent place of articulation analyses, where, in 2019, 77.6% of papers used spectral moments in the acoustic analysis. This coincides with anecdotal observations that spectral moments are common practice and are especially dominating acoustic analysis of obstruents. Spectral moments appear to be used more often in other acoustic categories, also coinciding with an intuition that spectral moments are regarded as generally useful. The exception to this is in sonorant place of articulation, where spectral moments are less commonly used in favor of other strategies, namely formants.

In contrast formant measurements are less common across almost all consonant categories, in particular for obstruent categories. However, there is an increase in the usage of formants for sonorant place of articulation. There is also an increase in voice quality measures used for sonorant consonant phonetic analysis, whereas voice quality measures decrease again for obstruents. This seems to show two different patterns emerging in the acoustic analysis of consonants, depending on this broad phonetic division.

The use of alternative spectral measures, including innovative measures, has decreased for sonorants and increased for obstruents. Of these measures, intensity is the most common one across all consonant categories- intensity measures, especially measuring the relative intensity between the neighboring vowel and a consonant, can sometimes be used as an indicator of differences between manners of articulation. However, measures such as peak frequency and spectral tilt only appear in a very few papers.

Vowel-related categories (Figure 2.12) show a different pattern within the survey time period. Tone analyses see an increase in the use of voice quality measures and a decrease in the relative use of pitch, although pitch is still a very common measurement. Voice quality also has an increase in voice quality measurements, which can be expected given the relatively recent establishment of robust acoustic measures of phonation. There is very little change in the kind of analyses used for vowels, which is dominated generally by formants. These data show how the acoustic analysis types for vowels are relatively well established, and have shown much less change than other types of analyses, consistent with findings for vowels in other areas of this analysis.

Alternative spectral measures for vowels have a very low rate, where intensity is occasionally used in these vocalic contrasts. For voice quality, there is also a decrease in alternative measures, where in earlier years a compound measure of voice quality was used, which appears to be replaced in this sample by the establishment of more transparent voice quality measures.

These results highlight a few key findings in the evolution of methodology in acoustic phonetics. Practices for consonants continue to change, while those for vowels are more stable. Where formants can be used, they are becoming the preferred approach, while in other cases spectral moments are used. Voice quality measurements have become quite established, and

Category	Parameter	1999	2019	Change
Tone	formant	0	3.3	3.3
	pitch	100	70	-30
	voice quality	0	26.7	26.7
	formant	0	11.3	11.3
Voice Quality	pitch	30	12.7	-17.3
	voice quality	30	64.8	34.8
	formant	86.6	84.2	-2.4
	other spectral	40	11.3	-28.7
Vowel quality	formant	86.6	84.2	-2.4
	other spectral	6.1	2	-4.1
	pitch	6.1	4	-2.1
	spectral moments	0	2	2
	voice quality	1.2	7.9	6.7

Table 2.12 Change in the frequency of each measure type across vowel categories

while limited in applications (are only used for voice) are considered quite robust in these cases. These results demonstrate that the field tends to cluster around a handful of common measures, supporting the claim that the actual usage of acoustic measures is limited to a relatively small subset of total possible approaches.

2.3.4 Trends in measure sampling

Another important set of methodological practices consist of how an acoustic measure is taken over the time course of the target segment. Table 2.13 reports the usage of different measurement timing strategies in the sample.

Notably, the use of trajectories, or a time series of measurements per segment, is increasing across many categories. This may be based on a recognition that speech is an inherently dynamic process, and the increase in access to statistical tools for measuring time series. Formants, which are traditionally associated with single midpoint or target measurements, see a slight increase in papers reporting trajectory or average-based measurements, although point measurements at individual time points are still the most common approach.

There is also a decrease in studies that don't specify how something is measured- this indicates a larger emphasis being placed on transparent and thorough documentation of the methodology. There is also a general increase in point measurements, which is most pronounced in the measurement of spectral moments, such as measurement at a specific point (e.g. midpoint

measurement). Voice quality and spectral moments see a decrease in averaged approaches, but they diverged in what replaced them - point moments for spectral moments and voice quality for trajectory.

The trends in measure sampling indicate that for spectral moments and voice quality, there is an emergence of best practices around sampling strategy- point measurements are most common for spectral moment analyses and trajectories are used more often for voice quality analysis. The fact that these measures result in different common best practices is another indication of how best practices evolve independently across different acoustic categories and may result in very little direct comparability or transferability of these measures to other conditions.

Measure	Sample	1999	2019	Change	Measure	Sample	1999	2019	Change
Formant	not specified	24.3	12.6	-11.7	Spectral moments	not specified	33.3	13.2	-20.1
	other	11.2	14.6	3.4		other	0	1.9	1.9
	point	46.7	38.2	-8.5		point	0	41.5	41.5
	target	2.8	1.5	-1.3		target	13.3	1.9	-11.4
	trajectory	6.5	15.1	8.6		trajectory	0	1.9	1.9
	window	8.4	18.1	9.7		window	53.3	39.6	-13.7
Pitch	not specified	16.1	15.2	-0.9	Other spectral	not specified	16.7	20	3.3
	other	25.8	15.2	-10.6		other	43.3	40	-3.3
	point	6.5	17.4	10.9		point	6.7	8	1.3
	trajectory	32.3	30.4	-1.9		target	23.3	8	-15.3
	window	19.4	21.7	2.3		trajectory	0	12	12
Voice quality	not specified	31.6	11.2	-20.4		window	10	12	2
	other	10.5	3.8	-6.7					
	point	5.3	15	9.7					
	target	5.3	7.5	2.2					
	trajectory	15.8	47.5	31.7					
	window	31.6	15	-16.6					

Table 2.13 Change in percent of papers in each subtype that use a duration-based measure

2.3.5 Measurement across manners of articulation

Of particular interest to this survey is how acoustic measurement is approached by studies regarding multiple manners of articulation, given that these categories sometimes have vastly different acoustic properties. Despite the large sample size in this survey, there are relatively few studies in this category, so data are pooled across both study years for this analysis. Figure 2.1 gives the most common strategy (or strategies, in case of a tie) for addressing contrasts in each combination of manner of articulation. The shade of the cell indicates the number of studies in that category. There is a distinct imbalance in the data towards measuring vowel-sonorant alternations over all other combinations. Within-category studies see a higher number of studies

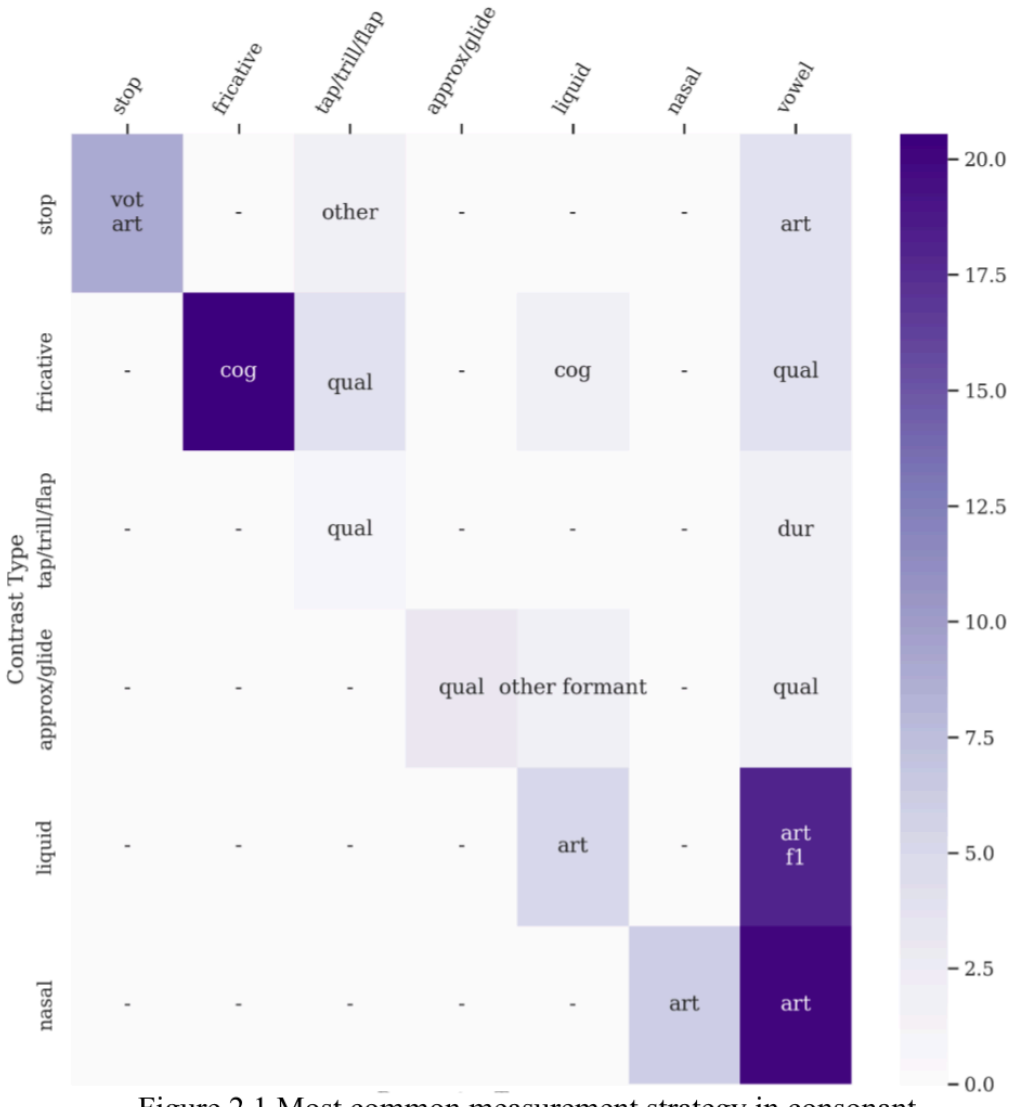

Figure 2.1 Most common measurement strategy in consonant analyses. The shading of each category indicates the number of papers in that category

regarding fricatives and vowels than other categories. While there may be multiple reasons for this, this also coincides with stronger methodological intuitions regarding how to measure these categories, and further suggests that methodology may be influencing the kinds of studies that are published in the field.

Almost all studies that included multiple manners of articulation included either articulatory or qualitative measures in the top methods used. The exceptions are formants, used in some sonorant contrasts, and CoG, used in some obstruent contrasts. This again reflects the confidence of the field in these two measures as particularly robust, and an intuition that best practice is to use formants when possible and spectral moments otherwise. Studies involving multiple manners of articulation are a prime opportunity to present innovation in acoustic methodology, but it is more common to observe alternative methods such as qualitative analysis than proposing a new quantitative method. This suggests that the default behavior when presented with a contrast with no established acoustic methodology is to use non-acoustic methods rather than innovating new measures. This again highlights the potential utility of developing robust acoustic measurement in phonetics in increasing the number of contrasts that can be quantitatively investigated.

2.4 Conclusion

This study has found, by examining the data from multiple angles, that there is generally less confidence shown in consonant parameters, particularly sonorants- they see a relative decrease in representation and a general increase in alternative measurements. In addition, consonants see the most non-spectral parameters, and a shift in parameterizations and measurement methods over the years.

These data also corroborate the general intuition of best practices in the field. Formants, spectral moments, and voice quality measurements are the most common. Generally, formants are used for vowels (and sonorants), f0 for tone, and moments for obstruents. These parameters are derived from the signal differently and are interpreted in contrast-specific ways. These mean that there is no comparability across these categories, and that there are limits on how descriptive the parameters are of the spectrum they describe. This is signified in the results for manner of articulation, where qualitative and articulatory measures were most frequent across manners of articulation in part because there is an absence of reliable acoustic correlates.

The most frequently used parameters also correlate with the most often studied contrasts - formants are used for vowels frequently and there is a marked increase in the proportion of studies studying vowels. There are also a number of studies targeting fricative place of articulation analysis, and spectral moments happen to be another established acoustic measurement. Trends in voice quality analysis also follow the emergence of robust voice quality measures. For these common measurements, we also see the emergence of common practices in terms of sampling strategies as well. While a more targeted study would be needed in order to properly establish a causative effect, if the availability of phonetic measures impacts the kinds of

studies that are done in the field, then it becomes even more important to quickly address gaps in effective acoustic measurement in the field, to expand the kinds of scientific questions that can be explored in the field.

These points pose serious concerns for the continued growth of acoustic measurement phonetics, especially in quantitative investigation of less common contrasts. In particular, in the absence of further development of acoustic measurements, it is likely that the patterns illustrated here will continue, resulting in a distinct imbalance in the scientific understanding of more vs. less commonly-studied contrasts. There is particular concern for contrasts that are less common overall in the system, which are still of significant scientific interest, but have less methodological research dedicated to them.

Four categories of methods have emerged as of particular interest as a result of this survey. Two of them, pitch and voice quality, are relevant particularly to suprasegmental research, and while there is still room for further refinement, capture the most important acoustic features of the sounds. The remaining two, formants and spectral moments, are primarily for segmental contrasts. All four of these categories are characterized by having tools available to facilitate measurement via software. Three of the four also have significant grounding in articulatory interpretability and the fourth (spectral moments) has also had some work in that area. In addition, they have all demonstrated at least some utility in separating out categories of interest for further analysis. These properties will be discussed further in Chapter 5 and appear to be important features for any further development of phonetic features.

Now that the general patterns of usage of acoustic measurement has been established, as well as qualitative discussion of the values of specific parameters, this dissertation now turns to developing a way to quantitatively compare acoustic measurement approaches, which will allow for a direct, stable understanding of how different acoustic measurements perform and further set the stage for consistent development and validation of new approaches to acoustic measurement.

Chapter 3: *The role of acoustic context in discriminability*

3.1 Introduction

Chapters 1 and 2 established current patterns in acoustic measurement in phonetics, and highlighted the need for further development in this area. However, before proposing new phonetic features, it is important to outline key considerations for designing a successful approach to acoustic measurement.

In particular, there are two conditions that must be met for an acoustic measure to be effective. First, there must be acoustic differences between the categories of interest, and second, the representation needs to meaningfully capture those acoustic differences. This chapter will investigate the first of these two statements, and Chapter 4 will address the second statement by presenting a quantitative benchmark of acoustic measurement performance.

Two sounds being perceptually distinct does not necessitate acoustic differences in the target segment itself. Speech perception happens in context, and it has been shown that neighboring segments can have an effect on perception (e.g. Stilp 2019, LaRiviere et al. 1975). This opens up the possibility that adjacent acoustic information is necessary in order to satisfactorily detect differences in some phonetic categories, and it is possible that further methodological development would also require consideration of context.

This chapter investigates the contribution of adjacent phonetic information to discriminability across major phonetic categories. Because conventional acoustic features intentionally capture a small slice of the total information in the sound, this chapter turns to neural net representations as a way to capture overall acoustic information in the segment. In a neural net, the input is passed through each layer, where it is progressively transformed into a

representation that is further removed from the original data. The model then produces a set of outputs for a particular task, e.g. speech recognition. The values from these intermediate layers have been usefully repurposed for uses other than those that the model was originally trained for, such as phoneme classification, and have been increasingly scrutinized to understand what types of phonetic content they contain. The primary purpose of these models is to perform a task, such as phoneme or word identification, and the models are iteratively trained by adjusting parameter values in order to optimize outputs on this task. These models contain high-dimensional internal representations that contain complex transformations of the original acoustic data.

Most neural net architectures applied to language use some type of context-sensitive architecture, and representations for individual frames have been found to contain enough information to determine word identity (Sanabria et al. 2023). These very complex representations retain a large amount of information from the original acoustic data, and represent the most complex acoustic representation available. Thus, if there is an acoustic difference present between two categories, a neural network representation is most likely to capture it without specifying *a priori* what that difference might be, which will allow for this study to investigate the general role of acoustic context without requiring significant abstraction.

