UC Irvine LSCI 117 Winter 2024

Acoustic Phonetics

Professor: Connor Mayer
Notes: Timothy Cho

March 2024
Lecture Note Series #**7**

Introduction

These notes come from both the lecture and the discussion, and are roughly sorted by content. Sections are numbered chronologically using the following scheme by taking the section number modulo 10. Note that we have occasionally merged two sections for continuity reasons.

Date	Lecture	Discussion
Monday	0	1
Tuesday	2	3
Wednesday	4	5
Thursday	6	7
Friday	8	9

No text was used for this course.

14 Overview

Phonetics is an interdisciplinary field focusing on speech, especially in the physical sense. This is in contrast to *phonology*, which concentrates on abstract representations of sounds. However, the two areas interact, in that our physical observations dictate abstract representations. Some subareas of phonetics include:

1. Speech production — how do humans make speech sounds using their body?
2. Speech perception — how do humans make sense of speech sounds?
3. Speech acoustics — how does linguistic structure come about using properties of speech sounds?

In this course, we will mainly focus on (3), with some foray into both (1) and (2). More precisely, a guiding question of this course is, *what are the critical acoustic properties* of a sound, and what does that mean?

Recall that we transcribe speech mainly using the International Phonetic Alphabet (IPA) and not using a language's orthography, as the IPA allows us to write down sounds in any refinement of transcription we desire. In general, we care about two levels of refinement:

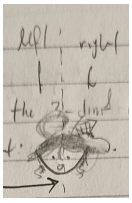
- A *phonemic transcription* marks only differences hat are contrastive in a given language. These are usually informed by a phonological study of the language.
- A *phonetic transcription* works more precisely, marking important allophonic differences.

We usually denote phonemic transcriptions between slashes $/\cdots/$, and phonetic transcriptions with brackets $[\cdots]$.

24 Physical Speech Production

In this section, we will discuss, in sweeping generality, how people produce sounds. When discussing speech sounds and the vocal tract, we will often project the three dimensional body onto a *mid-saggital projection*, which cuts the body into equal left and right pieces.

For example, here is Furina viewed from the front, showing a forward-facing slice of a mid-saggital projection:



This cross-section is useful, but we miss three important details:

1. Things off of the mid-line are missing — for example, if we have laterality, we cannot tell which way Furina points her tongue off the center.
2. Roundness of lips cannot be detected at all.
3. Information about muscle movement is missing.

Despite this, the mid-saggital view is predominant because it is useful — for other purposes, we might take a *coronal projection*, which cuts the body into front and back parts. This allows us to see laterality and rounding.

In general, through our observations, we subdivide the production of sound into 4 distinct segments, which are (relatively) independent of each other.

Airstream Process

We start our sound production process through airflow: air needs to flow through the vocal tract to make sounds, by the very nature of what sound is. In general, our sounds use the *pulmonic egressive* airflow type, where air flows out (egressive) of the lungs (pulmonic) as the volume of air is decreased in the lungs. This airflow process is predominant in **all** languages, so we will mainly focus on this. Other types of airflow do exist, but we will revisit them at the end of the course.

Phonation Process

Assuming that we are using a pulmonic egressive airflow process, air leaaves the lungs and enters the larynx, which contains the *vocal cords* and the *glottis* (the space between the vocal cords). Now, three things can happen here:

1. If the vocal folds are far apar (being *abducted*), no voicing will occur.
2. If the vocal folds are close together (being *adducted*), they will produce *modal voicing*.
3. If the vocal folds are completely adducted, they will produce the glottal stop [ʔ].

There are intermediate cases between (1) and (2), which will result in *breathy* and *creaky* voice, but we will revisit this at the end of the course.

Oro-Nasal Process

After air leaves the larynx, it enters the back of the vocal tract, consisting of three important parts: the *velum* opens and closes to let the air through the nose, the *pharynx* is the back of the throat, and the *velopharyngeal port* (VPP) is the space in between.

That is, if the velum closes, then the VPP is closed, and hence no air goes through the nose. If the velum is open, then air flows through the nose and we get nasality.

Articulatory Process

Finally, air flows into the oral cavity, which is the focus of this course. We will study this in far more detail, but in general, the objects in the mouth fall into two types:

1. Active articulators: these are the things that move — e.g., lips, tongue, jaw, etc.
2. Passive articulators: these are things that are contacted by active articulators to make sounds — e.g., teeth, alveolar ridge, etc.

Note that items like the soft palate or the pharyngeal wall could be both active and passive, depending on context.

28 Articulation of Consonants and Vowels

In this section, we consider the articulatory process more precisely, and we separate this among consonants and vowels.

Consonants

Intuitively, we know the distinction between consonants and vowels, but we make the following informal definition: *consonants* are sounds produced with a relatively narrow construction, and usually take up syllable margins.

We classify consonants using the five metrics below.

Voicing

This category is self-explanatory: consonants are either voiced [b, d, g] or voiceless [p, t, k].

Place of Articulation

We have three broad categories for these:

1. *Labial* consonants involve the lips: these include the bilabials [p, b, m, w] and the labiodentals [f, v].
2. *Coronals* involve the front of the tongue contacting structures in the mouth. These include dentals [θ], alveolars [t, d, s, n, l], retroflexes [ɹ], palatoalveolars [ʃ], and palatals [j].
    - The case of [ɹ] is actually far more complicated in English, and it is by far the most difficult sounds for children to produce, being acquired rather late. People differ drastically in their production of [ɹ], so it could be alveolar, retroflex, or palatoalveolar.
3. *Dorsals* involve the back of the tongue. These include velars [k, g], uvulars [q], and pharyngeals [ħ].

Manner/Nasality

Roughly, this category describes the width of the constriction made to produce the consonant. We will discuss this further when exploring consonant acoustics.

1. *Stops* (or *plosives*) are characterized by a complete closing of the vocal tract, followed by a release: [b, p, t]. This includes nasals, such as [m] or [n].
2. *Fricatives* have a constriction narrow enough to cause turbulence: [f, v, θ, s, z, h]. What this turbulence actually is will be considered later.
3. *Approximants* have a constriction too large to cause turbulence, yet too small to be considered a vowel: [j, ɹ, l, w].
4. *Trills* involve articulators repeatedly bouncing off each other, such as the Spanish [r]. In contrast, *taps* consist of a single such bounce, such as [ɾ].
5. *Affricates* are stops with a fricative release, functioning as one sound: [t͡ʃ, t͡s].

Laterality

This category refers to closure *off* the mid-saggital plane, but is also characterized by airflow off the midline: [l] is the main example. Sounds that are not lateral are called *central*.

Vowels

Vowels, in contrast, are sounds that involve very little constriction in the vocal tract, and often fill syllable nuclei. Typically, they are articulated with the tongue dorsum, and are classified under 3 main dimensions.

Height/Backness

These two categories refers to the position of the tongue in the mouth, vertically and horizontally. Very roughly, we have the following chart:

	Front	Central	Back
High	[i, y, ɪ]	[ɨ]	[u]
Mid	[ɛ]	[ə]	[o]
Low	[æ]		[ɑ]

Rounding

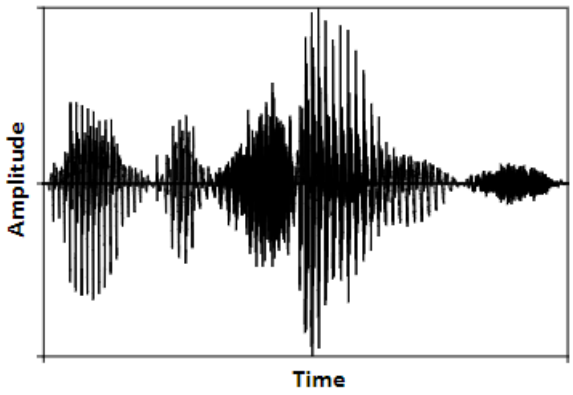
This refers to the shape of the lips. Typically, back vowels are more likely to be rounded, while front ones are not.

Suprasegmentals

Outside of specific sound segments, speech could be distinguished by other means as well, such as by syllabification, stress, or tone. We will discuss suprasegmentals at the end of this course.

30 Sounds on Computers I: Sampling

A main tool in phonetics studies are computers, as we often need to make many quantitative measures of sound. Thus, at this point, we should define what sound is: *sound* is a sensation that we experience when our auditory nerves are stimulated by vibrating air molecules. That is, sound is what soundwaves cause us to perceive mentally. When we record sounds, we *transduce* and store soundwaves — we will see that a microphone is simply an artificial eardrum, which converts variations in sound pressure caused by vibrating air into electrical voltage. The output of such a conversion is a waveform:



Here, time is plotted on the x-axis, while variations in pressure is plotted on the y-axis. *Amplitude* refers to the maximum difference in pressure, and is thus the "height" of the soundwave — the larger the amplitude, the louder we perceive the sound.

However, we are presented with a problem when we try to represent sound on a computer: both pressure and time are continuous variables, while computers are strictly discrete. It follows that any recorded sound we hear on a computer is a discrete *approximation* to the original audio signal — we take *samples* of the air pressure at evenly spaced points at time. That is, at each time t^*, we record an amplitude a^*, and we record the sequence (t^*, a^*). We thus have one consideration, phrased in two ways:

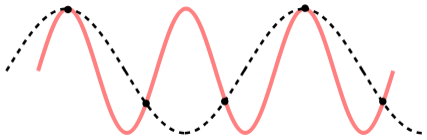
1. What is the *sampling interval* T, i.e., the time between two adjacent samples?
2. What is our *sampling rate* r, i.e., the number of samples we take per unit time?

Of course, we have the inverse relationship $r = 1/T$, and it follows that higher sampling rates are better approximations of the original audio signal, but take up a larger file size. Of course, the sampling rate will depend on the type of sound we want to measure, but we turn to the following theorem for a guideline.

