

# LIN 350: Experimental Phonetics

Scott Nelson





# Contents

<b>Overview of the course</b>	<b>7</b>
<b>I Review</b>	<b>11</b>
<b>1 Review of Articulatory Phonetics</b>	<b>13</b>
1.1 The International Phonetic Alphabet . . . . .	13
1.2 Articulatory Description . . . . .	15
1.2.1 Place of Articulation . . . . .	16
1.2.2 Manner of Articulation . . . . .	19
1.2.3 Voicing . . . . .	22
1.2.4 Putting it all Together for Consonants . . . . .	25
1.3 Articulation and Description of Vowels . . . . .	26
<b>2 Review of Acoustic Phonetics</b>	<b>27</b>
2.1 Spectrograms . . . . .	27
2.2 Resonance . . . . .	29
2.3 Silence . . . . .	34
2.4 Noise . . . . .	36
<b>II Main Lessons</b>	<b>39</b>
<b>3 f<sub>0</sub>-pitch-tone-intonation</b>	<b>41</b>
3.1 Sine Waves . . . . .	41
3.2 Fundamental Frequency in Speech . . . . .	44
3.3 Intonation . . . . .	45

<b>4</b>	<b>Lexical Tone + Contrast</b>	<b>51</b>
4.1	Contrast . . . . .	51
4.2	Lexical Tone . . . . .	54
<b>5</b>	<b>The Acoustic Theory of Speech Production</b>	<b>59</b>
5.1	Filters and Resonators . . . . .	59
5.2	Tube Models . . . . .	61
5.3	Vowels as multiple tubes . . . . .	63
<b>6</b>	<b>Acoustic Correlates of Stress</b>	<b>69</b>
<b>7</b>	<b>Voicing</b>	<b>71</b>
7.1	Voice Onset Time . . . . .	71
7.2	Other cues for Voicing . . . . .	76
<b>8</b>	<b>Acoustic Correlates of Place of Articulation</b>	<b>79</b>
8.1	Formant Transitions & Stop Bursts . . . . .	79
8.2	Spectral Moments in Fricatives . . . . .	84
<b>9</b>	<b>Introduction to Speech Perception</b>	<b>89</b>
9.1	Categorical Perception . . . . .	90
9.2	Perceptual Illusions . . . . .	90
9.3	Normalization . . . . .	90
<b>10</b>	<b>Theories of Speech Perception</b>	<b>95</b>
10.1	Objects of Perception . . . . .	95
10.2	Articulation of English /ɪ/ . . . . .	96
10.3	Perceptual Confusion . . . . .	97
<b>III</b>	<b>Labs</b>	<b>99</b>
<b>11</b>	<b>Lab 0</b>	<b>101</b>
<b>12</b>	<b>Lab 1</b>	<b>103</b>
<b>13</b>	<b>Lab 2</b>	<b>105</b>
<b>14</b>	<b>Lab 3</b>	<b>107</b>

<i>CONTENTS</i>	5
<b>15 Lab 4</b>	<b>111</b>
<b>16 Lab 5</b>	<b>113</b>
<b>17 Lab 6</b>	<b>115</b>
<b>18 Lab 7</b>	<b>117</b>
<b>19 Lab 8</b>	<b>119</b>
<b>Appendices</b>	<b>123</b>
<b>Appendix A Introduction to Praat</b>	<b>125</b>
A.1 What is Praat? . . . . .	125
A.2 Opening Praat/Sound Files . . . . .	126
A.3 Text Grids . . . . .	128
A.4 Measuring Things . . . . .	130
A.5 Other Resources . . . . .	132
<b>Appendix B Syllabus</b>	<b>135</b>
<b>Appendix C Example Lab Report</b>	<b>141</b>



# Overview of the course

These notes are from the Summer 2023 version of LIN 350: Experimental Phonetics at Stony Brook University. The official course description for LIN 350 is the following.