Neural network-derived representations have also been related to traditional acoustic features. Studies in this area of research are typically designed to investigate a specific property of the representation. Neural nets appear to have meaningful relationships with hand-engineered acoustic features including formant values, in particular F1/F2 (Reira et al. 2023; Vu et al. 2014), as well as f0 and CoG (Shah et al. 2021). They have also been related to Mel Frequency Cepstral Coefficients (MFCC), the most common speech representation in deep learning. In particular, when probing multiple layers of a neural network, the relationship between neural net representations and MFCCs are most similar in early layers, diverge (but phonetic information as measured by the probing task peaks) in later layers, and converge again with the MFCCs in the final (task specific) layers (Pasad et al. 2021; Pasad et al. 2023). Masking has been used as a probing experiment for CNNs to highlight the important parts of the spectrogram for classification (Ferragne et al. 2019)

Another approach has been to use neural nets to cluster speakers and varieties (Bartelds and Weiling 2022; Bartelds et al. 2022), to identify organization of speakers and languages (de Seyssel et al. 2022), or to identify speaker and language identity (Fan et al. 2021). These studies indicate that neural net representations appear to encode information about speaker and language identity beyond that of traditional acoustic features, which strive to abstract away from these properties.

Neural nets have a wide variety of architectures that have been developed for specific advantages, which means that properties detected in one model may not hold true for another. For this reason, it can be advantageous to investigate a specific architecture in some depth in order to get a complete understanding of how it behaves with regard to phonetic information. One architecture that has been investigated to some extent in the literature is wav2vec2.0 (Baevski et al. 2020), a transformer architecture that has featured prominently in phonetic

representation studies. This model takes raw audio (waveforms) as input values. These values are then fed through two sub-networks. The first includes several layers of convolution to extract local features. The second, the transformer, uses the attention to incorporate both short- and long-distance context into the representation.

Wav2vec2.0 utilizes a self-supervised pre-training step, which essentially allows for a significant amount of training on unlabelled data. This allows the model to incorporate a larger amount of data since it does not need to be labeled to be included in this step. The task in this case is the prediction of the identity of a segment of the audio that is masked- the typical length of a mask is 300ms. Typically, phonetics research is concerned with a shorter timescale, and this should be taken into account when working with pre-trained models. Pre-trained models can then be shared and subsequently fine-tuned by different researchers for different datasets or tasks with a much smaller amount of data.

The first module of wav2vec is often referred to as the feature extractor, which generates features from the original waveform to then be input into the transformer. This module is a convolutional neural net (CNN). In general, the early layers of convolutional neural nets have been found to act as matching filters (Palaz et al. 2019; Hoshen et al. 2015). This means that although they are not explicitly instructed to do so, they are similar to a cepstral transformation in the first few layers.

These early representations may be of particular interest in the examination of phonetic features, for example, f0 appears to be represented linearly, while formants were represented in a grid fashion (Choi and Yeo 2022). The representation also correlated more closely with MFCCs in later feature extractor layers (Pasad et al. 2021), with other audio features in the earliest transformer layers (Shah et al. 2021), and with articulatory electromagnetic articulography (EMA) representations (Cho et al. 2023)

Studies tend to focus on how each layer functions, for example by taking the representation at each internal layer of the transformer and feeding it into a secondary classifier. Neural representations perform differently in this context for different phoneme types: vowels perform the best, and other segments, in particular dynamic segments such as affricates, perform much worse in the secondary classifier (English et al. 2022), classification for obstruent sounds have also been found to be worse than sonorant sounds (Ma et al. 2021).

From these probing studies, a few themes emerge. First of all, the studies that separate out results for different phonetic categories find that there are differences based on phonetic category. Vowels perform better than fricatives, and dynamic sounds such as affricates and stops perform the worst. These differences align with previous work in other areas of phonetics and the general understanding that some contrasts are easier to measure than others.

In addition to the general research goal of quantifying the importance of acoustic context in phonetic representations, this experiment addresses an important consideration for the use of neural networks as acoustic representations. Neural nets are known for being able to include short- and long- distance dependencies in the representation, and have significant potential for advancing acoustic measurement in linguistics. However, there has only been a small amount of

work addressing how context is being incorporated into acoustic representations, in particular across different contrasts. Given the focus in phonetics on local acoustic context, any research incorporating neural network outputs into acoustic representations will require identifying and accounting for the influence of context into phonetic representation.

3.2 Methodology

The study in this chapter will probe a neural net in order to address the question: How much does contextual information improve discriminability across phonetic contrasts? The approach in this study is to extract intermediate representations from a pre-trained variant of the wav2vec2.0 model for datasets that are identical except for the amount of acoustic context included in the input to the model. A secondary model trained to classify for place of articulation within four phonetic contrasts (vowels, nasals, stops, and fricatives) will identify differences in discrimination across phonetic contrasts and amount of acoustic context.

3.2.1 Data and processing

The data used in this analysis is a subset of the CommonVoice corpus. The CommonVoice (Ardilla et al. 2020) project is an open source data set, where volunteers record their speech into a computer or phone by reading sentence prompts. VoxCommunis (Ahn and Chodroff 2022) produced time aligned phone-level representations across a significant subset of the Common Voice corpus using 38 languages by force-aligning and validating the sentence level transcriptions.

The number of hours of languages in this dataset varies widely (from <1 to >100). In previous literature (e.g. Vu et al. 2014, English et al 2022), a few hours (2-5 hours) of data appears to be the target for previous probing experiments within a given language. The other concern is that some segments are much more frequent than others, and this can result in imbalanced classification scores in a secondary classification task. To balance classes for phoneme frequency, 1000 tokens for each phoneme were randomly selected, and phonemes with fewer tokens were excluded from the analysis. Languages where less than half of one or more phonetic categories had sufficient tokens were excluded entirely from this set. This approach explores effects across a broad set of languages in a relatively comparable set, without the impact of frequency on the classification scores, although this means that some infrequent categories will not be represented in the target sample, and this is an area for future consideration.

This study focuses on four major phonetic categories: vowels, fricatives, nasals, and stops. These categories are typologically common and represent some key general types of phonetic sounds. Vowels and nasals are both sonorant sounds, characterized by having resonant frequencies in the vocal tract, while stops and fricatives are characterized by turbulent airflow. In addition vowels and fricatives are relatively temporally stable sounds, while stops have more dynamic articulation consisting of a closure and release and nasals also tend to rely more on

context for perception. Table 3.1 gives each language used in this study and the number of

Language	Fricative	Nasal	Stop	Vowel
Bashkir	11	3	7	9
Belarusian	14	6	8	5
Czech	8	4	8	10
Dutch	7	3	6	14
Georgian	5	2	10	5
Greek	8	2	4	5
Italian	9	5	11	8
Kyrgyz	5	3	6	8
Marathi	5	4	14	9
Polish	9	4	8	8
Portuguese	7	2	8	12
Romanian	5	2	6	7
Swedish	6	3	6	16
Tatar	8	3	6	10
Thai	3	3	9	17
Ukrainian	10	3	7	6
Uyghur	6	3	7	8

Table 3.1: Languages and number of categories in the dataset

segments per category that met the criteria described here.

For each segment in the sample, a window centered around the midpoint of the target segment was selected, where the size was determined by a ratio of the total duration of the segment. This method prevented the model from incorporating any more than the intended amount of acoustic information into the representation. The four window sizes are as follows (illustrated in Figure 3.1):

- **50-percent**- a window 50% the length of the segment centered around the midpoint of the segment. This avoids portions of the segment where coarticulatory influence from neighboring segments may be expected.
- **100-percent**- a window including the entire segment, including those areas expected to be influenced by neighboring segments.
- **150-percent**- the entire segment and a small amount of segmental data on either side. The total size of the window is 1.5x the length of the target segment.

- **300-percent**- entire segment and more data on either side- the total size of the window is 3x the length of the target segment

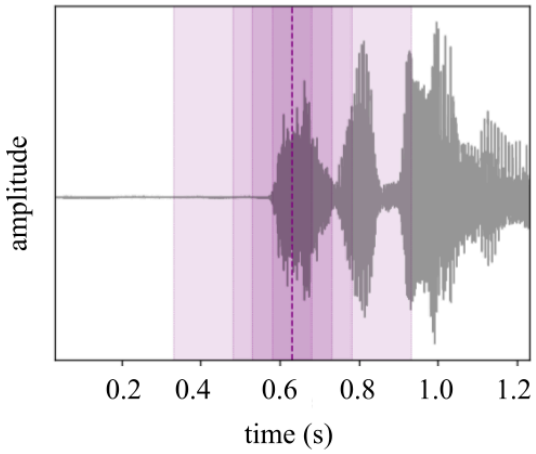

Figure 3.1: Window sizes for an example stimulus

This process results in a total of 16 conditions per language: 4 phonetic categories x 4 window sizes in order to address how acoustic context affects representations.

3.2.2 Model and representation extraction

The model used in this study is wav2vec2-xlsr (Conneau et al. 2020), a multilingual version of wav2vec 2.0 (Baevski et al. 2020). XLSR is a variant of the original wav2vec 2.0 architecture trained on a multilingual dataset of 53 languages and fine-tuned on more multilingual data including the Common Voice corpus, from which the data in this study originates. In the fine-tuning step, weights for all layers are updated. This contrasts with the base version of the wav2vec 2.0 model, where the majority of layers are frozen and only the final few layers are updated in fine-tuning.

For each of the 16 conditions described above, representations were extracted from the neural net model. Each audio sample was resampled to 16000 Hz to match with the expected input of the model. Then the audio was passed through each of the layers of the model including both the convolutional and transformer components, and all intermediate representations were extracted. For each layer, the representation for the middle frame of each token was extracted in order to identify how much additional context influences the most central representation of the target segment. This also guarantees that there will be an equally sized resulting representation for all tokens, regardless of the length of the original input. These representations were then

passed to the secondary classifier to investigate the effect of context in different phonetic categories.

3.2.3 Probing experiment

The overall focus of the probing task is to identify how well the representation performs in assigning place of articulation within each contrast x window size condition. This takes the form of a discrimination task that is performed within each condition.

The task in this study is pairwise classification for all phoneme pairs within each contrast. Pairwise classification has been used elsewhere as a way to evaluate the quality of representations for speech (Schutte 2009). This design is suited to multi-language comparison because it controls for different numbers and types of phonetic contrasts within each language.

For each phonetic category, a smaller neural network was trained to perform pairwise discrimination between all within-category pairs. High discriminability in the smaller window size conditions suggests that there is enough information to show the difference between them and supports the usefulness of phonetic representations for acoustic analysis. In particular, if there are increasing scores with more context, then the role of contextual phonetic information is indicated to be more important for this dataset, and if there are decreasing or level scores with increasing context, then there is more support for using less acoustic context in phonetic analysis.

3.3 Results and Discussion

3.3.1 Effect of phonetic category

The results for each phonetic category in the discrimination task are given in Figure 3.2. In the convolutional layers (0-6), there are two patterns of behavior across contrast types. Vowels and fricatives see high accuracies even in the first layers of the model, and the accuracy begins to level out towards the end of the convolutional layers. This suggests that the local features learned by a convolutional architecture are quite good at picking up on differences within these two categories. Nasals and stops on the other hand see overall lower accuracies in the convolutional component of the neural network, and see a steady increase in accuracies across CNN layers, with no leveling out towards the later layers. This suggests that the same types of local features are less effective in distinguishing between these types of sounds.

Within the transformer layers (layers 7-30), there are different behaviors across phonetic contrasts as well. Vowels and fricatives behave similarly and quickly reach a maximum point early on in the transformer, with only a slight improvement over the convolutional module. However stops and nasals see a much bigger benefit of the transformer, where they still have overall lower accuracies than the other two categories, but see a bigger increase as a result of adding layers of the transformer. Larger window sizes narrow the gap between the two sets of contrasts. This suggests that dynamic segments do appear to require more acoustic context in

order to have high discrimination relative to less dynamic segments, which will be discussed further in the next section. Generally speaking, differences between different phonetic categories support the idea that some categories are acoustically easier to distinguish than others, and therefore it may be more difficult to derive concise phonetic representations for certain types of phones.

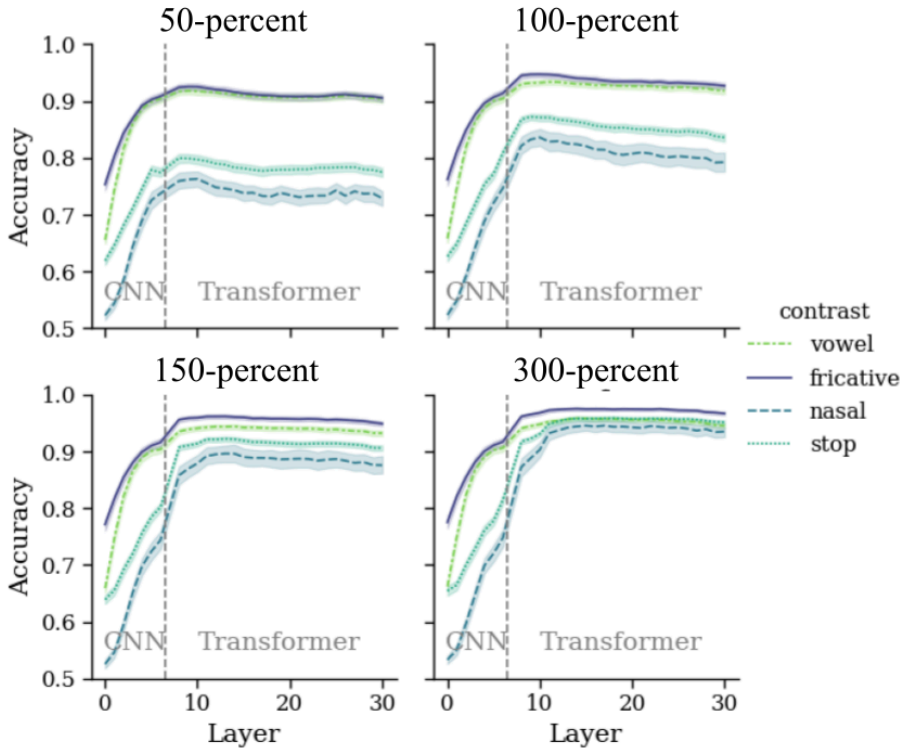

Figure 3.2: Aggregated pairwise discrimination accuracies across all model layers within each language for each phonetic contrast under study. The grey line indicates the transition between CNN and transformer layers.