Theorem 30.1 (Nyquist-Shannon Sampling Theorem)**.** *If the waveform we want to sample contains frequencies no higher than B, then a sample rate of at least $2B$ is sufficient to capture all of the information about the waveform.*

Proof. This involves a fair amount of Fourier analysis, which is outside the scope of this course. □

Given a sample rate r, the Nyquist frequency is the quantity $r/2$, which is the highest frequency that can be preserved using the sample rate r. The bound given by the Nyquist-Shannon theorem is in fact sharp: anything less will result in *aliasing*, where our measurements "miss" the shape of the waveform:



34 Sounds on Computers II: Quantization

In the previous lecture, we saw that sampling rates tell us when to sample, but now we need to know what amplitudes to sample. Again, amplitude is a continuous variable, so it must be discretized and approximated. We divide up the relevant range of amplitudes into an equal number of values, usually represented using some number of bits, and this number of bits is called the *quantization rate* or *bit depth*. It follows that the number of possible amplitudes, given a bit depth of n, is 2^n, so increasing the bit depth gives us more precision over our amplitudes, for a slight increase in file size.

This representation of sound thus informs us of how to make recordings. In general, recording should be as loud as possible, thus taking up most of the amplitude space we get from our bit depth, but not loud enough to cause *clipping*, i.e., exceeding the amplitudes which the recording device could handle.

We may also consider the *bit rate* of an audio file, defined as follows:

$$\text{bit rate} = \frac{\text{total number of bits}}{\text{length of audio in seconds}}, \tag{1}$$

which is related to both the sampling and quantization rates.

Complex Waves

The simplest periodic sound is a *sine wave*, which has an equation of the form $y = a\sin(bx + c) + d$. For sine waves, it is easy to read off their frequency: $b/(2\pi)$. However, most periodic sounds, including speech, are not sine waves but *complex waves*, which consist of multiple component frequencies, possibly at different amplitudes. Two mathematical techniques are used here: we will describe them, but not in detail.

1. *Fourier analysis* decomposes a function into a possibly infinite sum of sine waves.
2. *Fourier synthesis* produces a function by summing sine waves.

Decomposing a complex wave into a sum of finitely many sine waves gives us the following definition.

Definition 34.1. The *fundamental frequency* of a complex wave, denoted f_0, is the frequency of the lowest component wave. All other frequencies of component waves are called *overtones*, and overtones that are integer multiples of f_0 are *harmonics*.

Most speech (and most periodic sound in fact) contain harmonics, so we will be interested in analyzing them.

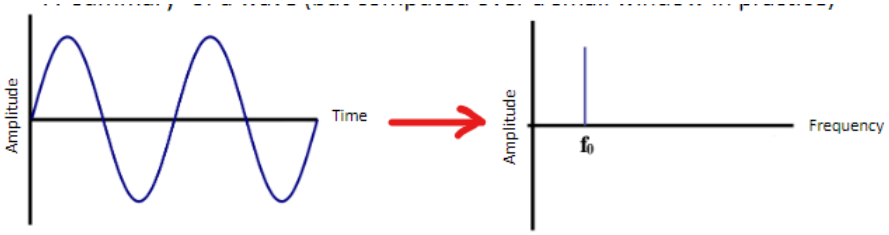
Something interesting occurs *perceptually* when dealing with harmonics. For example, consider the complex wave consisting of the frequencies $\{100, 200, 300, 400, 500\}$ Hertz — in our pitch perception, we perceive the fundamental $f_0 = 100$, and there are five harmonics. Now, consider the modified wave, consisting of the frequencies $\{300, 400, 500\}$. Here, the fundamental is $f_0' = 300$; however, the 400 Hz and 500 Hz overtones are not harmonics. Nonetheless, our brain perceives harmonics and subconsciously calculates

$$\gcd(300, 400, 500) = 100,$$

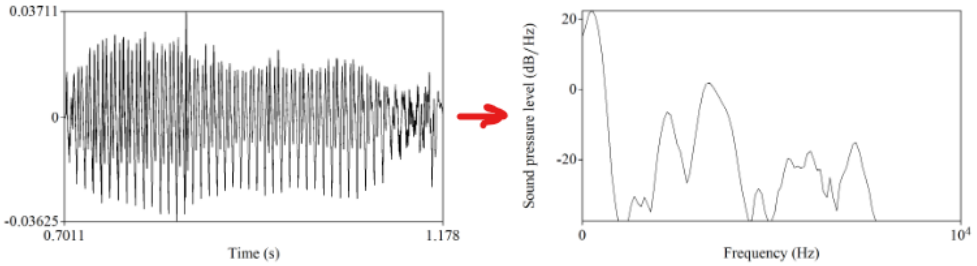
so we hear this wave with a fundamental at 100 Hz, but three strong overtones at $300, 400,$ and 500 Hz respectively.

Spectra

Often, we will need to work with visual data that shows the frequency information of a complex wave, and we will be interested in the relative amplitudes of the frequencies at hand. We do this via *spectra*: which are visualizations of the component frequencies of a complex wave, graphed by amplitude:

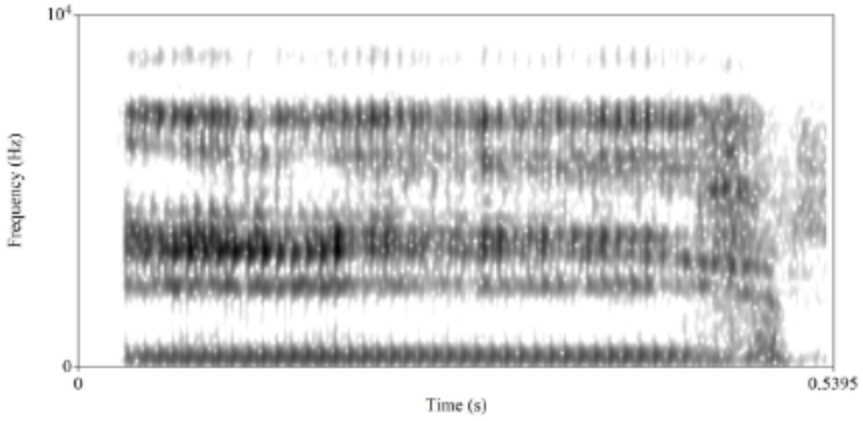


Here, we have a simple wave, decomposed into its *line* spectrum on the right, which, as expected, consists of a singular line corresponding to the frequency of the sine wave. However, line spectra in practice are idealized, so *power spectra* are used instead:



This gives a far more realistic, but probabilistic, view of the component frequencies of a wave. Aperiodic sounds may have spectra as well; however, these sounds are not cyclic in nature, and thus the air pressure variations look random.

However, a problem with spectra is that they capture a wave at a moment in time: if the wave changes dramatically, a spectrum, say taken at time t_0, will no longer be valid at time t_1. To compensate for this, we may overlay time against frequency, using a crucial visualization called a *spectrogram*:



Here, amplitude is described by the darkness of the chart: the areas of the spectrogram which are darkest have the highest amplitude concentration. Later in the course, we will see how to draw out acoustic observations from spectrograms.

38 Source-Filter Theory I

Now after our digression into representations of sound on a computer, we return to phonetics to discuss the production of sound in the vocal tract. We will do this through the lens of *source-filter theory*, which posits that acoustic energy is produced in the vocal tract by *sources*, which we will define, and this energy is then *filtered* by physical properties of the vocal tract, which either amplify or attenuate frequencies of a complex wave.

We first examine the four types of *sources* used in speech production.

1. Voicing: this is a complex periodic signal, containing a fundamental frequency f_0, as well as many important overtones. The fundamental frequency is given by the rate of vibration of the vocal folds, and this may be seen visually by regular waveforms, or dark bands on the bottom of a spectrogram.

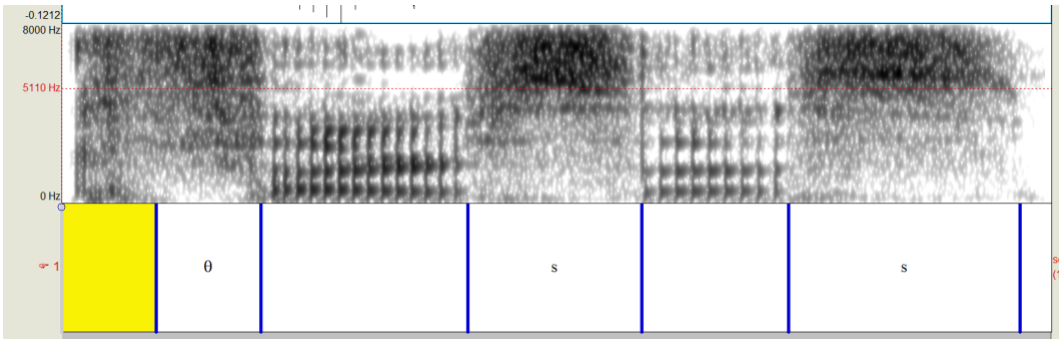
2. Frication: this is aperiodic noise, generated in the vocal tract due to turbulence in the airflow. More specifically, frication is generated at the place of articulation.

   - Some fricatives generate additional noise downstream, and thus are louder: for example, [s] has frication coming from the alveolar ridge, but this is pushed through the teeth, generating more noise.
   - Voiced fricatives combine the voicing and frication source; however, this ends up being difficult to produce: voicing slows the airflow, yet frication requires airflow to cause turbulence. Thus, voiced fricatives are usually softer, or may not be voiced at all in certain contexts.

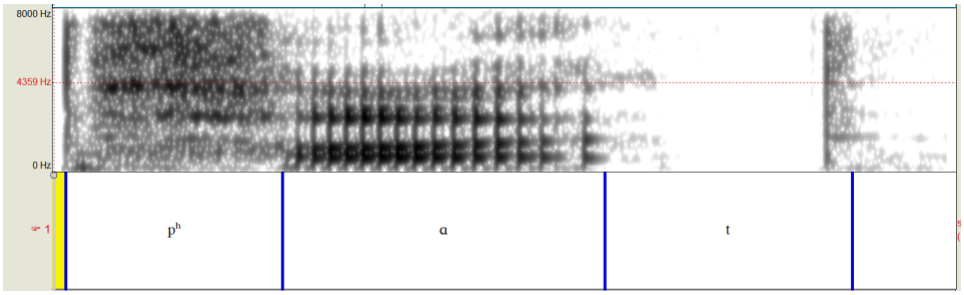
3. Aspiration: this is similar to frication, except that the noise is generated in the glottis. This occurs for [h] and aspirated consonants, and voiced fricatives may also be post-aspirated due to airflow through the glottis.