Introduction to common experimental methods for studying the sounds used in human language. Topics include basic speech acoustics, acoustic analysis, oral and nasal airflow, static palatography, linguography and electroglottography, as well as design of perception experiments. Students will learn the physical processes affecting each experimental variable and common methods of analyzing each kind of data. Students will get hands-on experience with each analysis method and will use two or more types of data to explore a hypothesis about sound structure in English or some other language of interest. Students will learn how to use software for making measurements and analyzing data. Students will learn to assess the validity of claims about language based on their understanding of the scientific method as applied to speech. The course will give students a solid foundation for further courses in laboratory skills relevant to assessment of normal and disordered speech and for pursuing research, either as undergraduate researchers, or in the early stages of graduate work.

In this version of the course we will primarily focus on acoustic and auditory analysis of speech. Articulatory analysis will be limited to discussion only. This is not because articulatory analysis is less important, but rather because it requires more resources and specialized equipment. On the other hand, acoustic and auditory analysis can be done (at least to a first approximation) with only a laptop and preferably some headphones.

The goal of this course is to teach students how to do their own analyses and the best way to learn is by doing. Therefore, this course is based around labs that will simultaneously guide students through an analysis, while also reinforcing the physical correlates of linguistic structure in the sound domain.

A schedule of topics and labs is listed on the following page. To conclude the course, each student will do an independent phonetic analysis on a topic of their choosing. The phonetic analysis must involve a comparison of some type between two speech communities. This can include speakers of different languages, speakers of the same language but different dialects, or disordered and non-disordered speakers. Each student will give a 15-20 minute presentation on their project as well as write a short paper explaining their findings. More details on the structure of the course can be found on the course syllabus.

## Class Schedule

Week	Day	Time	Topic
1	7/11 (T)	Asynchronous	Lecture - review of Articulatory Phonetics
		Asynchronous	Lecture - review of Articulatory Phonetics
		Asynchronous	Lecture - review of basic Acoustic Phonetics
	7/13 (Th)	Asynchronous	Lecture - review of basic Acoustic Phonetics
		Asynchronous	Praat - Introduction
		Asynchronous	<b>Lab 0: Analyzing Speech Data</b>
2	7/18 (T)	9:30–10:30	Lecture - f <sub>0</sub> , pitch, tone, intonation
		10:40–11:40	Praat - measuring f <sub>0</sub>
		11:50–12:55	<b>Lab 1: Intonation</b>
	7/20 (Th)	9:30–10:30	Lecture - lexical tone + contrast
		10:40–11:40	Praat - automatic analysis + drawing
		11:50–12:55	<b>Lab 2: Tone</b>
3	7/25 (T)	9:30–10:30	Lecture - the acoustic theory of speech production
		10:40–11:40	Praat - measuring formants
		11:50–12:55	<b>Lab 3: Formants</b>
	7/27 (Th)	9:30–10:30	Lecture - acoustic correlates of stress
		10:40–11:40	Praat - measuring intensity + duration
		11:50–12:55	<b>Lab 4: Stress</b>
4	8/1 (T)	9:30–10:30	Lecture - voice onset time
		10:40–11:40	Praat - point tiers + measuring vot
		11:50–12:55	<b>Lab 5: Voicing</b>
	8/3 (Th)	9:30–10:30	Lecture - acoustic correlates of place of articulation
		10:40–11:40	Praat - spectral measurements
		11:50–12:55	<b>Lab 6: Place of Articulation</b>
5	8/8 (T)	9:30–10:30	Lecture - intro to speech perception
		10:40–11:40	Praat - running experiments
		11:50–12:55	<b>Lab 7: Categorical Perception</b>
	8/10 (Th)	9:30–10:30	Lecture - theories of speech perception
		10:40–11:40	Praat - ???
		11:50–12:55	<b>Lab 8: Perceptual Similarity</b>
6	8/15 (T)	9:30–10:30	Lecture - more experimental techniques
		10:40–11:40	Project Presentations
		11:50–12:55	Project Presentations
	8/17 (Th)	9:30–10:30	Project Presentations
		10:40–11:40	Project Presentations
		11:50–12:55	Lecture - course recap