3.3.2 Effect of window size

The other key manipulation in this study is the incorporation of different window size conditions. Figure 3.3 shows the effect of window size across all phonetic categories. For the convolutional module (left of the vertical line) there is essentially no difference between window sizes, which is congruent with the structure of the module. Convolution is an architecture that specializes in local features, and this approach indicates that the types of filters used in convolution are small

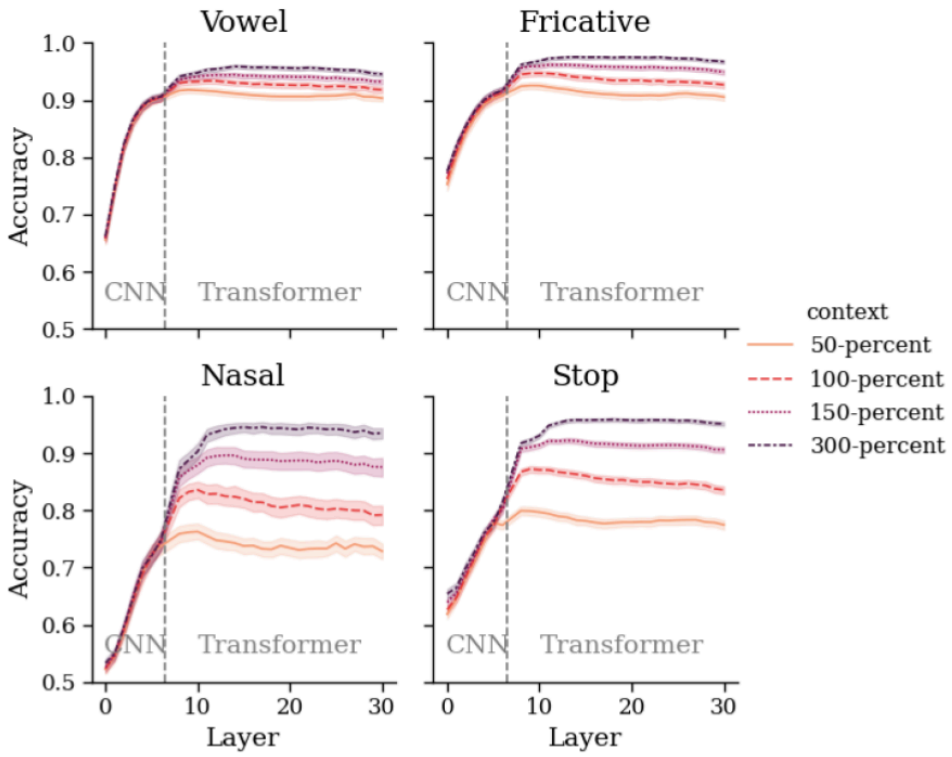

Figure 3.3: Aggregated pairwise discrimination accuracies across all model layers within each language for each phonetic contrast under study. The grey line indicates the transition between CNN and Transformer layers.

enough that the acoustic material from outside the middle window of the segment does not make it into the midpoint representation.

Context plays a larger role when it comes to the transformer layers to the right of the figure. There is, across all contrasts, a consistent increase in accuracy across all layers with larger window sizes. However, this also interacts with phonetic context. For vowels, window size has the smallest effect on the results, while there is a slightly larger effect for fricatives, a larger effect for stops, and the largest for nasals. This aligns with general observations about these segments in the field: vowels typically have less acoustic measurements associated with neighboring segments than fricatives, which have less than nasals.

These results suggest that acoustic context is more important in some phonetic contrasts than others - with nasals it seems more useful than it is for fricatives and vowels, perhaps because they already have plenty of acoustic information encoded in the segment itself. While context is useful for these contrasts, it is still possible to differentiate within these categories to some extent even without context. Different phonetic categories require different amounts of

context in order to reach equal levels of discriminability, and this coincides with challenges in the field to identify satisfying acoustic features for some phonetic categories, such as stops and nasals, and supports the idea that the segment itself does not contain as much acoustic information as other phonetic categories. Further development of good phonetic features, then, will likely require at least some incorporation of context in order to achieve results equal to more well-established acoustic methodology. This presents additional technical challenges to the field, but also a significant opportunity to develop robust, generalizable phonetic features.

3.3.3 Transformer Layer 5

It appears from Figures 3.2 and 3.3 that the largest gains in discrimination are made in earlier layers of the transformer. Figure 3.4 gives the aggregated pairwise results for transformer layer 5 across all conditions in the study. Lighter lines show the results for individual languages, while the darker lines show aggregate values. The lighter lines, although there is some variation between languages, show the same general trends across languages for window sizes - this suggests that these representations are picking up on general acoustic features that are useful for distinguishing between contrasts across all languages.

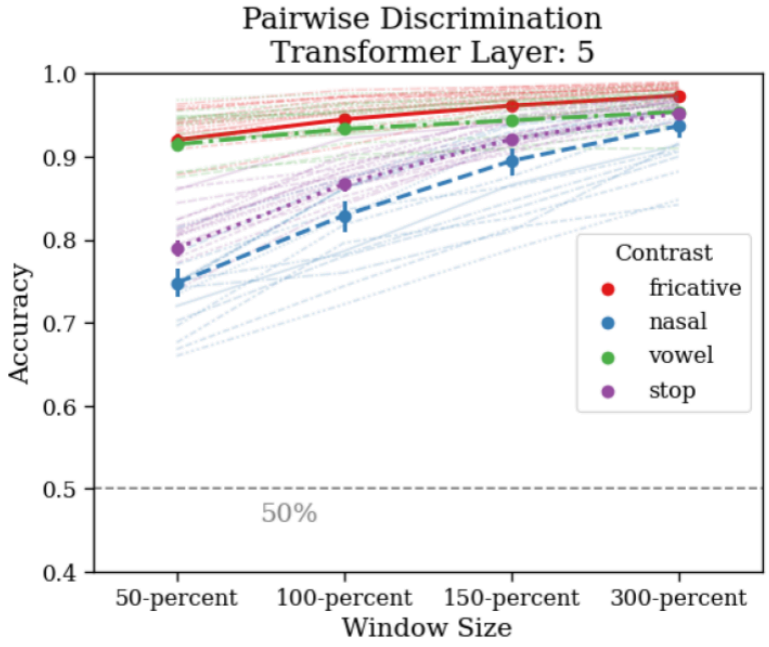

Figure 3.4: Aggregated pairwise discrimination accuracies across all contexts for layer 5 of the transformer

At layer 5, both fricatives and vowels perform with approximately the same high accuracy for all window sizes, suggesting that using the smallest window size is equally useful, whereas for nasals and stops there is still a marked benefit in using the larger window sizes. However, in the 300-percent window size condition, there is very similar performance across all three categories. This suggests that if one is interested in a single definition of a representation that works across all phonetic contrasts using this model, using the largest window at layer 5 is sufficient and provides a starting point for further investigating acoustic representations. In the next chapter, this neural representation will be quantitatively compared to traditional acoustic features to form a baseline analysis of the performance of acoustic measurement.

3.4 Conclusion

This chapter investigated the role of acoustic context in improving pairwise discriminability between phones in a multilingual dataset. The main result of this study is that there is a general increase in discriminability across layers, up to a relatively early layer of the transformer. Phonetic category does play a significant role in phone discrimination. In particular, if one considers the difference between more static (vowels/fricatives) and more dynamic (stops/nasals) sounds, it appears that context is required in different amounts in different contrasts, where some amount of acoustic context seems necessary to get equally well performance on the more dynamic categories.

Based on intuitions in the field, the order of discriminability, from high to low, is predicted to be vowels, fricatives, stops, and finally nasals. So, therefore, the fact that fricatives have comparable discriminability scores to vowels is interesting, because it highlights that particularly for fricatives, there is ample acoustic information in the acoustic signal, but that traditional acoustic approaches do not seem to adequately pick up on these differences.

Although the inputs had different window sizes, the representations used in evaluating these conditions all used the midpoint representation, rather than some segment-averaged approach. This means that the acoustic information of neighboring segments that is being revealed in the 300-percent case is being incorporated into the midpoint representation. This means that there is the opportunity for neighboring information to influence the representation and it is important to consider when using neural net representations - longer range context that may not be intended for use in the representation may be affecting the output representation in ways that may have been underestimated previously.

In general, the point at which all three phonetic categories in this dataset experiment converge is at the 300% window size and at layer 5 of the transformer. At this point, it seems there is enough context and complexity for each of the phonetic categories to be well represented. While this may be more complex than is necessary for fricatives and vowels, it performs well even on nasals, which are traditionally difficult phonetic categories to describe acoustically. Future work would include applying dimensionality reduction and visualization

techniques to better understand how these representations might relate to common phonetic and phonological properties.

Overall, these results suggest that without acoustic context, there is a ceiling for the amount of information that can be held by representations, especially for certain types of sounds. However, the majority of traditional phonetic representations do not include contextual information, and this might partially explain the long-term struggle to define satisfying acoustic features that capture relevant information in some phonetic categories, such as nasals. While it poses technical, theoretical, and computational challenges, the results of this study emphasize the need to find ways to treat speech as a dynamic process, and consider how context may be incorporated into acoustic analysis and measurement in phonetics.

Chapter 4: *Quantifying the performance of acoustic measures*

4.1 Introduction

Current common practices in acoustic measurement, while effective, cover only a small portion of the total phonetic variation of interest to the field (Chapter 2). While further development of features is desirable, there are many potential avenues of innovation of new features. Chapter 1 established the near-infinite number of possible ways to derive features from speech, and the most recent Chapter 3 established the importance of incorporating additional context to generate meaningful representations, which further expands the number of possible avenues for exploration. In order to have effective methodological advancement, however, it is important to establish a systematic approach to identifying which approaches perform better than others.

Past approaches to feature comparisons have taken a few different forms, including training a classifier to discriminate between categories using approaches such as discriminant analysis (Adank et al. 2004; Jongman et al. 2000), logistic regression (Ghaffarvand and Mahdinezhad 2020; Spinu et al. 2018; Styler 2017), or any of a vast number of classification algorithms (Ulrich et al. 2021; Cassidy and Harrington 2004; Ma et al. 2021). A few studies have also tested the model trained on the representation on an unseen subset of the data (e.g. Spinu and Lilley 2016; Gendrot et al. 2019), which helps account for overtraining. Another way to conceptualize the quality of a representation is that a good acoustic representation would aim to minimize variance within a target category, and maximize the distance between categories. This is implemented by comparing the distance in feature space between vs within the target categories, most commonly with Euclidean distance (as in Jannedy and Weirich 2017), or with perceptually-grounded distance metrics (e.g. Ghaemmaghami 1997).

A final way to validate phonetic measures is to discuss how well the feature correlates with a known benchmark. For example, Gradoville (2022) correlated between acoustic measures and perceptual measures of /s/. This approach may also use an acoustic benchmark, such as neural net representations and MFCCs (Shah et al. 2021), neural net representations and formants Reira et al. 2023; Vu et al. 2014), and MFCCs and formants (Hughes et al. 2020). A non-correlational ground truth benchmark of mutual information has also been applied to speech representations (Gump et al. 2020). This approach is limited to phenomena that have an established ground truth against which it can be compared.

Methodological research in phonetics has thus far been primarily concerned with validating a measure in a specific context. While targeted approaches to acoustic measure validation are useful for specific cases, the field lacks a broad benchmark to quantify performance, which limits the ability to compare across different approaches and methodologies. This chapter will present a design that can be applied to any acoustic measurement to evaluate performance across four common phonetic categories: vowels, fricatives, stops, and nasals. This single metric represents performance of the measure across multiple languages, speakers, and phones and allows for a broad comparison of acoustic measurement.

4.2 Methodology

4.2.1 Data

This benchmark uses the same subset of the VoxCommunis corpus (Ahn and Chodroff 2022) annotations to the CommonVoice Corpus (Ardilla et al. 2020) as described in Chapter 3. A few additional considerations for the dataset as it applies to a broader metric of phonetic discriminability performance are described below:

The CommonVoice corpus is a crowdsourced project where volunteers record reading an sentence at a time. These recordings can be made under a variety of conditions, including recording environment and microphone quality, and speakers can choose to contribute as much or as little data to the corpus as they would like. Because of the type of stimuli used, there are relatively few tokens per speaker, and very little control of the surrounding context. Traditionally, high-variability datasets pose a challenge for phonetic measurements by introducing noise in the resulting measurements. However, this is not necessarily an issue for the metric described here. In contrast, a set of measurements that scores high on this task despite variability in the data is also likely to be robust to contextual and inter-speaker variation, a desirable trait in a good phonetic feature.

The other consideration for this dataset is that the data are stored in the compressed .mp3 format. Although .mp3 is a common format, due to the method of compression there can be expected to be some decrease in the acoustic quality. In contrast, an uncompressed format such as .wav is most commonly used in linguistics and is usually the gold standard for acoustic phonetics research. However, .mp3 is very common in machine learning and speech research,

and most large multilingual datasets such as CommonVoice use the .mp3 format, so developing methods that are compatible with these data types is essential to be able to engage productively with these datasets. While there can be some loss of acoustic data quality based on the fact that the data were stored in this format, performing well in these conditions would be another desirable trait in a phonetic feature.

The corpus in its entirety includes time-aligned data from 36 languages, most of which have at least two hours of total speech from >20 speakers. The benchmark uses a subset of 17 languages that have at least 1000 tokens per segment for the majority of segments in each phonetic contrast under study (the same procedure as is detailed in Chapter 3). This is to allow for enough data to understand how that category works acoustically, and, because of the high variation in the dataset, having a number of tokens per segment will be important to establish the acoustic characteristics of the sound. In order to reduce the effect of frequency and class imbalance in the dataset, a 1000-token sample of each segment was randomly selected. This allows for the performance metric to not be biased towards one class because it is much more frequent. Since the goal of this study design is to understand which acoustic measures perform reasonably well in phonetic analysis, a (relatively) small number of tokens more closely approximates the type of data that would likely be the subject of phonetic analysis, as opposed to using all available data.

The advantage of this dataset is the large amount of data, and the broad coverage of a large number of languages and speakers relative to other similar datasets. One consideration for this dataset is that the forced-alignment approach of sentence level transcription may make the analysis a little more challenging. While the languages included in VoxCommunis are orthographically relatively transparent, there can be expected to be some acoustic variation that may be phonetically or phonologically meaningful, but has the same orthographic label, which will introduce some noise to the dataset. In addition, while 17 is a large number of languages relative to comparable multilingual studies, it is still a relatively small sample relative to the total range of language variation, and expanding this dataset would be a worthwhile further endeavor in this line of research.

4.2.2 Phonetic contrasts

The primary purpose of this study is to provide a broad understanding of the quality of current and new acoustic representations in phonetics. To that end, this study will focus on four phonetic categories that capture a broad range of sounds in phonetics: vowels, fricatives, stops, and nasals.

Vowels: Vowels have by far the most well-defined acoustic methods, and vowels are well understood from a theoretical, perceptual, and articulatory perspective. Vowel quality is also a central feature of both experimental and descriptive phonetics, and increasingly commonly studied in the field. Historically vowels have been the easiest category to capture acoustically, and conventional methods can be expected to perform the best on vowels.

Fricatives: Fricatives are a fairly temporally stable sound that have a robust, common measurement approach in common practice (spectral moments). Unlike vowels, the interpretation of spectral moments, and their efficacy in determining the differences between more similar fricatives, is still not well understood (discussed further in Chapter 1). In the survey in Chapter 2, fricatives were the most commonly studied consonant category, indicating a strong interest in quantitative comparison.

(Oral) Stops: Measurement strategies for stops look very similar to those for nasals - the main difference being that they have two distinct phases to articulation, closure and release. Since the closure period is near-silence, acoustic measurements for stops are typically made in the surrounding acoustic material, and measurements made during the stop burst look very similar to those for fricatives. Thus, stops give a good understanding of how acoustic measures work in a more dynamic, obstruent-type phonetic category.

Nasals: Nasal stops have long posed an issue for acoustic measurement due to their different articulatory and acoustic properties to other phonetic categories (discussed in Chapter 1). In the survey presented in Chapter 2, there are very few studies that consider nasals, even though they are typologically common. The studies that did primarily relied on non-acoustic approaches to measurement rather than acoustic approaches. This lack of established acoustic methodology represents an open question in the acoustic phonetic methodology, and finding acoustic measures that perform well on these categories are more likely to perform well in general.