4. Transience: this is a noise source, but for a very short, impulse-like noise usually termed *bursts* or *releases*. These exist in plosives, ejectives, implosives, and affricates, but not nasal stops. Generally, these are detected by an obvious spike in the waveform, or a short, dark band in the spectrogram.

We view some examples below.



The spectrogram above shows a production of the word "catharsis," with the fricatives segmented. Notice that the fricatives are clearly distinguished from the vowels (which have "voicing bars").



The spectrogram above shows a production of the word "pot," with all segments marked. Notice that aspiration looks quite similar to frication above, and the release of the [t] is marked by a dark band at the end of the spectrogram, after a period of silent closure of the vocal tract.

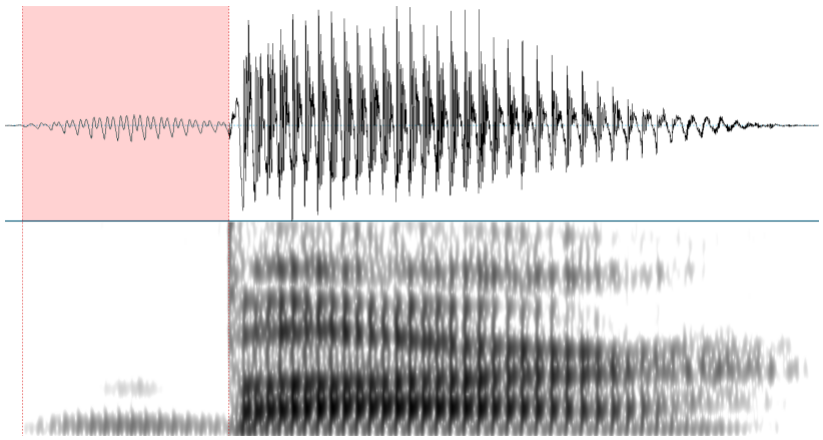
Voice Onset Time

The above example also highlights an important measurement we might desire to make. We have took note of the period of silence in the [t] above, followed by its release, and we have mentioned before that not all voiced sounds may be actually voiced. To quantify this, we use the following measurement:

Definition 38.1. The *voice onset time* is the elapsed time from the release of the stop to the onset of voicing:

$$\text{VOT} = (\text{time of voicing}) - (\text{time of burst}) \tag{2}$$

Here, we note that we are making a comparison between the voicing and transient sources of a stop. When VOT ≥ 0, we have *lag VOT*, and the burst comes before the voicing onset. When VOT < 0, we have *prevoicing*, where the stops are voiced before it is released. This latter case can be seen by examining the spectrogram for a f_0 bar, unaccompanied by anything else. The below is one such example of prevoicing:



In general, both the waveform and the spectrogram are important for identifying voicing onset:

- Time resolution is better in the waveform, as we can see the beginning of periodic cycles.
- The spectrogram gives information through the voicing bar, which marks out f_0.

However, there is a bit of guesswork to do in this as well, especially if the waveform is unclear — the idea is to choose a convention and use it consistently.

Finally, we note that some stops, especially those in the back of the vocal tract, can have multiple releases. The VOT is thus measured starting from the *first* release.

VOT Across Languages

VOT was first proposed as a measurement by Lisker and Abramson (1964) to compare voicing and aspiration distinctions across stops in different languages. In general, there are three distinctions: negative VOT (truly voiced), short lag (voiceless), or long lag (aspirated). Generally, most languages take two of these distinctions, though some have a three-way distinction. Occasionally, *short lag* stops are considered voiced (e.g. English, Mandarin), but unvoiced in others (e.g. French).

44 Source-Filter Theory II

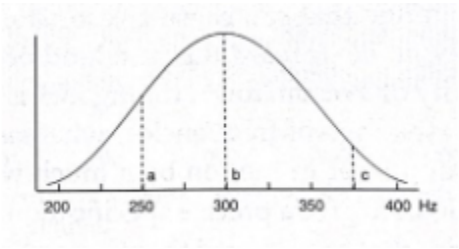
Recall that the voice source outputs a complex periodic wave. This wave consists of many harmonics of the form $n \cdot f_0$, where n is a positive integer — in general, we say that the frequency $n \cdot f_0$ is the nth harmonic. Now, we move into the "filter" part of source-filter theory, which describes what happens to harmonics in the vocal tract.

We have said before that filters do the job of either boosting or reducing certain harmonics after the sound exists the larynx — however, this presents us with a problem: the speech we hear is always filtered, and thus it may seem impossible to "unfilter" a sound to determine the original source material. Luckily, this is not a huge issue, as there are ways to approximately unfilter a sound, which is an advantage, as sources and filters are largely independent of each other.

Through this unfiltering, we will see that in the source, harmonics tend to decrease as frequency increases, and the rate of decrease depends on the phonation type (which we will discuss later). However, the filter will modify this decrease sequence of sounds, which differs based on the sounds that are produced. We note that filters *do not* change harmonics, nor do they add harmonics to the source material: they simply amplify or attenuate certain frequencies.

Resonances

The way that the vocal tract acts as a filter is through resonant properties. Resonance is a type of sympathetic vibration, which occurs when something near an object resonates at its *natural frequencies*. In general, the responsiveness to vibration depends on the *resonance curve* of that object:



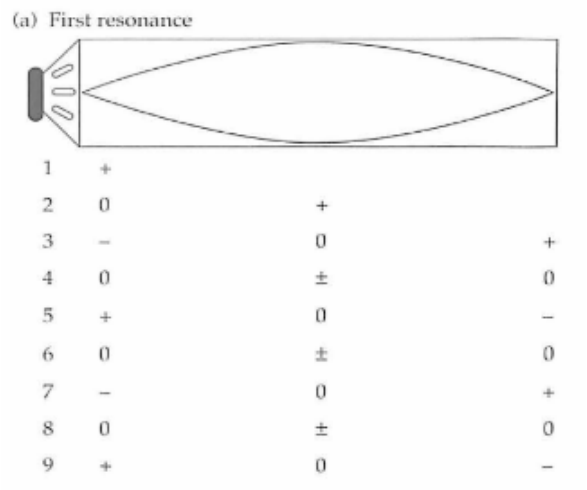
Here, the peak of the curve is 300 Hz, so at this *resonance peak*, the object's likelihood to sympathetically vibrate is at its maximum. Similarly, the vocal tract is an object, where changing the configuration of the vocal tract changes the resonance curve. With respect to the vocal tract, the resonance curve is called a *filter function*, and resonance peaks are known as *formants*. Thus, air in the vocal tract will vibrate only if a source frequency matches a formant frequency; otherwise, the frequency is decreased on amplitude — the filter function completely dictates what happens to a frequency as it passes through the vocal tract.

48 Tube Models I

In order to gain an understanding of the vocal tract, it is best to simplify our models, though we lose some precision. First, we model the vocal tract (the lips to the glottis) as a uniform tube, closed at one end (glottis — this opening is very small, so we treat it as a closure) and open at the other (lips). This is roughly the vocal tract configuration of the vowel [ə].

A *standing wave* describes the peak amplitude profile of a wave over the tube, and it is a sum of moving waves. In tube models, these are displacement waves, which are inversely related to pressure waves — high and low pressure fixes particles in place, and changing pressure moves particles. We first view what happens for a tube *closed* at both ends, in order to motivate our case of a tube closed at one end.

For a closed tube, we can define a family of resonant frequencies that are parameterized by the length L of the tube.



Consider the figure above. The + and − symbols refer to places of high and low pressure, respectively, as they vary along the tube of length L from time 1 to time 9. Notice that this wave is ideal, in that we get strong resonances due to the fact that the start of a new $+/-$ cycle coincides with the return of a $+/-$ cycle. For this to happen, the soundwave must travel twice the length of the tube (back and forth). Hence, if λ_1 is the wavelength of the lowest resonant frequency, we have $\lambda_1 = 2L$. In general, it is not too hard to show $\lambda_n = 2L/n$, where λ_n is the nth lowest resonance frequency of the tube.

For tubes open at one end, we can try the same thing, except that we need to account for *polarity change*, which flips a wave from + to −. Through a similar calculation, we can find the formula

$$\lambda_n = \frac{4L}{2n-1}. \tag{3}$$

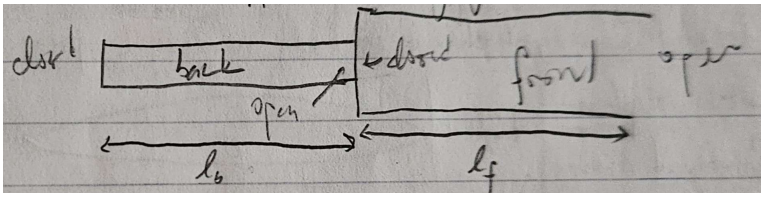
Now, we can easily solve for the formants F_n by using $\lambda_n F_n = c$, where c is the speed of sound (around $33000 - 35000$ cm/sec):

$$F_n = \frac{(2n-1)c}{4L}. \tag{4}$$

50 Tube Models II

In the previous section, we modeled the vocal tract as a uniform tube, closed at one end, and remarked that that was roughly the vocal tract configuration when producing the vowel [ə]. However, different vowels get their qualities from different filter functions, which rely upon more tube models.