# **Part I**

## **Review**



# Lesson 1

## Review of Articulatory Phonetics

In order to take this course, students are required to have received a grade of C or higher in LIN 201: Phonetics at Stony Brook University. The first two lessons in this course will provide a basic review of the material covered in LIN 201. It is expected that all students enrolled in this course have a strong familiarity with the [International Phonetic Alphabet](#). Being able to use IPA symbols in your typed documents is a useful skill. The [typeit](#) website is an easy way to copy and paste non-native symbols into your documents, but it's better to install an IPA keyboard directly to your computer. I recommend an IPA Unicode keyboard. Information on how to install this type of keyboard layout on your computer can be found [here](#). If you need help with installation, please let me know.

### 1.1 The International Phonetic Alphabet

The International Phonetic Association was formed in Paris in 1866 (originally called *Dhi Fonètik Tīcerz' Asóciécon*) and one of its main goals was to create a phonetic alphabet that could be used to help second language learners pronounce foreign words in a more native-like way. Since then, the International Phonetic *Alphabet* (IPA) has become one of, if not the, most commonly used phonetic alphabets.

Phonetic alphabets are a useful tool for representing language because their use ensures that anyone reading a word will know exactly how to pronounce it. The reason that this can happen is due to a one-to-one mapping between symbol and sound. What this means is that every symbol of the

phonetic alphabet signifies one and only one sound. This is the opposite of the English alphabet where you may have many symbols representing a single sound, or one symbol representing many sounds.

For example, the following sounds all contain an identical sound, but it is spelled differently in each form:

sea  
teeth  
me  
city  
achieve  
perceive

Notice that the sound corresponding to how we pronounce the letter ‘e’ is in all these words, but in each case, it is a different set of letters that represent that sound. In English spelling, ‘ea’, ‘ee’, ‘e’, ‘y’, ‘ie’, and ‘ei’ can all correspond to the same sound. This is a **many to one mapping**.

Now, let’s look at the following set of words that contain the same sequence of letters that correspond to a different sound in each word:

ghost  
rough  
though

In these examples, we see the letter combination ‘gh’ in each word, but it corresponds to a different sound (or lack of any sound) in each word. In ‘ghost’ the ‘gh’ is similar to what we might expect for a ‘g’ sound. In ‘rough’ the ‘gh’ is similar to what we might expect for an ‘f’ sound. And in ‘though’ the ‘gh’ corresponds to the absence of any sound. This is a **one to many mapping**.

The IPA mediates this problem by providing a **one to one mapping** between each sound and a corresponding symbol. This means that when we see a word transcribed this way, we know exactly how to pronounce it. When transcribing words in IPA, it is standard to put the symbols inside square brackets like this: [helou]. By doing so, we can differentiate between spelling and transcription.

Below you will find the symbols for English consonants and English vowels. Each symbol is paired with a set of words that contain the specific

sound. The sounds in the English word that contain the specific sound are bolded.

[p]	<b>pin</b> , lip, happy	[b]	<b>bin</b> , club, hobby
[t]	<b>tin</b> , mitt, atomic	[d]	<b>den</b> , hid, adorable
[k]	<b>king</b> , elk, tricky	[g]	<b>game</b> , flag, piggy
[tʃ]	<b>chip</b> , coach, itchy	[dʒ]	<b>jam</b> , huge, badger
[f]	fit, elf, stuffy	[v]	<b>vine</b> , glove, oven
[θ]	<b>think</b> , tooth, ethics	[ð]	<b>this</b> , smooth, brother
[s]	sale, mass, messy	[z]	zero, breeze, basil
[ʃ]	<b>ship</b> , fish, pushy	[ʒ]	(genre), measure, beige
[h]	hill, ahead	[ŋ]	<b>ring</b> , singer
[m]	<b>mouse</b> , gum, amish	[n]	<b>net</b> , win, honey
[l]	lost, <b>pill</b> , apollo	[ɹ]	<b>round</b> , core, curry
[w]	<b>witch</b> , awake	[j]	yellow, beyond

Table 1.1: English consonant symbols and corresponding words that contain each sound.