Focusing this analysis on these phonetic categories will form a baseline understanding of the quality of acoustic parameterizations across major acoustic types of sounds, including more sonorant and more obstruent sounds, and more dynamic and more static sounds. There is also variation in how well established best practices in acoustic measurement are for the field. These categories are also relatively typologically common, and have a reasonable number of contrasts within each category, which means it is possible to test these contrasts across a much greater range of languages. In order to properly test how general these features are, multiple speakers and multiple languages will be tested in the study, and, given the variability of corpus data, it is important to have a large number of tokens per each contrast to account for variation and noise in the data.

4.2.3 Acoustic measures

The goal of this study is to establish a general framework against which any set of acoustic representations can be compared. However, as a starting place, this chapter will present the following sets of representations to compare in this analysis:

Formants: Formants are a common approach to any sonorant-type analysis, and the most common acoustic measures used in any sonorant sound. In this case, the first four formants were

extracted for all segments in the dataset. Measurements for formants were taken both at the midpoint and as an average across the entire segment, resulting in two points for comparison. While traditionally formants are not a measure considered appropriate for non-sonorant segments such as fricatives, it is mathematically possible to derive potentially meaningful peaks in the spectrum through the same analysis, so this approach was applied to all phonetic contrasts in the study.

Spectral moments: Spectral moments are the default method for use in the analysis of fricatives, but have also been extended to other sounds (explored further in Chapter 1). This analysis extracted the first four spectral moments on the whole-segment spectrum using a low-pass filter at 300 Hz in order to prevent fundamental frequency from having a large impact in the shape of the spectrum, following the methodology in Rao and Shaw (2022).

MFCC: The third type of features tested is mel frequency cepstral coefficients (MFCCs). These are the most common representations used in speech recognition (Bhatt et al 2020). There are several variations on the implementation (some are discussed in Ganchev (2005)), but the basic premise is to make a spectrum on the mel scale (a perceptual scale of frequency) and then further transform it. This method has also been used in linguistics for vowels (Ferragne and Pellegrino 2010; Bailly and Martin 2014) and fricatives (Spinu and Lilley 2016). Perceptual work has also found a correlation between perceptual distance of MFCCs to perceived spectral shape (Terasawa et al 2012). MFCCs here are included as a more comprehensive acoustic characterization of the segments in this study.

Neural net representation: The final representation compared here is a neural network hidden layer representation. Neural networks that take waveforms as direct input have become very popular in recent years, and the intermediate representation that they generate has been explored for speech. For this analysis, the hidden layer representation from one layer (layer 5) of the Wav2vec2-XLSR53 model (Conneau et al. 2021) will be used as a point of comparison using the methodology described in Chapter 3.

4.2.4 Benchmark metric

Comparison of these measures will focus on performance of this model on a general phonetic classification task. The goal here is to determine how the model performs in a classification task across multiple languages given potentially significant variation: there are many speakers, different microphone/recording qualities, and because of noise expected in force-aligned data.

The classification approach that is taken here is a binary pairwise classification task within each phonetic category under study. For example, if a language has a three-way vowel contrast a-i-u, then there would be three different comparisons: a-i, a-u, i-u. While multiclass classification (as in English et al. 2022) is the most common approach, pairwise classification has also been used in evaluating acoustic representations (Schutte 2009). Pairwise classification

is suited to the multilingual evaluation context in that it does not depend on inventory size of contrast. In addition, while 1000 tokens is a large amount in phonetics, the larger the number of comparisons of a multiclass classification approach, the more data is needed in order to differentiate between all potential classes. While pairwise classification is a less common task in descriptive phonetic analysis, this approach is likely to give a broad understanding of relative performance of acoustic measures across multiple languages and phonetic categories. The pairwise classification algorithm used here is Random Forest, a common approach currently used in the field for discrimination analyses (e.g. Villareal et al. 2020), tend to be robust to overfitting, and are able to capture nonlinear relationships between variables.

4.3 Results

4.3.1 Evaluation on multilingual dataset

The results for pairwise within-category classification aggregated across all languages for each phonetic contrast is given in Figure 4.1. Conventional acoustic features (spectral moments and formants) show fairly similar results in terms of aggregated accuracies. Spectral moments are slightly advantageous for fricatives and stops, while formants perform slightly better for vowels and nasals. This aligns with qualitative intuitions in the field - phoneticians tend to use formants for sonorant sounds and spectral moments for obstruent ones. Overall, there is also a higher baseline on these measures for vowels and fricatives than for stops and nasals, which also aligns with intuitions that these categories are easier to measure with conventional practices in the field.

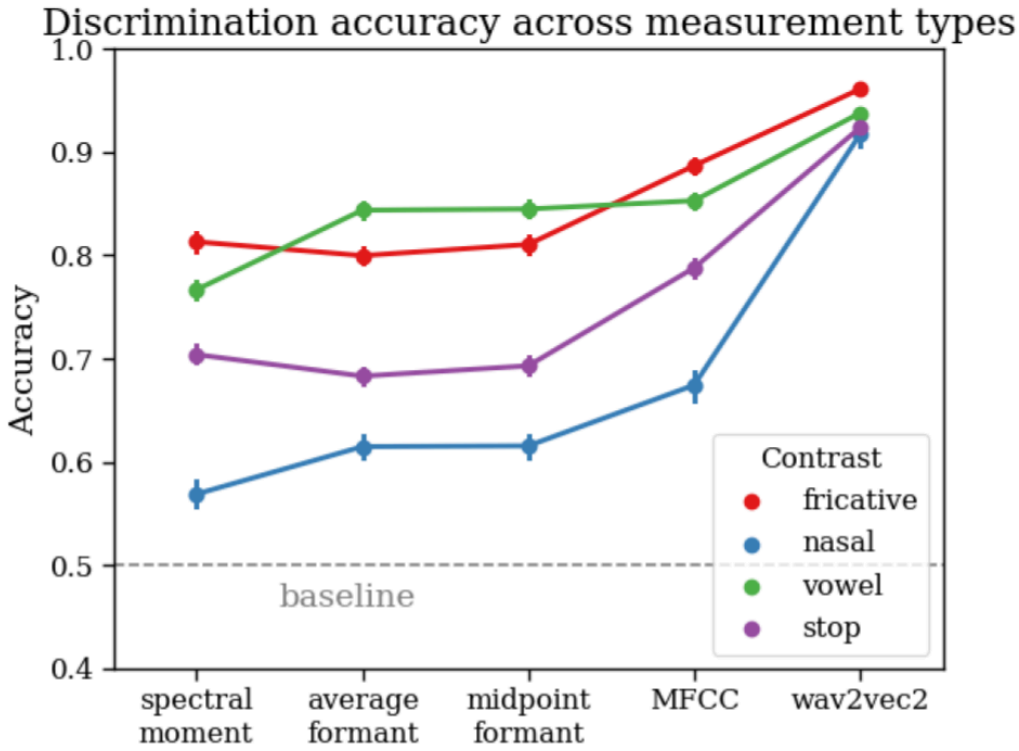

Figure 4.1: Aggregated pairwise discrimination accuracies across all phoneme pairs within each language for each phonetic contrast under study.

Within conventional features, there are two distinct patterns in the sonorant vs obstruent types of sounds included in this study. Spectral moments perform better on fricatives and stops, while formants perform better on vowels and nasals. Spectral moments are a global measurement of the overall shape of the spectrum, while formants are local measures of significant frequencies. The spectra of vowels are of similar overall shapes, where key differences are shown by small differences in the location of spectral peaks. The location of peaks is something that spectral moments wouldn't be sensitive to, so it makes sense that a global measure wouldn't accurately capture these features.

The results here align with general intuitions of the use of phonetic measurements in conventional acoustic features: formants work best for vowels and moments best for fricatives. The fact that the quantitative findings here align with the general intuitions regarding these methodologies in the field suggest that this approach is capable of capturing the relative quality of different acoustic representations for representing different phonetic contrasts.

MFCCs perform better than traditional phonetic features for fricatives, nasals, and stops. This indicates that there is additional acoustic information within the segment that is not captured by conventional acoustic measures and that contributes to higher classification accuracies. However, for vowels there is not an increase, which indicates that formants are equally good as a

much more complex representation for indicating the difference between the phonetic contrasts of interest. This highlights that formants are a uniquely useful measurement specifically for capturing the key acoustic features of vowels.

The final measurement strategy under consideration for this study is a representation taken from the hidden layer of a neural net. This approach is a powerful, complex, and high dimensional representation and there is an increase in accuracy for all four contrast types, almost to maximum accuracy. This means that if given a complex representation with access to context, it is possible to have near perfect discrimination between the phonetic categories in this study. The most dramatic increases are in stops and nasals, which are the categories in the study that can be most expected to rely on dynamic properties and neighboring acoustic context.

4.3.2 Phonetic similarity and classification

One of the discussions around phonetic features is whether they are sufficiently sensitive to distinguish between fine-grained phonetic differences. The mel frequency cepstral coefficients (MFCCs) have been used as a way to generally quantify phonetic distance (e.g. Gerosa et al. 2006). In this case the MFCC representational distance between the two items in each pair is correlated with accuracy of each feature set. All distance values were z-score normalized by language in order to prevent one dimension from dominating the distance metric. A positive correlation would indicate that more phonetic distance results in more accuracy, and that more similar comparisons have lower accuracy.

The results are given in Figure 4.2. For formants, moments, and MFCCs, there is a correlation between phonetic distance and accuracy. This means that the representations are picking up on meaningful phonetic differences between the sounds - it is easier to discriminate between two segments that are phonetically farther apart. However, this also means that the measurement may not be sensitive enough to capture differences between more phonetically similar sounds. In addition, for these three measures, nasals have a distinctly less strong correlation than any of the other categories. Nasals also have a much smaller range of phonetic distances as measured by MFCC distance. This suggests that nasals are uniquely difficult to characterize acoustically because any acoustic differences between them within the segment are very subtle.

The neural net representation has a slightly different profile. Although there are still generally correlations between phonetic distance and accuracy, the baseline for all of these comparisons are much higher. Here, accuracy is high, even when phonetic distance is low. This is a desirable trait because often very subtle acoustic differences are of interest in phonetic analysis, and suggests that the model is able to pick up on subtle differences that may be of interest to the model.

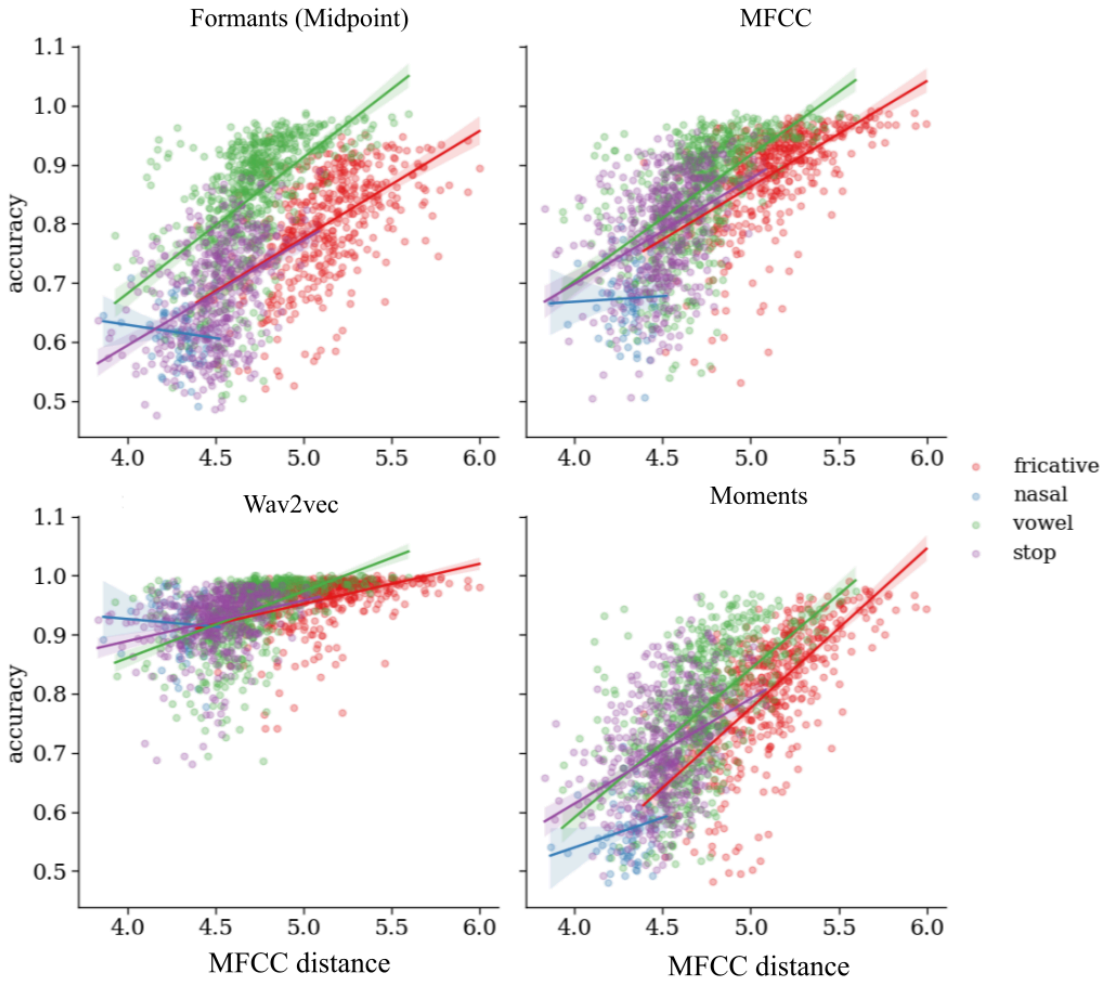

Figure 4.2: Correlation between average within-phoneme MFCC distance and accuracy

4.3.3 Behavior across common phonetic categories

Another question is how these parameterizations behave in the context of specific phonetic comparisons. This analysis took common contrasts for two phonetic categories that have the most well established acoustic measurement methodology: fricatives and vowels, and inspected how accuracy varied across different phonetic features.

Figure 4.3 compares spectral moments and wav2vec2 for fricatives. Moments (left) do particularly well at distinguishing between non-coronal and coronal fricatives. However, in terms of distinguishing within coronal fricatives, or distinguishing between/within non-coronal fricatives, moments do not perform as well. In contrast, wav2vec2 (right), while still performing well on coronal/non-coronal comparisons, also has higher accuracies for comparing between back fricatives and within coronal fricatives. However wav2vec2 also has relatively low accuracies in distinguishing between some fricative pairs that also seem more difficult with spectral moments, suggesting that these sounds are very similar to each other and are generally

hard to distinguish between. Even in these cases, baseline accuracy for wav2vec2 is higher than spectral moments, indicating that there is some advantage to the more complex representation.