For example, consider the low back vowel [ɑ], where the pharynx is narrowed by the tongue, and the oral cavity is enlarged. Now, using a uniform tube will do disservice to the accuracy of our model, but it is sufficient to use instead *two* uniform tubes of different diameters (but of similar length), both closed at one end and open at the other:

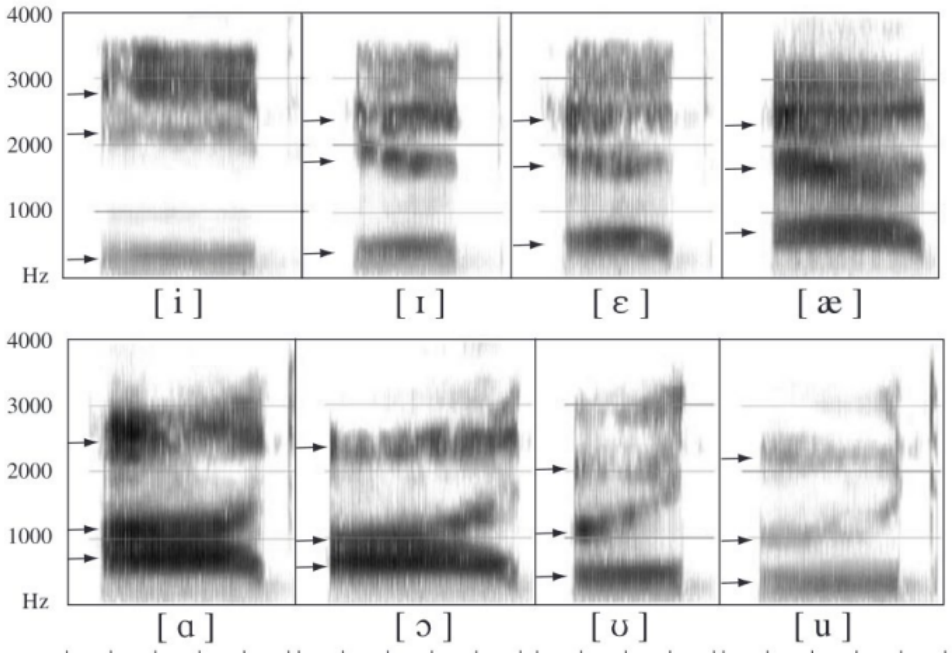


Hence, [ɑ] is almost like 2 productions of [ə]: the two sets of resonances are *similar*, but not identical. Each tube generates its own formants, and formants are aggregated from lowest to highest, regardless of source, so that the increase in frequency from formant to formant is nonlinear for [ɑ].

After compiling more models, we arrive at the following acoustic correlates for vowels.

- The first formant, F_1, is inversely related to vowel height: the higher the F_1, the lower the vowel.
- The second formant, F_2, is related to vowel backness: the higher the F_2, the fronter the vowel.
- Lip rounding lowers all formants, by extending the tube by a bit.

Hence, we can easily read off vowel information from spectrograms.



Certain features of this chart should be memorized: for example, the first two formants of [i] are very far apart, while the same formants of [ɑ] are essentially juxtaposed.

54 Experimental Design

The next two sections form a short interlude on experimental design and statistics that will be relevant for this course. The study of phonetics is built from questions to be explored, which usually involves these things:

1. Making a falsifiable hypothesis as to the solution of the question;
2. Testing the hypothesis using experiments.

In the realm of statistics, we usually make two hypotheses:

1. The *null hypothesis*, denoted H_0 states that there is no effect or difference between the control dataset and the experimental dataset. The *alternative hypothesis*, denoted H_1, states otherwise.
2. A *directional hypothesis* posits a directional effect on some experimental variable. This is usually what we hypothesize, but we usually test the null hypothesis to avoid bias.

We also have the following terminology for variables in experiments:

1. An *independent variable* is one that is manipulated between different experimental groups.
2. A *dependent variable* is a quantity that we want to measure.
3. A *controlled variable* is one that is kept constant between different experimental groups.

After collecting data using experiments, we move into statistical testing, which tells us whether we should accept or reject the null hypothesis. Alternatively, we could decide that our experiment had a design flaw, and then redo it.

64 Inferential Statistics

Suppose we have an experiment which gave us two datasets, corresponding to the control and experimental groups, and suppose the experiment was decidedly not flawed. At this point, we need to decide whether they two datasets are the same or different — i.e., do they come from the same distribution? To do this, we need to consider the central tendencies of both data sets, as well as the dispersion of the samples, and this is done through *significance testing*, which returns a probability that we should observe the current set of data, if we assume the truth of the null hypothesis. This probability is known as a p-value, and is a *conditional* probability:

$$p = \mathbb{P}(\text{observing this data under } H_0 \,\Big|\, H_0 \text{ is true}).$$

[In particular, the p-value is *not* the probability that H_0 is true, nor is it the probability that H_1 is false, and it is certainly not the probability that our current data could be generated by random chance.] After calculating p, we check it against some pre-chosen threshold α, and if $p < \alpha$ we reject the null hypothesis. In the social sciences, we take $\alpha := 0.05$, but this varies by discipline.

t-Tests

A t-test is a statistical comparison of 2 samples of data, done by calculating the t-statistic:

$$t := \frac{\text{difference between means of two categories}}{\text{measure of variability in categories}}.$$

There are many flavors of t-tests, which depend on what we are trying to do, but in general, a large t-statistic implies that the two samples are reliably different. From here, we can convert a t-statistic to a p-value through computer calculation.

There are two types of t-tests:

1. *Paired t-tests* are used when two samples are matched somehow (e.g. vowels produced by the same speaker, blood tests before and after a medical treatment, etc.)
2. *Unpaired t-tests* are used when the data lie in two unordered sets.
    - In addition to this, unpaired tests could be done under the assumption of *equal* or *unequal variance*: usually the first is used, unless the variability between the two sets are obviously different in some way.

Also, we can specify the *tailedness* of the test:

1. *One-tailed tests* assume that one sample mean is higher than the other.
2. *Two-tailed tests* assume that the sample means can differ in either direction.

We usually apply two-tailed tests, and there is little reason to apply a one-tailed test, except if it is clearly obvious to us that one mean might be higher than the other.

Typically, t-test results are reported in the following format:

$$t(df) = |t_0|, p = p_0,$$

where df is the degrees of freedom ($n-1$ for an unpaired test and $\frac{1}{2}n-1$ for a paired test), and t_0 is the t-statistic we found, with corresponding p-value p_0.

68 Measuring Vowel Formants

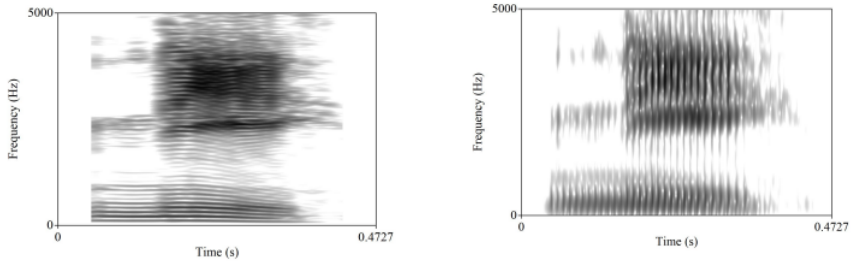
In this section, we outline useful tips for extracting vowel formant measurements in Praat.

The following are useful parameters for spectrogram visualization, which can be adjusted as needed:

1. Window length: we usually want this to be small enough to give us better time resolution.
2. Frequency range: typically, this is around 4-5 kHz.
3. Dynamic range: if this is too high, the spectrogram can appear too dark, but if it is too low, it could appear too light.

The window length dictates the type of spectrogram we are looking at:

1. A *wideband* spectrogram (with a window length of ≈ 5 ms) has clearly visible formants.
2. A *narrowband* spectrogram (with a window length of ≈ 50 ms) has clearly visible harmonics.



Here, the spectrogram on the left is a narrowband, and harmonics are the thin strips we see. The one on the right is a wideband, with clearly visible formants.

In general, we should measure formants in areas which fit some of the following criteria:

1. The steady part of the vowel — this is the best-case scenario, but this is not applicable for diphthongs.
2. At the most extreme formant values — this is especially important for F_1, and can also tell us information about jaw opening.
3. In the middle of the vowel.
4. Averaging over some interval.

We can measure formants by hand in Praat, but this is sensitive to human error and should be done as a last resort. However, hand approximations are useful in double checking automatic measurements made by Praat, which as a built-in *formant tracker*. This tracker works by measuring the vocal tract filter's effects on the sound source; however, this may be problematic if no harmonic from the sound source is close to the formant — this usually happens if f_0 is too high. Hence, formant trackers usually reconstruct the filter function based on reasonable assumptions of what the source may look likes — this technique is known as *linear predictive coding* (LPC), and works as follows:

1. We choose a time interval window, minimally one glottal cycle.
2. Predict the next value in the signal, using weighted averages of previous values — this allows us to estimate the filter function.

Usually, LPC gives us a filter function that is accurate; however, errors may still happen. We have three main problems, where we give troubleshooting techniques below.

Extra Formants

Occasionally, the formant tracker may identify the fundamental f_0 as a formant, and thus label the real F_1 as F_2, and inductively so. Alternatively, if the sound is a nasal, Praat can sometimes detect the resonance of the *nasal tract* as a formant, even if that is irrelevant to what we want.

Missing Formants

Alternatively, if two formants are too close together, as in the vowel [ɑ], Praat may mistake F_1 and F_2 as a single large F_1. Also, higher formants may be missing, simply because there is not enough resonant energy, and if f_0 is too high, the first formant can go missing.

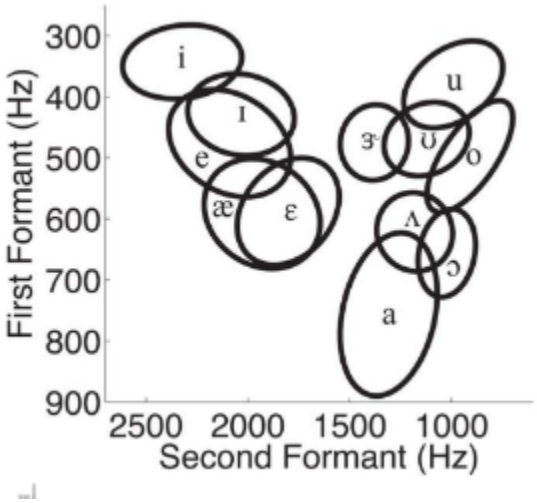
To fix this, we can either change the number of formants that Praat should be looking for, while maybe also reducing the frequency of the maximum formant. Also, if the bandwidth of the spectrogram is too wide, then Praat can easily mistake two close formants as one, so reducing the bandwidth is also an option. If this does not work, then measure the formants by hand.