[i]	sheep	[ɪ]	ship	[eɪ]	game	[ɛ]	pen	[æ]	cat
[u]	food	[ʊ]	foot	[oʊ]	soap	[ɔ]	cought	[ɑ]	cot
[ʌ]	cut	[ɜ]	bird	[aɪ]	tie	[aʊ]	house	[ɔɪ]	boy

Table 1.2: English vowel symbols and corresponding words that contain each sound.

If we return to our first example from above, we now have a way to show that, despite the spelling differences, each word contains the same vowel sound [i].

## 1.2 Articulatory Description

Recall that articulatory phonetics is the study of how we produce linguistic sounds. One thing that we can do is describe the different ways we position our articulators to make the individual speech sounds. Articulators can be broken down into two types: **active articulators** which move and **passive articulators** which don't/cannot move.

The IPA chart for consonants is given below. There are three basic terms we can use to describe any linguistic sound. **Place (of articulation)** refers to where along the vocal tract a constriction is being made. The different places of articulation are represented in the columns of the IPA chart. **Manner (of articulation)** refers mainly to the size of the constriction being made, but also can refer to secondary properties such as whether or not air is flowing through the nose and if air flows through the center of your mouth or through the sides of your mouth. The different manners of articulation are represented in the rows of the IPA chart. **Voicing** refers to whether or not your vocal folds are vibrating while making a sound. Certain sounds vary only in whether or not the vocal folds are vibrating while they are being made. That is, certain sounds can have the same place and same manner, but vary along this voicing dimension. Sounds are said to be **voiced** if the vocal folds are vibrating and **voiceless** if they are not. This contrast is represented by orientation within cells in the IPA chart. Sounds on the left side of the cell are voiceless and sounds to the right side of the cell are voiced.

THE INTERNATIONAL PHONETIC ALPHABET (revised to 2015)

CONSONANTS (PULMONIC) © 2015 IPA

	Bilabial	Labiodental	Dental	Alveolar	Postalveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Glottal
Plosive	p b			t d		ʈ ɖ	c ɟ	k ɡ	q ɢ		ʔ
Nasal	m	ɱ		n		ɳ	ɲ	ŋ	ɴ		
Trill	ʙ			ɾ					ʀ		
Tap or Flap		ⱱ		ɾ		ɽ					
Fricative	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	h ɦ
Lateral fricative				ɬ ɮ							
Approximant		ʋ		ɹ		ɻ	j	ɰ			
Lateral approximant				l		ɭ	ʎ	ʟ			

Symbols to the right in a cell are voiced, to the left are voiceless. Shaded areas denote articulations judged impossible.

Figure 1.1: Official IPA consonant chart

### 1.2.1 Place of Articulation

Every place of articulation on the IPA chart is really a combination of **active/lower articulators** and **passive/upper articulators**. The active articulators are the lower lip, the tongue tip/blade, the front of the (body of the) tongue, the back of the (body of the) tongue, and the tongue root.

The passive articulators are the upper lip, the upper teeth, the alveolar ridge, the (hard) palate, the velum, the uvula, and the pharynx. Figure 1.2 shows where each of these articulators are located. It also includes some landmarks which are not used for articulation (such as the upper teeth), as well as the glottis which is used for making the glottal speech segments [h], [ɦ], and [ʔ] that don't really adhere to the same active/passive combination as all the other sounds.