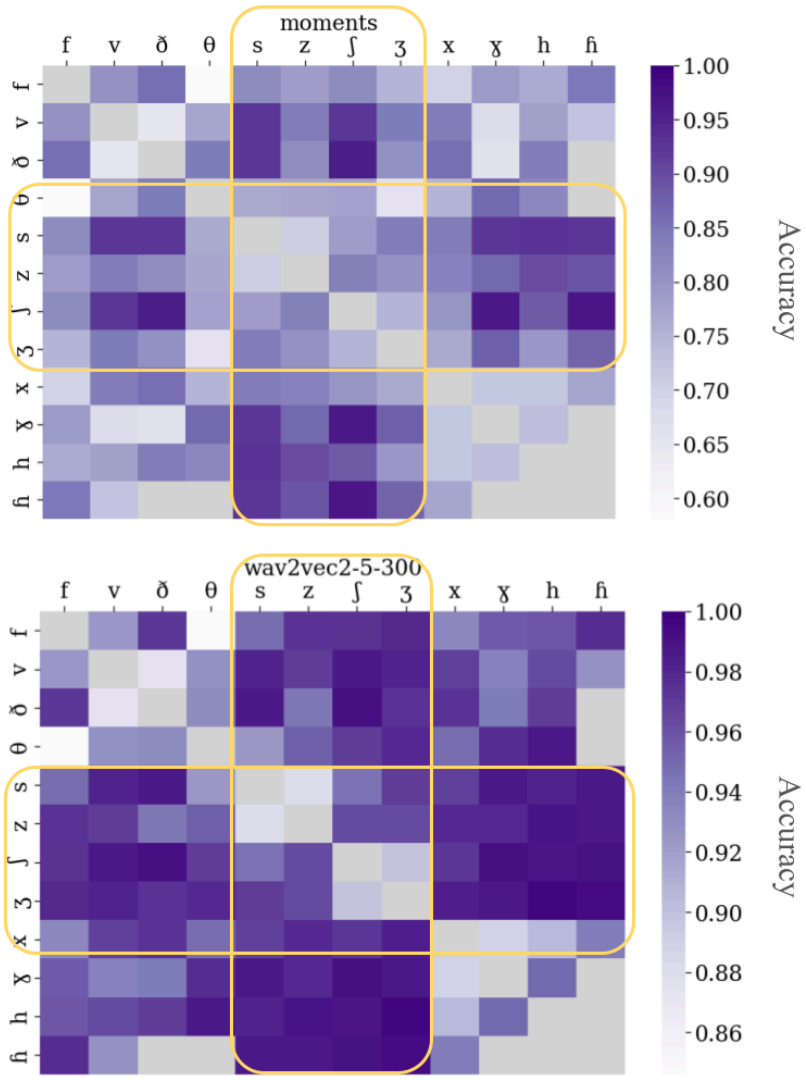

Figure 4.3: Accuracy of wav2vec2 (top) vs moments (bottom) on common fricative categories. Colorbar reports accuracy, and the highlighted box indicates coronal fricatives

Figure 4.4 compares the differences between midpoint formants and wav2vec2 for vowels. For formants and wav2vec2, there is a general lower classifications once again between more similar categories, in particular within front and back vowels. While wav2vec2 shows an overall improvement over the overall accuracies, these relative patterns appear largely the same. This highlights again the fact that conventional acoustic features pick up on important acoustic cues,

and that sounds that are difficult to distinguish between for traditional features are simply difficult to distinguish between in general.

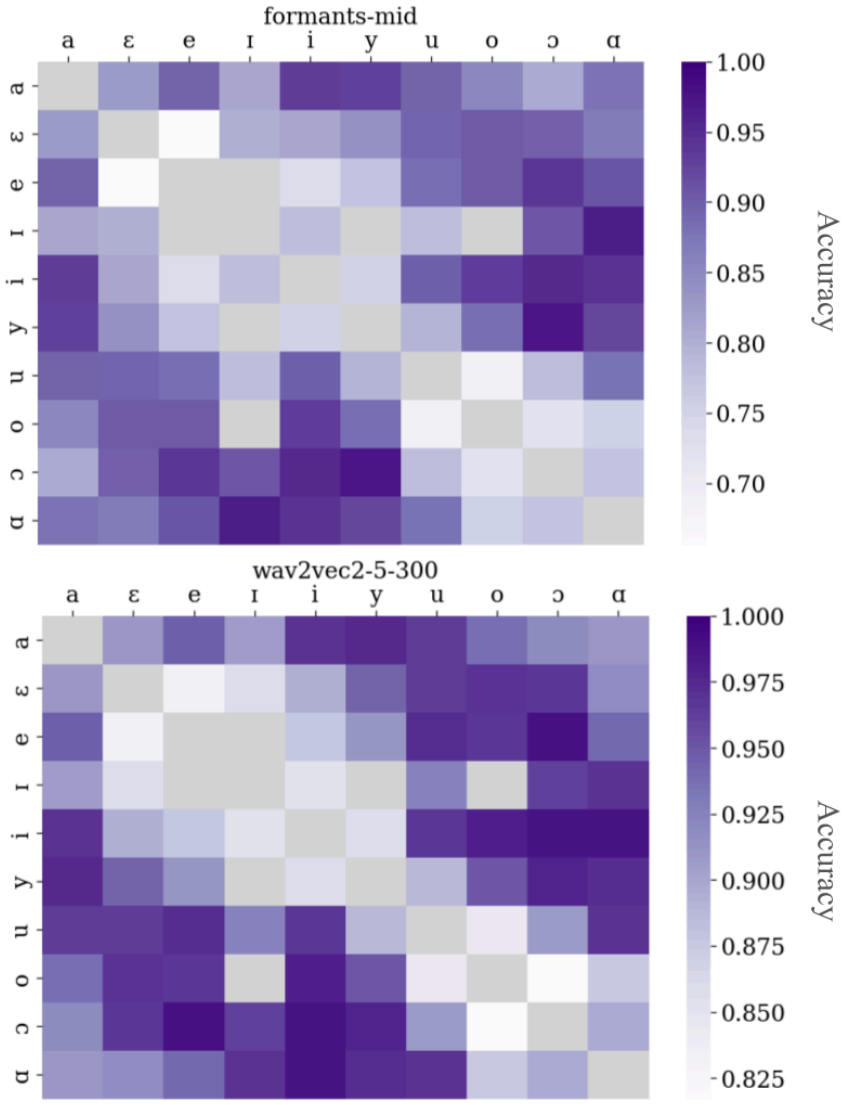

Figure 4.4: Accuracy of wav2vec2 vs formants on common vowel categories

4.4 Conclusion

This study presents a design that can be used to quantitatively compare the relative performance of different acoustic measures on a multilingual dataset applied to four common and acoustically informative phonetic categories. Performance on the task coincides with trends and intuitions in the field in terms of acoustic parameterizations.

Vowels are well-captured by formants, which is in line with previous work, and only sees a small improvement in the use of more maximal features. Fricatives, which have been the subject of recent methodological discussion and are conventionally measured by spectral moments, see significant improvements in classification with less traditional approaches, which suggests that the current accepted method is not as sensitive to the necessary acoustic differences. Nasals and stops are more difficult to distinguish between then other phonetic contrasts, in keeping with the general intuition in the field about these sounds. However, given sufficient complexity in the acoustic representation, it is possible to achieve high classification accuracy on these categories.

In particular, the only approach studied here that produced high results across all phonetic categories under study is the neural nets approach. Most strikingly, this approach was able to differentiate equally well and with high accuracy in pairwise comparisons across all phonetic categories.

Future directions for this work would be to extend this design to other phonetic categories or phenomena of interest. Less common phonetic categories may not be represented across all languages, resulting in less data and therefore less good estimates of how well a measure captures phonetic information in the data. Extending this approach to apply to datasets pertaining to specific phenomena of interest (where there may be less categories available, or less data) is also a worthwhile further extension of this analysis, and would allow for a more complete understanding of the overall phonetic environment.

Another further direction of interest is language sensitivity. Because different models are being trained for each pairwise comparison, it is not clear how each language might be using the same phonetic measures in a different way to capture the contrast. While this does not affect the overall question of feature performance, this extension would afford an opportunity to further understand how phonetic contrasts might be realized across languages.

The benchmark design presented here can be used to systematically compare and evaluate the performance of any type of acoustic measurement across a large number of phonetic contrasts. This metric can give objective insight into how well a measure captures acoustic information, and can be included with other strategies to aid in acoustic measurement selection in the field.

Chapter 5: *Towards the future of acoustic representations*

5.1 Introduction

The previous chapters of this dissertation have focused on understanding the current state of phonetic representations in the field. Chapter 1 identified the common numerical approaches to acoustic phonetic representations and outlined the vast number of approaches that have and can be taken to parameterize speech in linguistics. Chapter 2 investigated trends in the usage of phonetic measures, noting that measurements appear to be dominated by a small number of popularized methods, and that the kinds of contrast that are studied may be partially driven by the kinds of methods that are available. This chapter also presented evidence that indicated that there is lopsided confidence in the methods associated with different phonetic categories, in particular a lower confidence in consonant measures relative to vowels.

Chapters 3 and 4 explored current approaches to phonetic representations in a quantitative way. Chapter 3 found that acoustic context is not necessary for discrimination within vowels and fricatives, but is beneficial for stops and nasals. Chapter 4 presented a methodology for developing a benchmark that quantitatively compared the performance of acoustic measures across several phonetic contrasts. The results indicated very different profiles in how well conventional acoustic feature representations discriminate between phonemic contrasts. Of five representations that were compared, a neural net-based representation outperforms any traditional phonetic representation, and is the only representation that is able to perform similarly well across all phonetic contrasts of interest, particularly showing high performance on nasals and stops.

One of the key findings of this dissertation is that further development of acoustic measurement methodology is a significant area of future research. However, there is much

guidance to be found in the existing landscape of acoustic phonetic analysis in terms of what to look for in a good phonetic measurement. The following section synthesizes three properties that emerge as important to the adoption of a good phonetic representation, and should be centered in any discussion of new and alternative acoustic phonetic measurements.

5.2 A framework for evaluating acoustic measures

Based on the qualitative and quantitative analysis of the current state and trends of the field in this dissertation, the following three properties appear to be essential to a good phonetic feature:

1. **Performance:** It captures relevant variation and minimizes irrelevant variation.
2. **Interpretability:** It can be interpreted in a meaningful way for researchers, for example by being related to articulatory, perceptual, or acoustic theory.
3. **Accessibility:** It can be reasonably implemented given the tools available. A way to communicate the differences between phonetic groups of interest is available.

5.2.1 Performance

Performance for an acoustic measure may vary depending on context- the types of variation that are relevant and irrelevant may be partially dependent on the specific research question. A common formulation of this would be**:** The ideal correlate would be sensitive to perceptible, linguistically meaningful variation and insensitive to variation based on phonetic context, languages, speakers, and methodology.

The quantitative evaluation method described in Chapter 4, which was used to evaluate the performance of acoustic measures, gives a methodology by which to quantify performance into a single informative metric. Quantifying performance provides a clear, transparent method for choosing which acoustic features to use in an analysis. In addition, because of how the dataset used for that study was collected, results are by nature robust to variation from sources such as recording device and environment, phonetic context, and speaker. So, therefore, if a measure performs well under these conditions, it can be expected to perform well under a variety of conditions.

This metric is a strong initial approach to quantifying performance, but there is also room to further refine this metric in future work. While the phonetic contrasts evaluated in that study represent a relatively broad set of phonetic contrasts and languages, there are many phonetic phenomena of interest and comparisons that are not directly represented in the study design so far, including subphonemic or less common phonetic category representations. In addition, while a relatively small subset of the total possible dataset was used in the analysis, expanding this to a larger number of languages, but smaller number of tokens would improve typological breadth. This metric also requires significant computational power to calculate, given the many

comparisons and models involved. This may present another barrier to using this tool. Further development of this performance metric with a focus on a specific, balanced range of contrasts with minimal computational overhead will continue to refine this metric for future applications and easier use.

5.2.2 Interpretability

Interpretability, while harder to quantify, is another property that is key for scientific utility of an acoustic measure. Representations that don't additionally give insight into properties of interest of the sound have limited utility in phonetic research. The most desirable form of interpretability is to be able to relate the numbers in the representation to specific phonetic properties. For example, formants can be related to tongue position for many vowel categories. However, this type of interpretability is costly to develop and is only valid for a narrow slice of the acoustic space. In addition, for some acoustic properties, the articulatory relationship may not be as straightforward.

Another form of interpretability is a correlation between the representation and a more well-understood measure, such as the correlation between PCA measures and F1/F2 in vowels (Leinonen 2008), or between synthetic stimuli with known values and the acoustic representation (Choi and Yeo 2022). Perception measures can also be used, such as the correlation between perceived lenition and the acoustic measurements to validate their proposed feature (Bolyanatz and Brogan 2021), and between /s/ features and acoustic parameters (Gradoville et al. 2022). The benchmarks can also be articulatory, such as the correlation between Center of Gravity and electropalatography (Tabain 2001).

While correlational approaches to interpretability are useful, they are ultimately less satisfying in terms of grounding our interpretation of the representation scientifically. Interpretability is difficult to quantify - it's more of a qualitative assessment of how well a phonetic measure matches with the expectation based on the current understanding of the theory, and very few acoustic measures today have satisfying levels of phonetic interpretability. When thinking about exploring future phonetic features, one approach is to design representations with interpretability as a primary goal, rather than as an afterthought. This would increase the likelihood of developing effective acoustic measurements that are transparent and interpretable.

A final form of interpretability might relate the numbers back to the original acoustic signals - for example the coefficients of a DCT can be reconstructed into an approximation of the original signal. This approach does not require significant theoretical development and has the advantage of being broadly applicable across different phonetic categories. In addition, linguists are already accustomed to interpreting speech representations such as spectrograms and generating meaningful insights from these sources.

With this in mind, I will define an interpretable feature as something that either it can be directly related to other phonetic properties or related to a familiar acoustic representation of speech such as the waveform, spectrum, or spectrogram that can be interpreted.

5.2.3 Accessibility

Accessibility means generally the ease with which it is possible to use a particular parameterization. This includes the process of deriving the features - does the method require paid software? How transparent is the analysis process? Are there references in the field that describe the methodology in detail? What kind of technical knowledge is required in order to extract and analyze the features? In particular, the technical requirement of features is something to consider, and any feature extraction method needs to be accessible. Issues relating to technical knowledge (e.g. programming skills) necessary to extract the representation can be bridged by developing software tools or scripts that extract the representations of interest, and documenting the use of these systems. However, other issues are less easily surmounted, such as the amount of data necessary to use to the tool, or high computational resource requirements. The approach taken here for accessibility is based on the threshold reached by commonly used methods in the field - if it is possible to develop a script that can take a typical TextGrid/ WAV combination and generate the representations for future analysis.

This three-part definition of a good acoustic measurement can be used to critically evaluate the usage of current acoustic measures in the field. Formants are interpretable, accessible and perform well. In addition, they have a robust history of being able to capture differences in the field. Spectral moments are not sensitive enough or interpretable, but they are accessible. Voice quality measurements (H1-H2) have high performance, are partially interpretable, and are accessible. From this perspective, accessibility appears to be a driving factor, followed by interpretability and sensitivity - if it is impossible to implement a measure, then it won't be used enough to establish the other two properties.

The first property is related to the quantitative ability of the representation to capture phonetic contrasts, while the latter two are more related to the qualities that are necessary in order to use the method for scientific inquiry in phonetics. Within this framework, methods that may be excellent in terms of quantitatively discriminating between contrasts will still not be used if they can't be interpreted. In contrast, a method that is interpretable/accessible may be used even if it is less able to quantitatively capture the patterns observed in the data. In order to holistically evaluate phonetic representations, it is necessary to consider each of these three properties, rather than a single aspect of the representation.

Chapters 3 and 4 considered neural networks as an approach to phonetic representation. In terms of the framework presented here, this type of representation scored the highest when it comes to performance, but it is not a particularly accessible measure to the broader scientific community. In addition, these methods are not interpretable, given the current state of interpretability of this model. So it makes sense why, despite being popular in many adjacent fields of acoustics, neural nets have not caught on in phonetics.

5.3 The future of acoustic representations: interpretable machine learning

The current set of acoustic measures popular in phonetics, while useful for specific high-impact phenomena of interest, represent only a small slice of the vast amount of phenomena of scientific interest to the field. As a result, it seems that exploring additional methods of acoustic description is necessary for advancing phonetic research. In particular, for some categories, incorporating acoustic context and more complex patterns seems necessary to adequately characterize their acoustic properties. Machine learning approaches such as neural networks are well positioned to serve that purpose, but the framework of evaluating acoustic representations described in 5.2 emphasizes that interpretability and accessibility are perhaps more important than quantitative performance in terms of having high quality acoustic representations, and must be considered from the outset in designing approaches.