Jittery Formants

This refers to unexplained, significant fluctuations in Praat's formant measurements, even if the formant looks steady on the spectrogram. The usual way to fix this is to average the measurements over an interval, or to lengthen the window.

74 Vowel Spaces

We return from our digression into statistical and technical topics to discuss *vowel spaces*, which are part of the basic phonetic description of a language. In general, a vowel space is a two-dimensional plot of $F_2 \times F_1$, with the axes flipped to align with tongue position in the mouth. For example, this is a sample vowel space for English:



These *acoustic formant plots* differ from the IPA, which aim to express *auditory vowel spaces*, which is far more general — F_1 and F_2 are not the only cues to vowel identity.

Vowel Inventories

In general, it turns out that languages like sets of vowels that satisfy some set of spacing conditions. For example, one might pose the following questions:

1. Are the vowels in a language evenly dispersed throughout its vowel space?
2. Are the vowels in a language roughly symmetrically arranged in its vowel space?
3. Do some vowels cover more of the vowel space than others?

4. Do some vowels overlap in their vowel space more than others?
5. If a vowel is "crowded," does it have less variability?

Most languages choose their vowels based on these questions, and thus, vowel spaces across languages tend to be similar:

1. Vowels prefer to be spaced apart, or *dispersed*.
2. Vowels need to be *robust*, i.e., the vowel should allow for a high variability in pronunciation without affecting vowel identity.
    - To measure robustness, we describe a vowel's *acoustic quantality*, which describes a vowel's resistance to variations in actual pronunciation. For example, the vowels [i, u, ɑ] exhibit a lot of quantality, and hence are the three most common vowels cross-lingistically.
    - We can also consider *biomechanical quantality*, which measures how lax or tense our muscles can be for producing the right sounds.

Plotting Vowel Spaces

We have seen how vowels can vary in F_1 and F_2 values. This subsection suggests methods in analyzing formant data, given the existence of this variability.

1. In general, we should plot speakers separately to look for outliers, when they exist.
2. Diphthongs should be plotted as two points, connected by an arrow.
3. It may be helpful to plot consonant contexts separately, in the case that a consonant has a major effect on the pronunciation of the vowel in question.
4. Occasionally, plotting $F_3 \times F_1$ *and* $F_2 \times F_1$ is useful: F_3 is more directly affected by rounding than F_1 and F_2.

78 Consonant Acoustics I

In this section, we give the acoustic properties of many consonants.

Fricatives

Recall that fricatives are characterized by the frication source, which is produced at the place of articulation, as well as voicing.

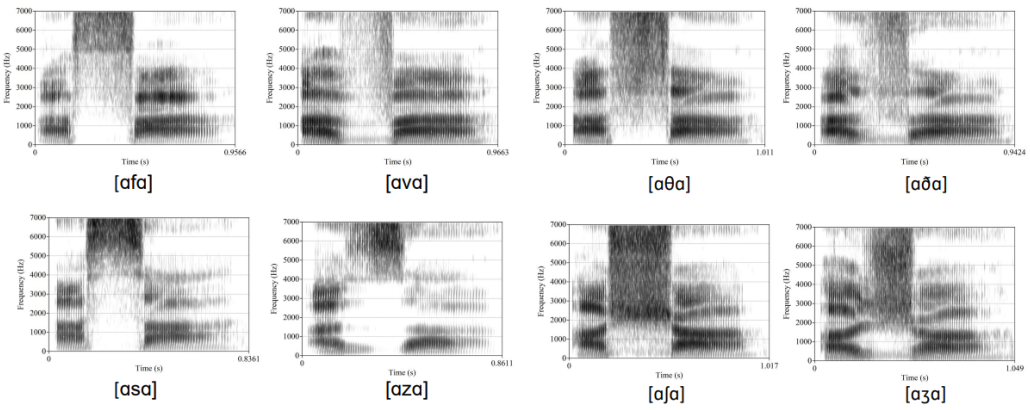
- Voiceless fricatives tend to be longer and louder than voiced fricatives, and as their name suggests, their spectrogram representations lack a voicing bar.
- Voiced fricatives **may or may not** have a voicing bar, and they can often be pronounced as sonorants instead. This is especially true if the fricative is not a strident (e.g., [s]).

The place of articulation of the fricative dictates the quality of the frication that arises. A general rule for consonants is that *the filter is always in front of the source* — this means that while the vocal tract is still the filter for the voice and aspiration sources, this is not true for frication and transience, which are produced at the place of articulation.

Thus, it follows from here that backer fricatives tend to be *lower* in frequency, due to the frication having a longer tube to travel through:

- Labial fricatives have no front cavity, so these sounds lack a filter and thus come out as unfiltered, *broadband* noise.
- Labiodental fricatives [f, v] have a tiny cavity between the teeth and the lower lip, so there might be some boost for higher frequencies, but are in general similar to labials.
- Dentals either lack a front cavity, or have a small one, and hence sound similar to labials and labiodentals — for example, compare [θ] and [f]. The vowel formants next to such a fricative may show a high F_4, and $F_2 \approx 1500$ Hz.
- Alveolar fricatives have a small front cavity, but this cavity is significant enough to cause the noise to be concentrated in high frequencies, usually above the F_4 of a neighboring vowel. These fricatives also may have multiple noise sources, so these tend to be the loudest fricatives. Vowels near an alveolar fricative could have higher F_2's and F_3's.
- Palatoalveolar fricatives have a larger front cavity, and possibly space under the tongue, so the frequencies tend to be concentrated in the middle, near the F_3 of a neighboring vowel.

A chart of sample spectrograms is given below.



Usually, we will quantify fricatives using *center of gravity*, which is the mean frequency of the fricative, weighted by the amplitude. It thus follows that alveolar fricatives have high centers of gravity.

Sonorants

In contrast to fricatives and stops, sonorants are vocalic in nature, so we can use formants directly to classify them.

- Nasals tend to be very obviously distinct from their neighbors, as they seem more stop-like.

- Liquids are somewhat vowel-like, in a way we will quantify shortly.
- Glides are the most vowel like.

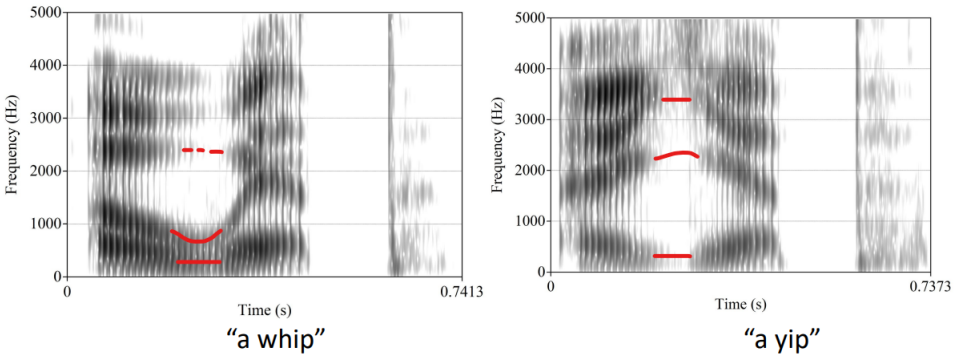
However, all three types of sonorants typically have less acoustic energy than vowels.

Glides

In general, glides match high vowels in formants:

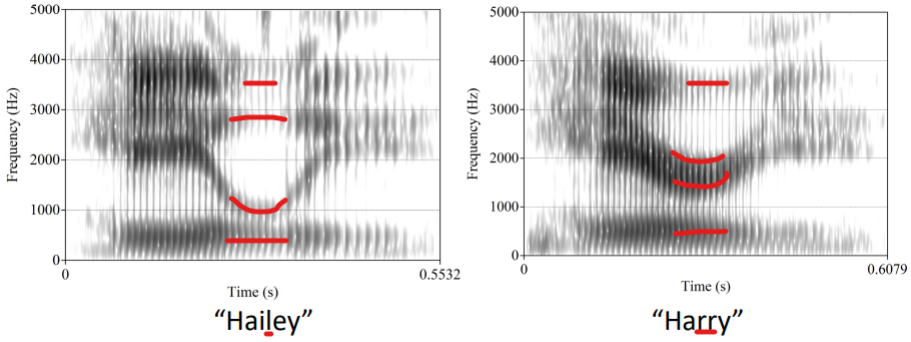
- Like [u], the glide [**w**] has a low F_1, a low F_2, and a low-mid weak F_3.
- Like [i], the glide [**j**] has a low F_1, a high F_2, and a high strong F_3.

In the chart below, the first three formants of the consonant have been marked.



Liquids

Liquids generally have a fairly low F_2, and their F_3 values are *extreme*: [ɹ] has a very low F_3, while [l] has a very high F_3. In the chart below, the first four formants of the consonant are marked.



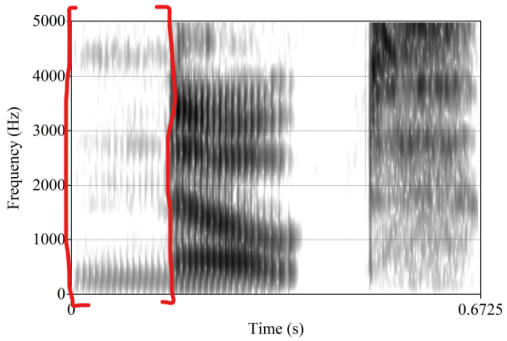
In general, our approximants satisfy the following:

	Lower F_3	Higher F_3
Higher F_2	[ɹ]	[j]
Lower F_2	[w]	[l]

80 Consonant Acoustics II

Nasals

As we have mentioned in the previous section, nasals have formants, but they are weak and often have gaps. Despite this, it is easy to identify them, but the main difficulty lies in identifying their place of articulation, though *formant transitions* into adjacent vowels can be helpful.