In particular, a challenge for machine learning approaches to acoustic representations is interpretability. Quantifying interpretability has also been taken in the machine learning context- by feeding controlled stimuli to the algorithm and study the impact of the output on the organization of the representations (Choi and Yeo 2022), occluding parts of the signal (Ferragne et al. 2019), correlating to other activations (Zhou et al. 2023) or correlating to acoustic features (Riera et al. 2019). This indirect investigation gives some insight, but given the high degree of complexity in the representation, cannot be expected to capture the total range of possible variation in the data.

While it is still possible to account for interpretability after the fact to some degree, other attempts to establish interpretability in the field have not yet proven satisfactory enough to be taken as a definitive interpretation of the representation. For example, despite having significant research into the post-hoc interpretability of spectral moments, the interpretation of each of the components of the representation are not widely understood. In contrast, other popular parameterizations such as formants can be related back to an easily visualized relationship between the signal and the number. From these kinds of representations, jumping to the much less interpretable representation of neural nets means accepting a large reduction in interpretability.

Another way to approach interpretability is to investigate machine learning approaches where it is possible to relate the features back to significant elements of the original speech signal. For example, some deep learning approaches use whole-spectrogram input, which is a natural evolution from the template matching approach from previous decades (discussed in Chapter 1). In this case, it is possible to treat the spectrogram as essentially an image and therefore leverage models and techniques developed in computer vision, which has been used with some success in clustering (Hajarolasvadi and Demirel 2019) and deep learning approaches (Gong et al. 2022). Some work has also focused on deriving local features from spectrograms (e.g. Denis et al. 2013, Schutte 2009). The following section outlines one such type of model with potential to contribute to the advancement of acoustic representations in phonetics.

5.3.1 The Audio Spectrogram Transformer

The Audio Spectrogram Transformer (AST) (Gong et al. 2021) is a transformer architecture that has been developed recently for use in audio classification. A characteristic of this architecture is that it is purely attention based, with no convolutional component, which is a departure from other audio models. Some key properties of this model that make it of interest for acoustic phonetic representations are:

Input: The (mel) spectrogram. The feature extraction component of this model computes the mel spectrogram of the data. The mel transformation of the spectrogram is a common perceptual contribution used frequently in speech, and is based on filterbanks meant to mimic the relative importance of different frequencies for human auditory perception. This spectrogram is then used as the input to the transformer, and gives a visually interpretable input data type. This particular model used 12 mel bins for the spectrogram in the frequency domain, but a more fine-grained approach may be more appropriate to the phonetic task of classification.

Training data: The base form of this model is trained on the AudioSet dataset (Gemmeke et al. 2017). This dataset is based on a task for classifying 10-second clips of audio in terms of broad acoustic category (e.g. chirp, bell, whistle, sigh, speech). This is not trained specifically on speech recognition type tasks, and the timescale of the audio that is being used is much longer than typically for speech. The specific model reported here was fine-tuned on the speech commands v2 dataset (Warden 2018). This dataset has a limited vocabulary of English speech commands, which is closer to the target type of data that is common in phonetics, but is still significantly limited relative to the types of multilingual data that is frequently used in other models applied to speech.

Architecture: The key innovation of this model over other similar acoustic neural net models is the use of attention rather than convolution (or a combination of both). Popular transformer models such as wav2vec2 (Baevski et al. 2020) have a convolutional component followed by a transformer component. The introduction of convolutions prior to the attention component in that case means that it is more difficult to relate the weighted attention back to the original signal. The AST architecture has 12 transformer layers, with 12 attention heads in each layer - a total of 144 attention heads. Another feature of the attention-only mechanism is that it is possible to mask different attention heads for attention layers. The two points above taken together mean that the weighted attention maps can be mapped back onto the original spectrogram relatively easily. While this means that individual feature values cannot necessarily be related to components of the spectrogram, it is possible to get an overall visual impression of which parts of the spectrogram are influencing the representation. In addition, the complexity of the model can be reduced by masking attention heads, resulting in a much smaller model. Thus, it may be possible to reduce the complexity of the model by pruning attention heads, therefore resulting in a simpler but still effective representation.

5.3.2 Evaluating the AST as a phonetic representation

Given, based on the description above, that the Audio Spectrogram Transformer is a candidate for useful, interpretable phonetic representation, it is possible to evaluate the AST representation within the framework established in this chapter for determining a high-quality acoustic feature.

Performance: Figure 5.1 gives the results for the performance metric described in Chapter 4 for all four window sizes and phonetic categories in the dataset from Chapter 3. For fricatives and vowels, there are relatively high classification rates across the first three window sizes in the data. Both categories reach a maximum accuracy of around 90%. Nasals and stops see an increase in classification rates across the first three window sizes, with a maximum accuracy around 150-percent window size.

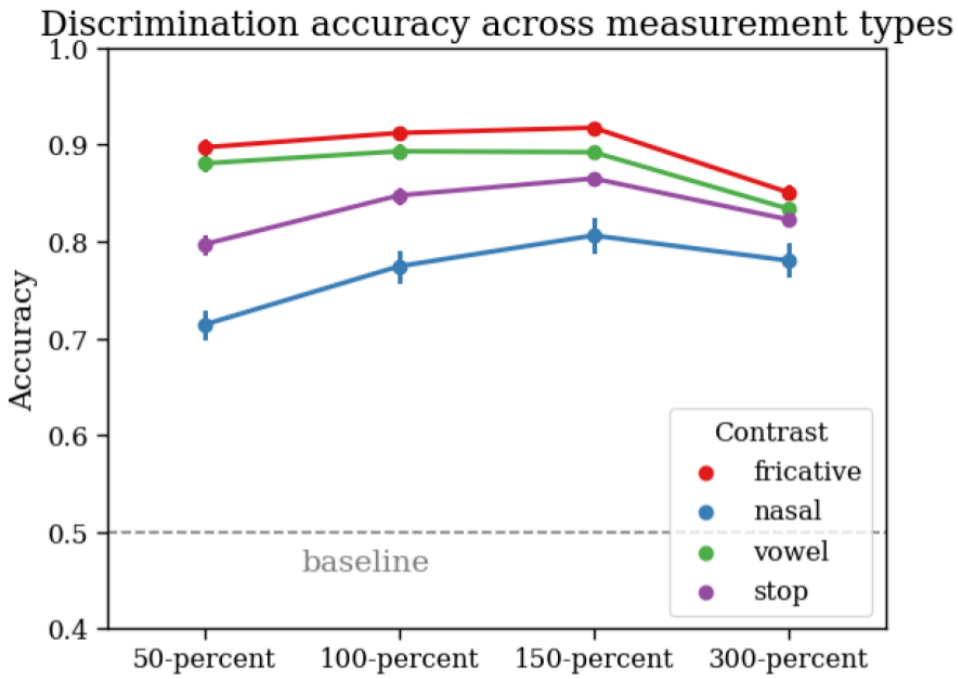

Figure 5.1: Aggregated pairwise discrimination accuracies for all phonetic categories with different window sizes

Unlike the previous neural net architecture used, there is no advantage of using the largest window size. This may be because the proportion of the signal that is relevant for classification (the phonetic contrast) is a smaller part of the data when using fixed-width spectrogram input as is leveraged in this study. Rather in this case, the context with the highest accuracy across all phonetic contrasts is 150%, which includes the entire segment and a small amount of information on either side.

Overall, the results for this model fall somewhere between MFCCs and wav2vec2 when compared to the results from Chapter 4. AST-based representations are not the most sensitive,

but given that the model has much less direct training on speech, and no multilingual input, it seems likely that the performance of the representations will increase if the model is fine-tuned on multilingual phonetic input. In addition, this model is able to achieve good accuracy with significantly less model complexity - just six transformer layers, rather than 12 total layers as was the case for the previous neural model in Chapters 3 and 4.

Interpretability: The main potential advantage of this approach to neural nets would be a more interpretable approach to phonetic differences. The main part of this interpretability would be in visualizing how attention maps highlight which parts of the spectrogram are used in the analysis. While a full treatment of interpretability of this model is outside the scope of this dissertation, this section will briefly explore the directions possible here.

The structure of the model means that it is possible to visualize the attention of the model as mapped onto the input data. These attention maps can be visualized to show which part of the signal is having the most influence over the representation at that point of the model. The figures below compare the original spectrogram to the attention maps for each attention head in layer 6 of the model. Figure 5.2 gives the attention maps for an utterance of /f/ and Figure 5.3 gives the same values for an utterance of /s/.

Each attention head in both figures appears to be primarily concerned with a very specific part of the signal, which suggests that it may be possible to pinpoint the influence of relatively subtle parts of the signal to speech. Also, because of attention, there are no locality restrictions on context. The attention maps between the two different fricatives are qualitatively very similar to each other, but there is a difference in the values associated with the attention map for head 5- this has the highest activation in lower frequencies of the frication in /f/, and highest activation in high frequency frication in /s/. This is also an area where one might expect to see differences when qualitatively analyzing the spectrograms as a trained phonetician. This provides a qualitative alignment of which parts of the spectrogram are driving any representational differences between the two sounds.

That this kind of visualization of attention maps provides a level of intuitive interpretability that may make these models a viable source of acoustic representations,

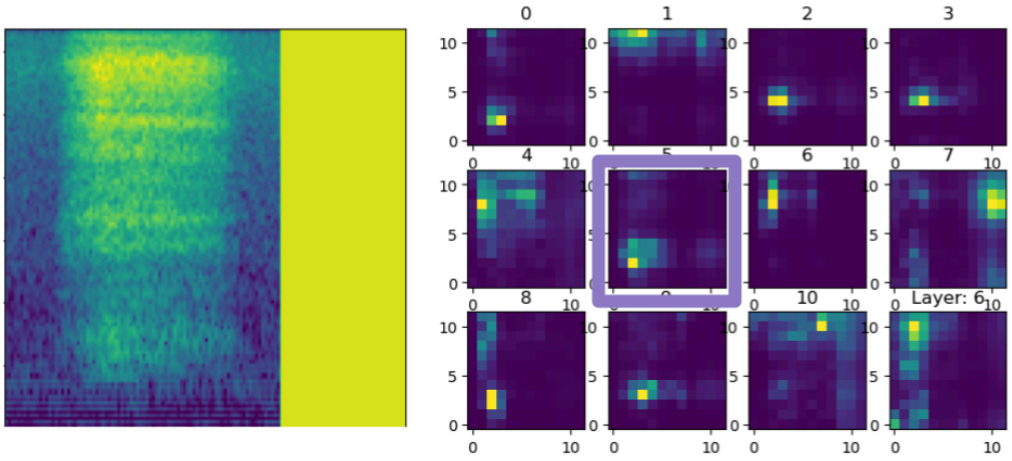

Figure 5.2: Original spectrogram (left) and attention maps for layer 6 (right) for /f/. Highlighted box is the attention map for head 5 in layer 6.

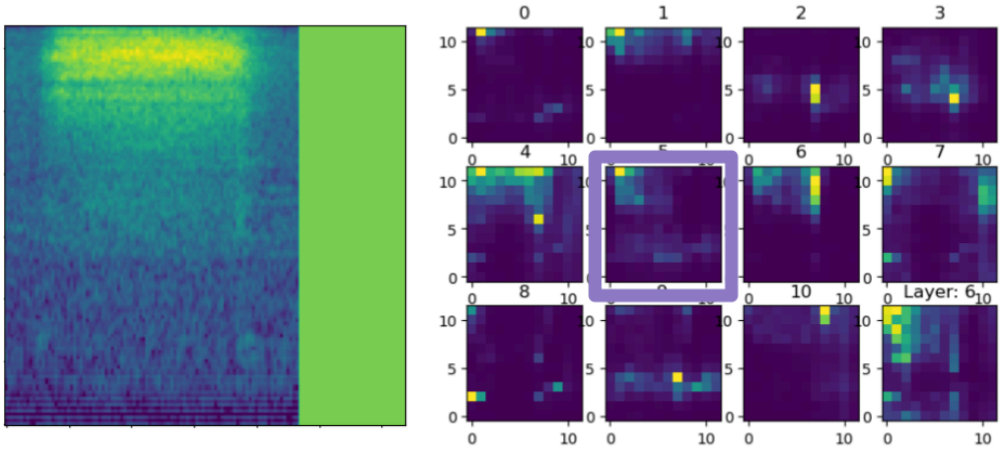

Figure 5.3: Original spectrogram (left) and attention maps for layer 6 (right) for /s/. Highlighted box is the attention map for head 5 in layer 6.

especially since interpreting spectrograms has been a long standing tradition in the field. In addition, these approaches are still compatible with the many indirect probing techniques typically used for understanding the inner workings of neural network-based representations. In addition, it is possible to entirely mask certain attention heads and therefore their input into a given layer of the representation. This affords a degree of fine-grained control over the complexity of the representation that is a potential avenue for fruitful future investigation.

Accessibility: While working with neural networks can be an intimidating prospect, there are a number of packages and tools designed to disseminate pre-trained models in a way such that a script + time-aligned audio paradigm can be reasonably developed to meet that standard for accessibility. If the model needs to be further trained or fine-tuned, the requirements for data and computational power do go up. However, the fine-tuned model can then be shared in the same way as a pre-trained model, meaning that technical burden does not need to be assumed by the end user. Thus, while this method doesn't currently match the criteria of accessibility, it is possible to develop the tools to support this method such that it is more generally accessible as an acoustic measurement strategy.

5.4 Conclusion

The overall goal of this dissertation was to understand the landscape of phonetic features and propose directions for productive future development in this area. First, this dissertation outlined the common categories of acoustic measures conventionally used in the field (Chapter 1) and compared this scope of possibility to actual usage (Chapter 2). This analysis revealed a pattern of interest and potential concern to the field: phonetic categories with established quantitative methodologies are more frequently studied than those without. In particular, interesting but less frequent consonant contrasts see very little quantitative research, and qualitative methods still are very common in less commonly studied phonetic contrasts. These findings highlight the need for further development of robust acoustic methodology in the field.

This dissertation then turned to ways of quantifying performance of acoustic measurements in the field. First, this necessitated the development of a multilingual benchmark dataset that can be used as a testing ground to compare different acoustic representations. Using this benchmark dataset, this dissertation investigated two questions. First, Chapter 3 quantified the role of contextual information in a highly complex acoustic representation, and the methodology was broadened into a general benchmark design that can be applied to any set of acoustic representations to quantify performance. The results of these chapters highlighted the importance of considering how the acoustic properties of different phonetic categories may influence performance of acoustic methodologies, and suggest that a representation that contains at least some local contextual information may be necessary in order to achieve satisfactory performance in some acoustic categories.

The final section of this dissertation (Chapter 5) distills these findings into a concrete set of recommendations for future design and evaluation of acoustic representations in the field: performance, interpretability, and accessibility. In particular, this section highlighted the benefit of explicitly considering these properties when designing new acoustic representations in phonetics. The current landscape of phonetic parameters makes it clear that both interpretability and quantitative performance need to be balanced in exploring phonetic representations. The architecture explored here - the Audio Spectrogram Transformer- appears to balance both

interpretability and sensitivity, and has significant potential to improve in both dimensions through pruning of attention heads and fine-tuning on a more specifically tuned task.

The AST model presented here is one of many transformer architectures that could be leveraged to improve acoustic measurements in the field. In particular, the visual interpretability of the attention heads demonstrate the potential for constructing features that can be to some extent related back to the original acoustic information. More generally, designing features with interpretability in mind can help guide development in the field to yield useful and usable acoustic measures. As computational methods continue to evolve, a deep understanding of the priorities for designing new acoustic features will help to focus future research on areas of high potential benefit to the field.

Bibliography

Abdi, H., & Williams, L. J. (2010). Principal component analysis. *WIREs Computational Statistics*, *2*, 433–459.

Adank, P., Smits, R., & Van Hout, R. (2004). A comparison of vowel normalization procedures for language variation research. *The Journal of the Acoustical Society of America*, *116*, 3099–3107.

Ahn, E., & Chodroff, E. (2022). Voxcommunis: A corpus for cross-linguistic phonetic analysis. *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 5286–5294.

Ainsworth, W. A., & Paliwal, K. K. (1984). Correlation between the production and perception of the English glides /w, r, l, j/. *Journal of Phonetics*, *12*, 237–243.

Ardila, R., Branson, M., Davis, K., Kohler, M., Meyer, J., Henretty, M., Weber, G. (2020). Common Voice: A Massively-Multilingual Speech Corpus. *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 4218–4222.

Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. *Advances in Neural Information Processing Systems*, *33*, 12449–12460.

Barreda, S. (2021). Fast Track: Fast (nearly) automatic formant-tracking using Praat. *Linguistics Vanguard*, *7*, 20200051.

Bartelds, M., De Vries, W., Sanal, F., Richter, C., Liberman, M., & Wieling, M. (2022). Neural representations for modeling variation in speech. *Journal of Phonetics*, *92*, 101137.

Bartelds, M., & Wieling, M. (2022). Quantifying Language Variation Acoustically with Few Resources. *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 3735–3741.

Berkson, K. (2012). Acoustic correlates of breathy voice in Marathi sonorants. *The Journal of the Acoustical Society of America*, *132*, 2001–2001.

Best, C. T., & Strange, W. (1992). Effects of phonological and phonetic factors on cross-language perception of approximants. *Journal of Phonetics*, *20*, 305–330.

Bhatt, S., Jain, A., & Dev, A. (2020). Acoustic modeling in speech recognition: A systematic review. *International Journal of Advanced Computer Science and Applications*, *11*.

Blackwood Ximenes, A., Shaw, J. A., & Carignan, C. (2017). A comparison of acoustic and articulatory methods for analyzing vowel differences across dialects: Data from American and Australian English. *The Journal of the Acoustical Society of America*, *142*, 363–377.

Boersma, P. (1993). Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. *Proceedings of the Institute of Phonetic Sciences*, *17*.

Boersma, P., & Weenink, D. (2011). *Praat: Doing phonetics by computer [Computer program]*. Retrieved from http://www.praat.org/

Bolyanatz, M., & Brogan, F. D. (2021). Acoustic differences between Chilean and Salvadoran Spanish /s/. *The Journal of the Acoustical Society of America*, *150*, 2446–2460.

Bough, I. D., Heuer, R. J., Sataloff, R. T., Hills, J. R., & Cater, J. R. (1996). Intrasubject variability of objective voice measures. *Journal of Voice*, *10*, 166–174.
Brotherton, C., & Block, A. (2020). Soft d in Danish: Acoustic characteristics and issues in transcription. *Proceedings of the Linguistic Society of America*, *5*, 792.
Carter, P. (2004). Extrinsic phonetic interpretation: Spectral variation in English liquids. In L. John, O. Richard, & Rosalind Temple (Eds.), *Phonetic Interpretation: Papers in Laboratory Phonology VI*. Cambridge University Press.
Cassidy, S., & Harrington, J. (1995). The Place of Articulation Distinction in Voiced Oral Stops: Evidence from Burst Spectra and Formant Transitions. *Phonetica*, *52*, 263–284.
Cheng, R., & Jongman, A. (2019). Acoustic analysis of nasal and lateral consonants: The merger in Eastern Min. *The Journal of the Acoustical Society of America*, *145*, 1828–1828.
Cho, C. J., Wu, P., Mohamed, A., & Anumanchipalli, G. K. (2023). Evidence of Vocal Tract Articulation in Self-Supervised Learning of Speech. *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. Rhodes Island, Greece: IEEE.
Cho, T., Jun, S.-A., & Ladefoged, P. (2002). Acoustic and aerodynamic correlates of Korean stops and fricatives. *Journal of Phonetics*, *30*, 193–228.
Choi, K., & Yeo, E. J. (2022). *Opening the Black Box of wav2vec Feature Encoder*. *Arxiv*.
Chu, W., & Alwan, A. (2012). SAFE: A Statistical Approach to F0 Estimation Under Clean and Noisy Conditions. *IEEE Transactions on Audio, Speech, and Language Processing*, *20*, 933–944.
Conneau, A., Baevski, A., Collobert, R., Mohamed, A., & Auli, M. (2021). Unsupervised Cross-Lingual Representation Learning for Speech Recognition. *Interspeech 2021*.
Crosby, D., & Dalola, A. (2021). Phonetic variation in the Korean liquid phoneme. *Proceedings of the Linguistic Society of America*, *6*, 701.
Davidson, L. (2021). Effects of word position and flanking vowel on the implementation of glottal stop: Evidence from Hawaiian. *Journal of Phonetics*, *88*, 101075.
de Seyssel, M., Lavechin, M., Adi, Y., Dupoux, E., & Wisniewski, G. (2022). Probing phoneme, language and speaker information in unsupervised speech representations. *Interspeech 2022*, 1402–1406.
Delattre, P., & Freeman, D. C. (1968). A Dialect Study of American R's by X-ray Motion Picture. *Linguistics*, *6*.
Denis, B., Côté, J., & Laprise, R. (2002). Spectral decomposition of two-dimensional atmospheric fields on limited-area domains using the discrete cosine transform (DCT). *Monthly Weather Review*, *130*, 1812–1829.
Dunn, H. K. (1950). The Calculation of Vowel Resonances, and an Electrical Vocal Tract. *The Journal of the Acoustical Society of America*, *22*, 740–753.
English, P. C., Kelleher, J., & Carson-Berndsen, J. (2022). Domain-informed probing of wav2vec 2.0 embeddings for phonetic features. *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, 83–91.
Esposito, C. M. (2010). The effects of linguistic experience on the perception of phonation. *Journal of Phonetics*, *38*, 306–316.
Fan, Z., Li, M., Zhou, S., & Xu, B. (2021, January 14). *Exploring wav2vec 2.0 on speaker verification and language identification*. arXiv.

Fant, G. (1971). *Acoustic theory of speech production: With calculations based on X-ray studies of Russian articulations* (Vol. 2). Walter de Gruyter.
Ferragne, E., Gendrot, C., & Pellegrini, T. (2019). Towards phonetic interpretability in deep learning applied to voice comparison. *ICPhS*, ISBN-978.
Fischer-Jorgensen, E. (1967). Phonetic analysis of breathy (murmured) vowels in Gujarati. *Annual Report of the Institute of Phonetics University of Copenhagen*.
Fischer-Jorgensen, E. (1954). Acoustic Analysis of Stop Consonants. *Le Maître Phonétique*, *32*, 42–59.
Forrest, K., Weismer, G., Milenkovic, P., & Dougall, R. N. (1988). Statistical analysis of word-initial voiceless obstruents: Preliminary data. *The Journal of the Acoustical Society of America*, *84*, 115–123.
Frid, A., & Lavner, Y. (2010). Acoustic-phonetic analysis of fricatives for classification using SVM based algorithm. *2010 IEEE 26-Th Convention of Electrical and Electronics Engineers in Israel*.
Fujimura, O. (1962). Analysis of Nasal Consonants. *The Journal of the Acoustical Society of America*, *34*, 1865–1875.
Fulop, S.A., Kari, E., & Ladefoged, P. (1998). An Acoustic Study of the Tongue Root Contrast in Degema Vowels. *Phonetica*, *55*, 80–98.
Fulop, Sean A., Ladefoged, P., Liu, F., & Vossen, R. (2003). Yeyi Clicks: Acoustic Description and Analysis. *Phonetica*, *60*, 231–260.
Garellek, M. (2022). Theoretical achievements of phonetics in the 21st century: Phonetics of voice quality. *Journal of Phonetics*, *94*, 101155.
Garellek, M., Keating, P., Esposito, C. M., & Kreiman, J. (2013). Voice quality and tone identification in White Hmong. *The Journal of the Acoustical Society of America*, *133*, 1078–1089.
Gemmeke, J. F., Ellis, D. P. W., Freedman, D., Jansen, A., Lawrence, W., Moore, R. C., Ritter, M. (2017). Audio Set: An ontology and human-labeled dataset for audio events. *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 776–780. New Orleans, LA: IEEE.
Gendrot, C., Ferragne, E., & Pellegrini, T. (2019). Deep learning and voice comparison: Phonetically-motivated vs. Automatically-learned features. *ICPhS*.
Gerfen, C., & Baker, K. (2005). The production and perception of laryngealized vowels in Coatzospan Mixtec. *Journal of Phonetics*, *33*, 311–334.
Gerosa, M., Lee, S., Giuliani, D., & Narayanan, S. (2006). Analyzing Children's Speech: An Acoustic Study of Consonants and Consonant-Vowel Transition. *2006 IEEE International Conference on Acoustics Speed and Signal Processing Proceedings*, *1*, I-393-I–396. Toulouse, France: IEEE.
Ghaemmaghami, S., Deriche, M., & Boashash, B. (1997). Comparative study of different parameters for temporal decomposition based speech coding. *1997 IEEE International Conference on Acoustics, Speech, and Signal Processing*, *3*, 1703–1706. Munich, Germany: IEEE Comput. Soc. Press.
Ghaffarvand Mokari, P., & Mahdinezhad Sardhaei, N. (2020). Predictive power of cepstral coefficients and spectral moments in the classification of Azerbaijani fricatives. *The Journal of the Acoustical Society of America*, *147*, EL228–EL234.

Giles, P. (2019). Using crowd-sourced data to analyse the ongoing merger of [ɕ] and [ʃ] in Luxembourgish. *Proceedings of the International Congress of Phonetic Sciences*. Presented at the ICPHS 2019, Melbourne. Melbourne.
Gobl, C., & Ni Chaside, A. (1999). Perceptual correlates of source parameters in breathy voice. *Proceedings of the International Congress of Phonetic Sciences*. Presented at the ICPHS, San Francisco. San Francisco.
Gold, E., & French, P. (2011). International Practices in Forensic Speaker Comparison. *International Journal of Speech, Language and the Law*, *18*, 293–307.
Goldberg, H. G., & Reddy, R. (1977). Phonetic labeling by template matching. *The Journal of the Acoustical Society of America*, *61*, S69–S70.
Gong, Y., Lai, C.-I., Chung, Y.-A., & Glass, J. (2022). SSAST: Self-Supervised Audio Spectrogram Transformer. *Proceedings of the AAAI Conference on Artificial Intelligence*, *36*, 10699–10709.
Gordon, M., Barthmaier, P., & Sands, K. (2002). A cross-linguistic acoustic study of voiceless fricatives. *Journal of the International Phonetic Association*, *32*, 141–174.
Gradoville, Michael S, Brown, E. K., & File-Muriel, R. J. (2022). The phonetics of sociophonetics: Validating acoustic approaches to Spanish/s. *Journal of Phonetics*, *91*, 101125.
Gradoville, Michael Stephen. (2011). Validity in measurements of fricative voicing: Evidence from Argentine Spanish. *Selected Proceedings of the 5th Conference on Laboratory Approaches to Romance Phonology*. Somerville, MA: Cascadilla Proceedings Project.
Gump, M. H. (2020). *Unsupervised methods for evaluating speech representations* (PhD Thesis). Massachusetts Institute of Technology.
Hajarolasvadi, N., & Demirel, H. (2019). 3D CNN-Based Speech Emotion Recognition Using K-Means Clustering and Spectrograms. *Entropy*, *21*, 479.
Halle, M., Hughes, G. W., & Radley, J.-P. A. (1957). Acoustic Properties of Stop Consonants. *The Journal of the Acoustical Society of America*, *29*, 107–116.
Hargus, S., Levow, G.-A., & Wright, R. (2021). *Acoustic characteristics of Deg Xinag fricatives*. University of Washington Working Papers in Linguistics.
Harrington, J. (2009). Acoustic Phonetics. In W. J. Hardcastle (Ed.), *The handbook of phonetic sciences* (1. publ. paperback, [Nachdr.]). Oxford, UK: Blackwell.
Harrison, P. (2013). *Making Accurate Formant Measurements: An Empirical Investigation of the Influence of the Measurement Tool, Analysis Settings and Speaker on Formant Measurements*. University of York.
Hillenbrand, J. M., & Houde, R. A. (2003). A narrow band pattern-matching model of vowel perception. *The Journal of the Acoustical Society of America*, *113*, 1044–1055.
Hillenbrand, J. M., Houde, R. A., & Gayvert, R. T. (2006). Speech perception based on spectral peaks versus spectral shape. *The Journal of the Acoustical Society of America*, *119*, 4041–4054.
Hirst, D. (2007). A Praat plugin for Momel and INTSINT with improved algorithms for modelling and coding intonation. *16th International Congress of Phonetic Sciences*. Presented at the ICPhS XVI.
Ho, S. Y. B. (2021). *The phonetics and phonology of hong kong english: A study of fricatives* (Westfaelische Wilhelms-Universitaet). Westfaelische Wilhelms-Universitaet, Muenster, Germany.

Hoshen, Y., Weiss, R. J., & Wilson, K. W. (2015). Speech acoustic modeling from raw multichannel waveforms. *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4624–4628. South Brisbane, Queensland, Australia: IEEE.

Hualde, J. I., Beristain, A., Isasa, A. I., & Zhang, J. (2019). Lenition of word-final plosives in Basque. *Proceedings of the International Congress of Phonetic Sciences*. Presented at the ICPHS 2019.

Hughes, G. W., & Halle, M. (1956). Spectral Properties of Fricative Consonants. *The Journal of the Acoustical Society of America*, *28*, 303–310.

Hughes, V., Clermont, F., & Harrison, P. (2020). Correlating Cepstra with Formant Frequencies: Implications for Phonetically-Informed Forensic Voice Comparison. *Interspeech 2020*, 1858–1862. ISCA.

Ingemann, F., & Mermelstein, P. (1975). Speech recognition through spectrogram matching. *The Journal of the Acoustical Society of America*, *57*, 253–255.

Itakura, F. (1975). Minimum prediction residual principle applied to speech recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *23*, 67–72.

Jassem, W. (1965). The Formants of Fricative Consonants. *Language and Speech*, *8*, 1–16.

Jassem, W. (1995). The Acoustic Parameters of Polish Voiceless Fricatives: An Analysis of Variance. *Phonetica*, *52*, 251–258.

Jaworski, S., & Baran, M. (2021). Acoustic Features of Burst Release: A Study of Welsh Plosives. *Roczniki Humanistyczne*, *69*, 89–105.

Johnson, K. (2012). *Acoustic and auditory phonetics* (3. ed., 1. publ). Malden, Mass.: Wiley-Blackwell.

Johnson, K., & Sjerps, M. J. (2021). Speaker Normalization in Speech Perception. In J. S. Pardo, L. C. Nygaard, R. E. Remez, & D. B. Pisoni (Eds.), *The Handbook of Speech Perception* (1st ed., pp. 145–176). Wiley.

Jongman, A., Wayland, R., & Wong, S. (2000). Acoustic characteristics of English fricatives. *The Journal of the Acoustical Society of America*, *108*, 1252–1263.

Kashani, H. B., Sayadiyan, A., & Sheikhzadeh, H. (2017). Vowel detection using a perceptually-enhanced spectrum matching conditioned to phonetic context and speaker identity. *Speech Communication*, *91*, 28–48.