"Note"

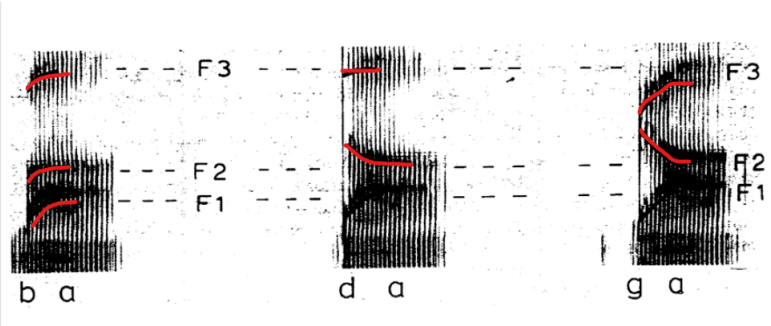
Stops

Stops are classified by their place of articulation, their release, and voicing.

- Voiceless stops consist of a silent, long closure, followed by a strong release with possible aspiration.
- Voiced stops **may or may not** have a voicing bar depending on VOT, have a weaker release, and usually have no aspiration.

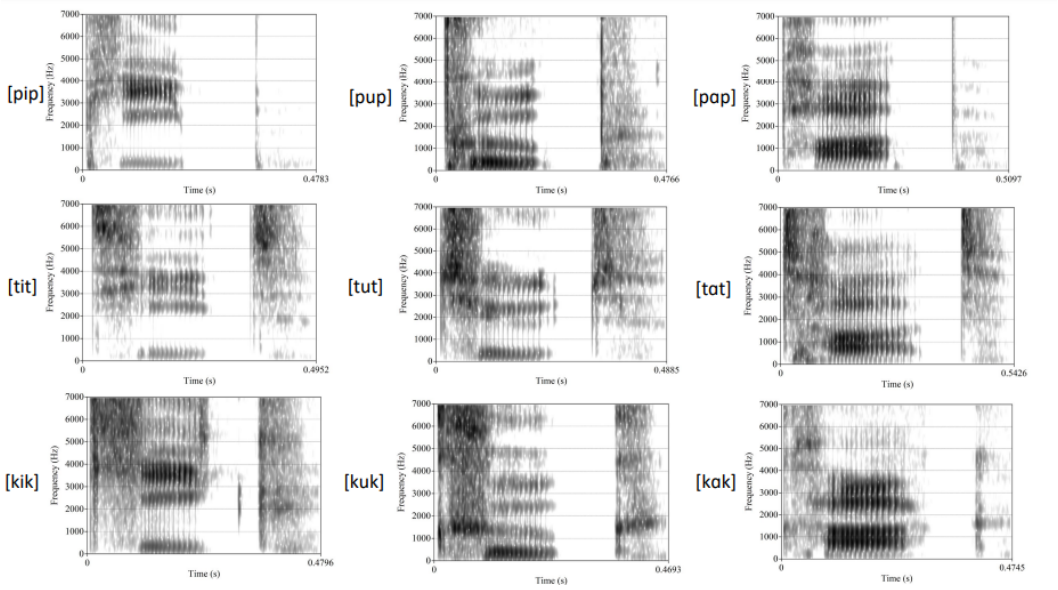
The place of articulation of a stop (both oral and nasal) is usually most apparent in formant transitions to and from adjacent vowels.

- Labials: all formants in the next vowel start low, and then rise, and vowels before a labial have falling formants.
- Alveolars: F_2 and F_3 in the next vowel start high.
- Velars: F_2 and F_3 start similar, then diverge. This is known as *velar pinch*.



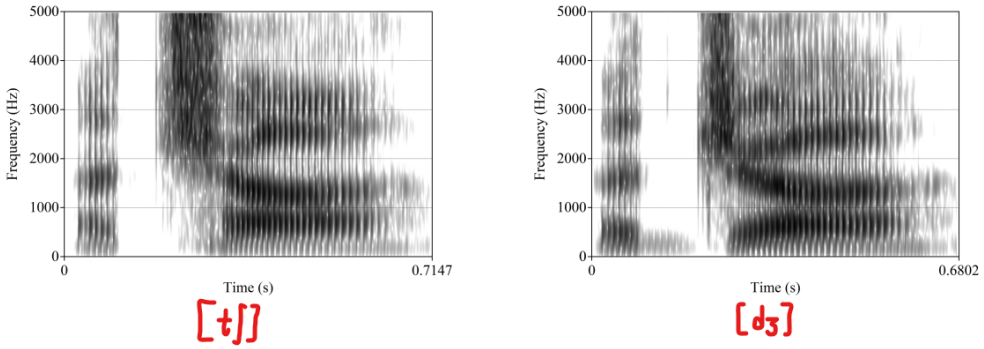
Stop bursts, like fricatives, vary with the cavity size caused by the place of articulation, and thus can be informative as well.

- Labial stops have a broadband, but weak burst, due to the lack of a filter.
- Alveolar stops have bursts that are strong in high frequencies.
- Velars have stops that are "compact" at F_2 and F_3 vowel frequencies, and often can consist of double- or triple-bursts due to the slow action of the tongue back.



Affricates

Affricates have the properties of both stops and fricative, but the frication noise part is generally shorter than those of fricatives. Additionally, frication is rarely actually voiced in voiced affricates. An example is given below.



Non-pulmonic Consonants

Up until this point in the course, we have assumed a pulmonic egressive airstream, where the volume of the lungs decreases and pushes air in the vocal tract, thus producing sound. However, several other airstream mechanisms are used in human language and can basically be placed into two main categories: *glottalic* and *lingual* airstreams.

Ejectives

Stops and fricatives made using a *glottalic egressive* airstream are known as *ejectives*: [k', t', b', ...]. These are produced as follows:

1. Close the glottis and form a closure in the vocal tract corresponding to the place of articulation.
2. Raise the glottis to build pressure.
3. Release the oral closure.
4. Release the glottal closure.

Ejectives are characterized by their loud burst, and a long silence between the burst and a succeeding vowel.

Implosives

Stops and fricatives made using a *glottalic ingressive* airstream are known as *implosives*: these are similar to ejectives, except that where we raise the glottis, we lower it instead. Because the glottis in this case is not completely closed, these are *voiced* sounds, and the amplitude of voicing during the closure itself *increases*. Other linguistic correlates of implosives vary cross-linguistically, and are being actively studied.

Clicks

Clicks are sounds produced using a *lingual airstream*. Because clicks are not similar to many of the sounds we have seen before, they are transcribed in the IPA by special symbols:

[ʘ ǀ ǃ ǂ ǁ].

The steps in producing a click are as follows:

1. Form a velar closure, and one further forward in the vocal tract.
2. Lower the tongue body to decrease air pressure.
3. Release the front closure.
4. Release the velar closure.

Similar to plosives, clicks could be voiced, voiceless, nasalizes, or aspirated, and IPA indicates these distinctions with preceding velar symbols if necessary.

84 Phonation Type

Throughout a majority of this course, we have assumed that sounds are voiced and voiceless. However, (modal) voicing is only one quality of *phonation* (sound production by the larynx) that is possible. We know that a pitch of someone's voice is determined by the rate of vibration of the vocal folds; in this section, we consider *how* the vocal folds vibrate.

Voice quality refers to the way the vocal folds vibrate — we have already discussed two types of voice quality: *voiceless* and *modal voicing*, the latter which refers to regular, periodic vibration. These two types of voicing occur in all languages

However, *non-modal voicing* exists as well, and are can be contrastive within a language:

- *Breathy voice* occurs when more air is allowed through the glottis when speaking, but the vocal folds are still adducted enough to produce voicing. The vocal folds are thus *open* for most, if not all, of the glottal cycle, and the additional airflow generates noise. Breathy voice cna be phonemically contrasted on stops and vowels, and is usually indicated in the IPA by two dots on the bottom of the symbol: [a̤, b̤].

- *Creaky voice* occurs when the glottis is constricted more than usual during voicing, and thus only partial, abrupt, and irregular vibration occurs. The glottis spends less time being open in this case, and similarly to breathy voice, both consonants and vowels can contrast for creakiness, indicated in IPA as follows: [a̰, b̰].

We may view breathy and creaky voicings as extreme versions of modal voicing, which mainly differ by the amount of time the vocal folds stay open for in the glottal cycle. This measurement is called the *open quotient*, and we get the following relationships between the open quotients of the three types of voicing we see:

$$\text{creaky} < \text{modal} < \text{breathy}. \tag{5}$$

We also note that difference phonation types change the *source* directly, and not the filter. In particular, since f_0 is given by the rate of vibration of the vocal tract, and not the method of vibration, f_0, as well as the harmonics of f_0, are constant in frequency across the three types of voicing. However, the *amplitudes* of the harmonics are independent of f_0, so they can vary. It is these differences between harmonic amplitudes that are perceived as changes in voice quality.

If we denote by H_n the *amplitude* of the nth harmonic, we can compare the measurement of $H_1 - H_2$ between voicing types:

$$\begin{array}{ccc} H_1 - H_2 < 0 & H_1 - H_2 \geq 0 & H_1 - H_2 \gg 0 \\ \text{creaky voice} & \text{modal voicing} & \text{breathy voice} \end{array} \tag{6}$$

Alternatively, we know that the amplitudes of the harmonics decrease as the frequency increases. Thus, we can mesure the rate at which the amplitudes decrease, known as the *spectral tilt*: modal speech has a spectral tilt of 12 dB per octave; creaky voice shows a slower decline, while breathy voice shows a faster decline.

There are other sorts of measurements to consider as well: if we denote by A_n the amplitude of the frequency closest to F_n, then $H_1 - A_3$ tends to be larger for breathy voice and smaller for creaky voice. It follows from $H_1 - H_2$ and $H_1 - A_3$ that *breathy voice is strong in low frequencies*, and *creaky voice is strong in high frequencies.*

Relationship with Formants

We know that phonation type is a property of the source, and not the filter; however, all sound needs to go through the filter in order to exit the vocal tract. This again causes an issue, as formants change harmonic amplitudes, so once again, we find ourselves needing to invert the effects of the filter function. This is doable, and we can find *formant-corrected harmonic amplitudes* via hand calculation and the right formulas. These are notated with an asterisk: $H_1^*, A_5^*, \ldots$.

Alternatively, we can instead control for formant effects by using similar vowels, so that the effect of the formants stay similar across experiments, or we can examine $H_1 - H_2$ only for low vowels, as low vowels have a high F_1 and thus has less likelihood to interfere with H_1 or H_2.

88 Tone and Intonation I

We continue our study of suprasegmentals by discussing *pitch*, which is a psychological percept based on the fundamental frequency of a sound. Hence, pitch is not identical to f_0 for reasons we shall see later, for our purposes in this section, this is close enough.

Recall that f_0 is the rate of vibration of the vocal folds, which depends on three factors: the mass of the folds, the stiffness of the folds (stiffer is faster), and the amount of air that flows through the glottis. Roughly, with these things considered, the vocal folds vibrate at 80 to 250 Hz, which is faster than any voluntary motion we can do. The reason for this is due to physics, which is beyond the scope of this course.

Pitch Tracking in Praat

Computers are able to track pitch through *autocorrelation analysis*, which takes copies of the waveform and attempts to place them on top of each other, such that their overlays are highly correlated. [Another type of analysis, called *cross-correlation analysis*, works in more precise contexts, such as at the level of a glottal period, but is not too applicable for intonation contours.] This explains the errors that can occur during pitch tracking, such as pitch halving or pitch doubling (where autocorrelation "misreads" one cycle for two, and vice versa). Occasionally, creaky voice could cause the pitch tracker to be unable to find a period. To fix this, we should adjust the frequency range which the computer to search in.

Variation in f_0

We know that people have voices of different pitches, and this is generally caused by the size of someone's vocal folds. All things considered, a person usually has a *range* of comfortable f_0's spanning 100 Hz. This is largely achieved by tensing the vocal folds. Interestingly, entire languages may have typical, characteristic ranges of f_0, and bilingual speakers often reveal this to us: for example, a study reveals that English-Japanese bilinguals speak Japanese at a higher f_0 than in English.

However, the f_0 variation we are mainly interested in is *phonemic*, where pitch of a syllable is contrastive in meaning. In general, there are two types of tones: *level tones* (with roughly constant f_0) and *contour tones* (changing f_0). The actual f_0 that a speaker uses in a tonal language is not particularly important, due to individual speakers' naturally differing f_0's, but speakers of tonal languages discern the change in f_0.

90 Tone and Intonation II

Despite f_0 not being used to distinguish between different tones, examining f_0 can give us a more exact characterization of tones, and we can also see allophonic variation of tones across different contexts.

Intrinsic f_0

As a result of glottal and oral configurations, consonants and vowels have their own intrinsic, or typical, f_0 values:

- High vowels have a higher f_0 than low vowels, and there is generally a strong correlation between f_0 and F_1.
- Voiced obstruents also have a lower f_0 than voiceless obstruents or sonorants: the build-up of air pressure behind a consonant constriction during voicing results in less airflow, and less airflow gives a lower f_0. Conversely, glottal tensing for voicelessness raises f_0 for the vowel that follows.

These intrinsic f_0's result in allophonic variation of tones, which adapt to different segmental contexts.

Tonogenesis

Tonogenesis refers to the appearance of new contrastive tones, and usually comes in two forms:

1. *Tone splits* occur when allophonic variations of tones split into two phonemically constrastive tones.
2. *Tone mergers* occur when two separate tones merge into one.

Occasionally, non-tonal languages develop tones through the following process:

1. Two sounds are distinguished by some other acoustic cue other than tone.
2. A redundant tone distinction emerges for some reason, usually not linguistically related.
3. The tone distinction becomes larger, but remains redundant.
4. The original distinction begins to disappear, but the tone distinction remains.
5. The two sounds that were once distinguished by some other cue are now distinguished by tone.

We see this in two modern languages:

- Punjabi once had a [d^h]/[t] distinction, but now /d^h/ is realized as [t] word-initially and [d] elsewhere, *but* with a falling pitch contour.
- Korean has three types of voiceless stops, which are aspirated, fortis, and lenis, but the VOT for aspirated stops became shorter, while f_0 differences are currently increasing.

Utterance-level f_0

In a broader view, pitch is used across an entire utterance to signal linguistic information. For example, question phrases often have a pitch rise at the end of the utterance, but in general, languages approach this differently.

One universal property of utterance-level f_0 is *declination*, where the f_0 decreases over the utterance as airflow lessens. However, this is subconsciously factored into our perception of pitch.

Other than natural declination, we use *intonation* to mark certain words or f_0 targets as prominent (*head-marking tones*), and changes in f_0 mark off certain phrases (*boundary tones*). These are modeled using the *autosegmental metrical* (AM) *model*, which breaks a pitch contour of an utterance into a series of discrete pitch targets that align with prominent words and prosadic boundaries, and interpolates everything else.

In transcription, we use the *tones and break indices* (ToBI) framework, where head-marking tones are marked with an asterisk: $H^*, L^*, \ldots$, boundary tones are marked $H\%, L\%$, ..., and numbers mark break strength. Note that ToBI is not standardized across languages.

Stress

In languages that use stress, such as English, f_0 is often correlated wht stress, though other acoustic cues may exist, like loudness, duration, and changes to pronunciation. Which cues are used varies between languages: for example, English stress is longer and louder, while stressless syllables often have heavily reduced vowels. Allophony may also be sensitive to stress, so figuring out a language's stress patterns can help in understanding the relationship between segments, and vice versa.

94 The Auditory System

We have focused on speech acoustics for a majority of this course, so in the last few sections, we will discuss speech perception. We begin with an overview of the auditory system, which does the job of transforming physical vibrations in air to electrochemical signals which the brain uses. Because of the way the auditory system does things, non-linearities are introduced in this process.

Outer Ear

The outer ear consists of the *pinna* (external part of the ear) and the *ear canal* (the tube leading into the ear). The purpose of the outer ear is to bring in sound pressure waves, and also to protect the eardrum. At this point in the auditory process, we are still working with air vibrations, but the ear canal is a tube that contains resonances. These resonances boost sounds between 2 and 5 kHz by around 15 dB, thus introducing our first non-linearity.

Middle Ear

The middle ear contains the *eardrum*, the *ossicles*, and the *oval window*. These structures vibrate mechanically due to the waves coming in. Going from the large eardrum to the small ossicles, sounds especially around 0.8 to 2.5 kHz is boosted by around 30 dB. This is seemingly a problem, because 30 dB is quite significant — it should be the case that the

sound of our own voice destroys our ears after years of speaking. However, the muscles of the ossicles tense before we speak to protect against this.

Inner Ear

The inner ear contains multiple vital structures needed for hearing and audition. The most important of these is the *cochlea*, a snail-shaped structure containing tubes consisting of vibrating fluid (*hydro-dynamic energy*). The middle of the cochlea contains the *basiliar membrane* (BM), which essentially contains a biological Fourier transform — frequencies, from high to low, are mapped along parts of the BM: its base is narrower and responds to higher frequencies, and it grows wider within, thus giving us access to lower frequencies. Because of the structure of the BM, our perception of pitch is logarithmic, as more space is devoted to lower frequencies than higher ones: $\frac{3}{5}$ of the length of the BM is for frequencies under 4000 Hz.

Finally, after the sound travels into the cochlea, the *Organ of Corti*, which sits on the BM, contains hair cells which pick up vibrations from the BM to synapse with auditory nerve fibers, which gives electrochemical signals that the brain interprets. These hair cells are crucial to allowing hearing, and strong movement in the BM (through loud sounds) can kill these hair cells, which cannot regrow. For deaf people, *cochlear implants* exist, which bypass the cochlea to send electrical signals to the auditory nerve.

98 Speech Perception I

Perceptual Frequency Scales

Recall that the basilar membrane devotes more area to lower frequency sounds, and thus our pitch resolution is better in lower ranges than in higher ranges. This means that an objective measurement like absolute frequency (in Hertz) is likely not a good representation of our subjective perceptions of pitch. To account for this, we use one of three perceptual scales for pitch:

- The *Mel scale* measures subjective pitch, by defining a doubleing in mels to be a doubling in perceived pitch on a sine tone. This does not necessary correspond to a doubling in frequency.
- The *Bark scale* considers the basilar membrane as a filter band of 20 so-called "filter bands." This is similar to the mels, but it is more linear at lower frequencies.
- The *ERB scale* is similar to the above, and can seen as an update to the two.

Loudness

Similar to pitch (and thankfully), our perception of loudness is also non-linear. Thus, we use the *decibel* (dB) *scale*, which is logarithmic and compares the power values of different amplitudes. Taking the amplitude to be the pressure value at the peak of a waveform, we take a window, say some time interval I of amplitude values and compute its root mean squared amplitude:

$$x := \sqrt{\frac{1}{|I|}\sum_{k\in I} k^2}. \tag{7}$$

From this, we compute the decibel value basd on some reference amplitude r:

$$\text{dB} := 10 \log_{10} \left(\frac{x^2}{r^2} \right) = 20 \log_{10} \frac{x}{r}. \tag{8}$$

Relative to sound pressure, the computation in decibels match judgments of relative loudness better than using absolute amplitude. Typically, we use "dB SPL," which uses "just perceptible sound" as our reference level r.

Thresholds

Loudness perception also depends on the specific sound. Some sounds are said to have a *high threshold*, in that it must be objectively louder in order for a person to perceive it, and different frequencies have different thresholds. Thresholds for low frequencies are fairly consistent with age, but it is well-known that high-frequency thresholds increase by 10 dB every decade, past the age of 30.

Correlates and Cues

Now, we have sufficient background to discuss speech perception. Speech perception refers to how listeners use an acoustic signal to recover linguistic information, and are based off of acoustic *correlates* and *cues*:

- Correlates are properties observed in the acoustic signal that are correlated with a linguistic feature or sound.
- Similarly, cues are correlates that are actively used by listeners to recognize a linguistic feature or sounds.