Kent, R. D., & Vorperian, H. K. (2018). Static measurements of vowel formant frequencies and bandwidths: A review. *Journal of Communication Disorders*, *74*, 74–97.

Khan, S. U. D. (2012). The phonetics of contrastive phonation in Gujarati. *Journal of Phonetics*, *40*, 780–795.

Klatt, D. (1982). Prediction of perceived phonetic distance from critical-band spectra: A first step. *ICASSP '82. IEEE International Conference on Acoustics, Speech, and Signal Processing*, *7*, 1278–1281. Paris, France: Institute of Electrical and Electronics Engineers.

Klatt, D. H., & Klatt, L. C. (1990). Analysis, synthesis, and perception of voice quality variations among female and male talkers. *The Journal of the Acoustical Society of America*, *87*, 820–857.

Kochetov, A. (2020). Research methods in articulatory phonetics I: Introduction and studying oral gestures. *Language and Linguistics Compass*, *14*, 1–1.

Kochetov, A., Tabain, M., Sreedevi, N., & Beare, R. (2018). Manner and place differences in Kannada coronal consonants: Articulatory and acoustic results. *The Journal of the Acoustical Society of America*, *144*, 3221–3235.
Koenig, L. L., Shadle, C. H., Preston, J. L., & Mooshammer, C. R. (2013). Toward Improved Spectral Measures of /s/: Results From Adolescents. *Journal of Speech, Language, and Hearing Research*, *56*, 1175–1189.
Koenig, W., Dunn, H. K., & Lacy, L. Y. (1946). The Sound Spectrograph. *The Journal of the Acoustical Society of America*, *18*, 244–244.
Kreiman, J., Gerratt, B. R., & Antoñanzas-Barroso, N. (2007). Measures of the Glottal Source Spectrum. *Journal of Speech, Language, and Hearing Research*, *50*, 595–610.
Kreiman, J., Shue, Y.-L., Chen, G., Iseli, M., Gerratt, B. R., Neubauer, J., & Alwan, A. (2012). Variability in the relationships among voice quality, harmonic amplitudes, open quotient, and glottal area waveform shape in sustained phonation. *The Journal of the Acoustical Society of America*, *132*, 2625–2632.
Kurowski, K., & Blumstein, S. E. (1987). Acoustic properties for place of articulation in nasal consonants. *The Journal of the Acoustical Society of America*, *81*, 1917–1927.
Lamel, L. (1988). *Formalizing knowledge used in spectrogram reading: Acoustic and perceptual evidence from stops*. Cambridge: Massachusetts Institute of Technology.
LaRiviere, C., Winitz, H., & Herriman, E. (1975). The Distribution of Perceptual Cues in English Prevocalic Fricatives. *Journal of Speech and Hearing Research*, *18*, 613–622.
Lee, S.-H., Yu, J.-F., Hsieh, Y.-H., & Lee, G.-S. (2015). Relationships Between Formant Frequencies of Sustained Vowels and Tongue Contours Measured by Ultrasonography. *American Journal of Speech-Language Pathology*, *24*, 739–749.
Leemann, A., Schmid, S., Studer-Joho, D., & Kolly, M.-J. (2018). Regional Variation of /r/ in Swiss German Dialects. *Interspeech 2018*, 2738–2742. ISCA.
Leinonen, T. (2008). Factor Analysis of Vowel Pronunciation in Swedish Dialects. *International Journal of Humanities and Arts Computing*, *2*, 189–204.
Li, F., Edwards, J., & Beckman, M. (2007). Spectral measures for sibilant fricatives of English, Japanese, and Mandarin Chinese. *Proceedings of the International Congress of Phonetic Sciences*. Presented at the ICPHS 2007.
Li, S., & Gu, W. (2015). *Acoustic analysis of Mandarin affricates*. Presented at the Sixteenth Annual Conference of the International Speech Communication Association.
Ma, D., Ryant, N., & Liberman, M. (2021). Probing Acoustic Representations for Phonetic Properties. *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 311–315. Toronto, ON, Canada: IEEE.
Maddieson, I. (1984). The effects on F0 of a voicing distinction in sonorants and their implications for a theory of tonogenesis. *Journal of Phonetics*, *12*, 9–15.
Maddieson, I. (2008). Glides and gemination. *Lingua*, *118*, 1926–1936.
Maeda, S., & Honda, K. (1994). From EMG to Formant Patterns of Vowels: The Implication of Vowel Spaces. *Phonetica*, *51*, 17–29.
Malecot, A. (1956). Acoustic Cues for Nasal Consonants: An Experimental Study Involving a Tape-Splicing Technique. *Language*, *32*, 274.
Maniwa, K., Jongman, A., & Wade, T. (2009). Acoustic characteristics of clearly spoken English fricatives. *The Journal of the Acoustical Society of America*, *125*, 3962–3973.
Martínez-Celdrán, E., & Regueira, X. L. (2008). Spirant approximants in Galician. *Journal of the International Phonetic Association*, *38*.

Matsumoto, H., & Wakita, H. (1986). Vowel normalization by frequency warped spectral matching. *Speech Communication*, *5*, 239–251.
McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., & Sonderegger, M. (2017). Montreal forced aligner: Trainable text-speech alignment using kaldi. *Interspeech 2017*. Presented at the Interspeech.
McCarthy, D. T. P. D. (2019). *The acoustics of place of articulation in English plosives* (Newcastle University). Newcastle University.
Metfessel, M. (1929). Experimental phonetics. *Psychological Bulletin*, *26*, 305–323.
Morris, J., & Hejná, M. (2020). Pre-aspiration in Bethesda Welsh: A sociophonetic analysis. *Journal of the International Phonetic Association*, *50*, 168–192.
Nakata, K. (1959). Synthesis and Perception of Nasal Consonants. *The Journal of the Acoustical Society of America*, *31*, 661–666.
Narayanan, S., Byrd, D., & Kaun, A. (1999). Geometry, kinematics, and acoustics of Tamil liquid consonants. *The Journal of the Acoustical Society of America*, *106*, 1993–2007.
Nartey, J. N. A. (1982). *On fricative phones and phonemes: Measuring the phonetic differences within and between languages*. University of California, Los Angeles. es
Ohde, R. N., & Stevens, K. N. (1983). Effect of burst amplitude on the perception of stop consonant place of articulation. *The Journal of the Acoustical Society of America*, *74*, 706–714.
Orellana, S., & Ugarte, J. P. (2021). Vowel characterization of Spanish speakers from Antioquia–Colombia using a specific-parameterized discrete wavelet transform analysis. *Applied Acoustics*, *172*, 107635.
Palaz, D., Magimai-Doss, M., & Collobert, R. (2019). End-to-end acoustic modeling using convolutional neural networks for HMM-based automatic speech recognition. *Speech Communication*, *108*, 15–32.
Pasad, A., Chou, J.-C., & Livescu, K. (2021). Layer-Wise Analysis of a Self-Supervised Speech Representation Model. *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 914–921. Cartagena, Colombia: IEEE.
Pasad, A., Shi, B., & Livescu, K. (2023). Comparative Layer-Wise Analysis of Self-Supervised Speech Models. *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. Rhodes Island, Greece: IEEE.
Perkell, J. S., Klatt, D. H., & Stevens, K. N. (Eds.). (1986). *Invariance and variability in speech processes*. Hillsdale, N.J: Lawrence Erlbaum Associates.
Peterson, G. E., & Barney, H. L. (1952). Control Methods Used in a Study of the Vowels. *The Journal of the Acoustical Society of America*, *24*, 175–184.
Rao, D., & Shaw, J. A. (2021). The role of gestural timing in non-coronal fricative mergers in Southwestern Mandarin: Acoustic evidence from a dialect island. *Journal of Phonetics*, *89*, 101112.
Rebernik, T., Jacobi, J., Jonkers, R., Noiray, A., & Wieling, M. (2021). A review of data collection practices using electromagnetic articulography. *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, *12*, 6.
Recasens, D. (1983). Place cues for nasal consonants with special reference to Catalan. *The Journal of the Acoustical Society of America*, *73*, 1346–1353.

Riera, P., Cerdeiro, M., Pepino, L., & Ferrer, L. (2023). Phone and speaker spatial organization in self-supervised speech representations. *2023 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, 1–5.

Sanabria, R., Tang, H., & Goldwater, S. (2023). Analyzing Acoustic Word Embeddings from Pre-Trained Self-Supervised Speech Models. *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. Rhodes Island, Greece: IEEE.

Schutte, K. T. (2009). *Parts-based models and local features for automatic speech recognition* (PhD Thesis). Massachusetts Institute of Technology, Department of Electrical Engineering ….

Scobbie, J., Punnoose, R., & Khattab, G. (2013). Articulating five liquids: A single speaker ultrasound study of Malayalam. In L. Spreafico & A. Vietti (Eds.), *Rhotics: New data and perspectives* (1st edition). Bozen: Bozen-Bolzano University Press.

Shadle, C.H., & Mair, S. J. (1996). Quantifying spectral characteristics of fricatives. *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP '96*, *3*, 1521–1524. Philadelphia, PA, USA: IEEE.

Shadle, Christine H., Chen, W.-R., Koenig, L. L., & Preston, J. L. (2023). Refining and extending measures for fricative spectra, with special attention to the high-frequency range. *The Journal of the Acoustical Society of America*, *154*, 1932–1944.

Shah, J., Singla, Y. K., Chen, C., & Shah, R. R. (2021, July 12). *What all do audio transformer models hear? Probing Acoustic Representations for Language Delivery and its Structure*. arXiv.

Sharifzadeh, H. R., McLoughlin, I. V., & Russell, M. J. (2012). A Comprehensive Vowel Space for Whispered Speech. *Journal of Voice*, *26*, e49–e56.

Shue, Y.-L., Keating, P., Vicenik, C., & Kristine, Y. (2011). VoiceSauce: A program for voice analysis. *Proceedings of the International Congress of Phonetic Sciences*. Presented at the ICPhS XVII, Hong Kong. Hong Kong.

Soli, S. D. (1981). Second formants in fricatives: Acoustic consequences of fricative-vowel coarticulation. *The Journal of the Acoustical Society of America*, *70*, 976–984.

Spinu, L., Kochetov, A., & Lilley, J. (2018). Acoustic classification of Russian plain and palatalized sibilant fricatives: Spectral vs. cepstral measures. *Speech Communication*, *100*, 41–45.

Spinu, L., & Lilley, J. (2016). A comparison of cepstral coefficients and spectral moments in the classification of Romanian fricatives. *Journal of Phonetics*, *57*, 40–58.

Stevens, K. N. (1985). Spectral prominences and phonetic distinctions in language. *Speech Communication*, *4*, 137–144.

Stevens, K. N. (2000). Diverse Acoustic Cues at Consonantal Landmarks. *Phonetica*, *57*, 139–151.

Stilp, C. (2020). Acoustic context effects in speech perception. *WIREs Cognitive Science*, *11*, e1517.

Strevens, P. (1960). Spectra of Fricative Noise in Human Speech. *Language and Speech*, *3*, 32–49.

Styler, W. (2017). On the acoustical features of vowel nasality in English and French. *The Journal of the Acoustical Society of America*, *142*, 2469–2482.

Tabain, M. (2001). Variability in Fricative Production and Spectra: Implications for the Hyper- and Hypo-and Quantal Theories of Speech Production. *Language and Speech*, *44*, 57–93.

Tabain, M., Butcher, A., Breen, G., & Beare, R. (2016). An acoustic study of nasal consonants in three Central Australian languages. *The Journal of the Acoustical Society of America*, *139*, 890–903.

Tahiry, K., Mounir, B., Mounir, I., & Farchi, A. (2016). Energy bands and spectral cues for Arabic vowels recognition. *International Journal of Speech Technology*, *19*, 707–716.

Takeuchi, S., Kasuya, H., & Kido, K. (1975). A Method for Extraction of the Spectral Cues of Nasal Consonants. *Journal of the Acoustical Society of Japan (E)*.

Thomas, E. (2017). *Sociophonetics: An Introduction*. Bloomsbury Publishing.

Ulrich, N., Allassonnière-Tang, M., Pellegrino, F., & Dediu, D. (2021). Identifying the Russian voiceless non-palatalized fricatives /f/, /s/, and /ʃ/ from acoustic cues using machine learning. *The Journal of the Acoustical Society of America*, *150*, 1806–1820.

V. Latoszek, B. B., Maryn, Y., Gerrits, E., & De Bodt, M. (2018). A Meta-Analysis: Acoustic Measurement of Roughness and Breathiness. *Journal of Speech, Language, and Hearing Research*, *61*, 298–323.

Villarreal, D., Clark, L., Hay, J., & Watson, K. (2020). From categories to gradience: Auto-coding sociophonetic variation with random forests. *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, *11*, 6.

Vizza, P., Mirarchi, D., Tradigo, G., Redavide, M., Bossio, R. B., & Veltri, P. (2017). Vocal signal analysis in patients affected by Multiple Sclerosis. *Procedia Computer Science*, *108*, 1205–1214.

Vu, N. T., Wang, Y., Klose, M., Mihaylova, Z., & Schultz, T. (2014). Improving ASR performance on non-native speech using multilingual and crosslingual information. *Interspeech 2014*, 11–15. ISCA.

Vu, N. T., Weiner, J., & Schultz, T. (2014). Investigating the learning effect of multilingual bottle-neck features for ASR. *Fifteenth Annual Conference of the International Speech Communication Association*.

Warden, P. (2018). Speech commands: A dataset for limited-vocabulary speech recognition. *arXiv Preprint arXiv:1804.03209*.

Watson, C. I., & Harrington, J. (1999). Acoustic evidence for dynamic formant trajectories in Australian English vowels. *The Journal of the Acoustical Society of America*, *106*, 458–468.

Whalen, D. H., DiCanio, C., & Dockum, R. (2022). Phonetic documentation in three collections: Topics and evolution. *Journal of the International Phonetic Association*, *52*, 95–121.

Wikse Barrow, C., Włodarczak, M., Thörn, L., & Heldner, M. (2022). Static and dynamic spectral characteristics of Swedish voiceless fricatives. *The Journal of the Acoustical Society of America*, *152*, 2588–2600.

Wu, Z., Qian, Y., Soong, F. K., & Zhang, B. (2008). Modeling and Generating Tone Contour with Phrase Intonation for Mandarin Chinese Speech. *2008 6th International Symposium on Chinese Spoken Language Processing*, 1–4. Kunming, China: IEEE.

Yoshii, K., Goto, M., & Okuno, H. (2005). *Adamast: A drum sound recognizer based on adaptation and matching of spectrogram templates*. Presented at the 1st Annual Music Information Retrieval Evaluation eXchange (MIREX).

Yu, A. C. L., Lee, C. W. T., Lan, C., & Mok, P. P. K. (2022). A New System of Cantonese Tones? Tone Perception and Production in Hong Kong South Asian Cantonese. *Language and Speech*, *65*, 625–649.

Zhou, M., Liu, X., Liu, D., Wu, Z., Liu, Z., Zhao, L.,. (2023). Fine-grained artificial neurons in audio-transformers for disentangling neural auditory encoding. *Findings of the Association for Computational Linguistics: ACL 2023*, 7943–7956.

Zue, V. W. (1982). Acoustic-Phonetic Knowledge Representation: Implications from Spectrogram Reading Experiments. In J.-P. Haton (Ed.), *Automatic Speech Analysis and Recognition* (pp. 101–120). Dordrecht: Springer Netherlands.

Zygis, M., & Hamman, S. (2003). Perceptual and acoustic cues of Polish coronal fricatives. *Proceedings of the International Congress of Phonetic Sciences*. Presented at the ICPHS 15, Barcelona. Barcelona.