Interestingly, it follows that nearly every correlate that exists is a cue, though its usefulness is highly language-dependent. However, a correlate can only become a reasonable cue if it is *perceptible*, which contains three components:

1. *Detection*: the correlate needs to be able to be heard by a person in the first place.
2. *Discrimination*: the correlate needs to be distinguished from another one.
3. *Identification*: the listener needs to be able to tell exactly what the distinguishing feature of the correlate is.

We also note that cues may or may not be acoustic: when communicating, our eyes bring us visual information that can help discern what a person is saying.

We have already discussed detection of signals in the previous part of this section, and identification was the majority of the course. We now examine detection more closely.

Just-Noticeable Differences

A *just-noticeable difference* (JND) is the smallest detectable difference in a signal. This can refer to differences in frequency, loudness, duration, and more. It follows that a correlate can only pass as a cue if it can be discriminated from another correlate past some JND. These are somewhat influenced by language experience, but in general a few rules of thumb exist:

- The JND for duration is roughly 30 ms. This implies that a difference of VOT less than 30 ms will most likely remain undetected.
- Frequency distinctions in lower ranges (below 5 kHz) are more easily perceived.

100 Speech Perception II

Acoustic cues can be studied or identified in the following ways:

1. *Partial stimuli*: if listeners can recognize a sound from a subset of the frequency range for example, that range likely contains a cue.
2. *Natural acoustic variation* can be correlated with perceptive variation.
3. *Speech synthesis* techniques can keep correlates fixed, and thus compare the relative strength of two opposing correlates.
4. *Speech editing* of recordings is similar to above.

Categorical Perception

Our discrimination of sounds is constrained by categories — sounds in different categories are easier to distinguish. This is stronger for consonants than for vowels or tones, which is what we should expect.

- Interestingly, we also do categorical perception on non-speech sounds, but clearly, we are not getting any linguistic information here.
- Hence, part of our categorical perception is probabilistic, as we also filter out noise when we perceive speech sounds. This makes us rely on our expectations — for example, if we expect the next vowel in an utterance to be an [i], we will be looking for a low F_1 and a high F_2.
  - We also *shift* our perceptions of boundary values to "repair" the speech signal if necessary: for example, if we expected the vowel [i] but got an unusually low F_2, we are more likely to shift that percept up a bit in order to match our expectations.
  - Lexical expectations also modify our perceptions: we tend to prioritize real words over fake ones, for obvious reasons.

Perceptual Similarity

Sounds in a language may be similar, and thus confusable. To quantify this, we use *confusion matrices*, where we play syllables (with some noise), and ask listeners to identify what they heard. This allows us to create the confusion matrix P, where the main diagonal records correct responses, and the entries off the diagonals, say P_{ij}, records the number of times sound i was mistaken for j. This allows us to calculate a similarity matrix S, by defining

$$S_{ij} := \frac{p_{ij} + p_{ji}}{p_{ii} + p_{jj}}. \tag{9}$$

It is easily verified that S is symmetric, so that we may refer to **the** similarity between i and j. From here, we take $D_{ij} := -\ln(S_{ij})$ (for perceptual reasons) and plot these values of D using *multi-dimensional scaling*, which takes a set of pairwise dissimilarities and plots out those points in relation to the others. From here, it is easy to reveal *perceptual warping*, which is related to language acquisition.

Cues and Sound Inventories

The best speech sounds for languages have multiple, salient cues, as they are robust against environmental conditions; similarly, the best speech sound contrasts ones that differ based on multiple easy-to-discriminate cues. This shapes a language's phonotactics, as segments are more likely to appear in positions or combinations where their cues can be perceived.

As an example, stop clusters are difficult to distinguish because their stop cues include the closure, burst, and a formant transition, which are very sensitive to overlapping segments of this type. For example, the string /pt/, as in "pterydon," a Greek borrowing, is often simplified to [t] at the beginning of words in English, as the [p] is phonotactically disadvantaged in English. However, the string /pt/ does not become simplified in words such as "helicopter," where /pt/ is word-medial. Here, the [p] is not phonotactically disadvantaged, as it occurs as the coda of a syllable, while the [t] is the onset of the next syllable.

104 Coarticulation, Instrumental Phonetics

Coarticulation

Speech sounds often have many overlapping cues, as boundaries between segments need not be clearly distinguished. Hence, at any moment, the state of the vocal tract reflects more than one segment — this makes sense, as segments are influenced by their neighbors, and thus rarely appear in their "canonical" forms. This is known as *coarticulation*, and is usually an advantage because it spreads out the cues, and thus makes them easier to detect. For example, vowel rounding can also lower preceding fricatives when the rounding "bleeds" into the fricative:

- The [s] in [su] is typically lower in frequency than that in [sɑ].

The only problem with this, of course, is that listeners need to map cues onto their intended segments, and they need to be able to undo the effects of coarticulation — in the example above, the [s] in [su] is typically closer to [ʃ]. Are listeners able to *perceptually compare* [s] and [ʃ]? To test this, we use speech synthesis to make two acoustic continua of fricatives: one from [sa] to [ʃɑ], and another from [su] to [ʃu]. If indeed the listeners are able to adjust for coarticulation, the decision boundary between the two continua should be different, and this is in fact the case.

Another similar example occurs in nasal vowels, which have an extra nasal resonance in the F_1 range. This has the effect of making nasal high vowels sound lower, while nasal low vowels rise. Nasal effects from stops [m, n] occasionally bleed into a neighboring vowel, but listeners can tell this apart fairly easily as well.

However, listeners can be occasionally "fooled" by coarticulation, and thus perceive segments incorrectly. This is a fairly common catalyst for language change, and occurs in two forms:

1. *Hypocorrection* occurs when listeners do not correct for coarticulation, thus perceiving cues for the wrong segment.

2. *Hypercorrection* occurs when listeners overcompensate for coarticulation, thus "undoing" what was not intended as coarticulation.

Instrumental Phonetics

Thus far, we have been focused on acoustic measurements of speech; however, there are other measurements we can make:

1. Articulator position: where do body parts move to when producing a speech sound?

2. Airflow: how much air, or how does air, move through the vocal tract in speech?

3. Muscle activation: which muscles are used in producing a certain sound?

4. Brain imaging: which parts of the brain are used to produce or perceive a sound?

We often have a wide range of measuring instruments we can use, but a consistent theme consists of these four questions:

1. How expensive is this technique?

2. How unpleasant or dangerous is this for the person being measured?

3. What is the trade-off between temporal and spatial resolution?

4. What are we trying to do in the first place?

There are also challenges to consider when making articulatory measurements:

1. The vocal tract contains hard and soft structures, which move differently, and can deform.

2. Many structures in the vocal tract are hard to access (e.g. larynx, velum), and access might disrupt the speech that we are trying to study.

3. Structures in the vocal tract move at different speeds.

4. Parts of the vocal tract can be coupled together.

5. Instruments are hard to affix inside the vocal tract, due to its moist, warm nature.

Because of these challenges, we usually use audio instead, but occasionally getting through these challenges is fruitful. Now, we give a list of possible techniques, which fall into two categories: *direct* (the instruments contact the structure of interest) and *indirect* (the instruments are remote).

1. *X-ray imaging* generates a saggital projection of the vocal tract. However, this image is not necessary *mid-saggital*, and thus can be hard to read. Additionally, this exposes the subject to radiation, which is bad.

2. *Tomographic imaging* projects a thin, flat beam through tissue along a plane, and *computed tomography* (CT) uses x-rays to image slices. This is an improvement from the above, but still exposes the subject to radiation.

3. *Magnetic resonance imaging* (MRI) is already fairly well-known. However, for the purposes of speech, MRI is too slow, so an alternative called *real-time MRI* is used instead. This provides good temporal and spatial resolution, but is extremely expensive.

4. *Ultrasound* produces images based on the reflective properties of sound waves, and gives good temporal resolution, while being non-invasive and inexpensive. The trade-off is poorer spatial resolution.

5. *Electromagnetic articulography* (EMA) places subjects into a plastic assembly, which generates an alternating magnetic field which interacts with pellets affixed to articulators. This is a relatively non-invasive technique, which gives very high temporal and spatial resolution; however, the experience in affixing pellets is time-consuming and often unpleasant, and it only allows imaging of specific points.

6. *Palatography* measures the contact between the tongue and the palate. There are two forms: *static* (which involves smearing charcoal over the tongue), or *electropalatography*, which uses a custom prosthetic.

7. *Electroglottography* (EGG) measures the contact between vocal folds by affixing two electrodes on opposite sides of the larynx, from which a small current is passed between the two.

8. *Airflow measurements* are usually done by a specially-designed mask, which filters oral and nasal airflow. Alternative techniques include using airtype chambers, putting tubes in the subject's mouth (which often affects speech), or putting a needle in the trachea to measure sub-glottal pressure (very dangerous).

9. *Electromyography* (EMG) measures muscle activation by using surface electrodes, or thin needle electrodes into muscle fibers. This gives us a sense of muscle activiation, but it can disrupt speech (especially if the needle form is used).

108 Speech Technology

We end with some applications of phonetics into computationally-based fields. *Speech technology* refers to computation that operate on spoken languages, and fall into two categories.

1. *Automated speech recognition* (ASR) takes speech and converts it to text, for the computer to use. Implementing an ASR system requires a computer to understand speech and spot words, which is difficult, so a large corpus of annotated speech is used to computationally train the computer to do this. Sample applications include computer interface interaction, or automated closed captioning.

2. *Speech synthesis*, or *text-to-speech* (TTS), should be self-explanatory. A problem we run into is *text normalization*, where different strings of letters can be read in different ways. For example, the string "1750" could either be a year (seventeen-fifty), a number (one thousand seven hundred fifty), or a string of digits (one seven five zero). Types of synthesis include rule-based, parameter-based, or concatenation-based, which assembles speech into units based on diphone midpoints (transitions between) segments. Sample applications include customer service numbers, audio books, document readers, and the like.

References

Images were taken from the lecture slides, which were written by the professor. Some images may originally be borrowed from other texts.