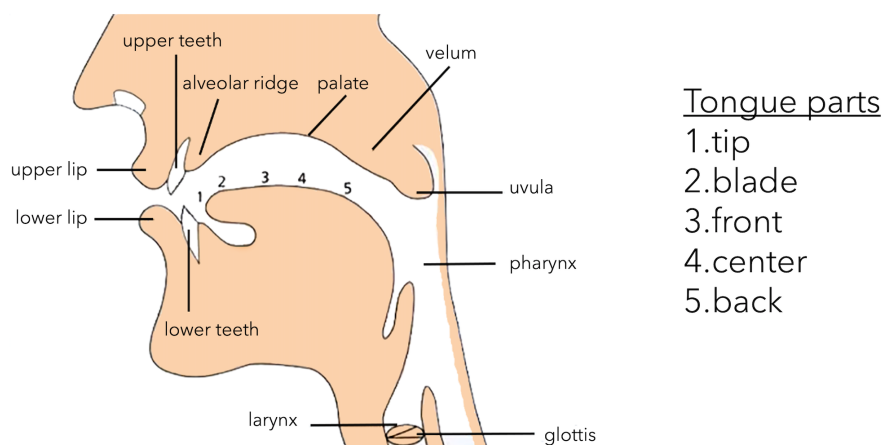


Figure 1.2: A list of most articulators used in speech production.

Some articulator pairs can be deduced from the name alone. For example, a bilabial sound uses the lower lip and upper lip as its articulators (*bi*-meaning two and *labial* meaning lips). Other pairs cannot be deduced from the name alone since only the passive articulator is given. We know that a palatal sound is made at the palate which is a passive articulator (listed above), but what is the active articulator involved in making these types of sounds? The answer here is that it is the active articulator that is closest to it (in this case, the front of the tongue). So Dental/Alveolar/Postalveolar sounds are all made with the tongue tip/blade as the active articulator, Palatal sounds are made with the front of the tongue as the active articulator, Velar and Uvular sounds are made with the back of the tongue as the active articulator, and Pharyngeal sounds are made with the root of the tongue as the active articulator. You may be wondering why we did not include Retroflex sounds anywhere in this discussion. This is because Retroflex is not a place of articulation in the same way that these other

sounds are (despite being lumped in with it on the IPA chart). Instead, a Retroflex sound is made in the same general area as an alveolar sound, only the tongue tip curls backwards and the underside of the tongue is used as the active articulator.

[Seeing Speech](#) is an interactive website created by phoneticians in Scotland. It is an extremely useful resource for students and anyone trying to learn about the phonetic properties of speech. The link here takes you to a clickable IPA consonant chart. Here, you are able to click on a speech sound which will trigger a video that provides audio and video of that specific consonant in between two vowels. At the top of the page you are able to choose between three video types: ANIMATION, MRI, and ULTRASOUND. I suggest ignoring the ULTRASOUND setting as it is hard for newcomers to discern what is going on. As an activity, click on one of the alveolar plosive sounds ([t] or [d]). Notice how the tip of the tongue makes full contact with the alveolar ridge. Now, click on one of the retroflex plosive sounds ([ʈ] or [ɖ]). Here, you should see that the tongue tip curls backwards and it is the underside of the tongue that is making contact in the area between the alveolar ridge and the palate (it may be useful to use the ANIMATION setting here to get a clearer view of what is going on).

We end our discussion of place of articulation by talking about the tongue. The tongue is arguably the most important speech organ since it is involved in creating a majority of the consonant sounds in the world's languages and in creating all of the vowel sounds. Broadly, the tongue can be divided into three areas: the blade of the tongue, the body of the tongue, and the root of the tongue. The tip of the blade also gets its own special term. Many types of sounds can be articulated using either the tip or the blade. In LIN 201 they were lumped together which resulted in the tip/blade being referred to as almost a singular unit. Later in this course we will discuss the variation in tip/blade articulations in more detail.

The body of the tongue can also be divided into three parts: the front, the center, and the back. Only the front and the back are used for making speech sounds. This is a result of the way this part of the tongue moves. The back of the tongue can be thought to move diagonally back and up towards the velum and uvula, while the front of the tongue can be thought to move diagonally forward and up towards the palate. Finally, there is the root of the tongue, which is not pictured in the diagram above, but can be moved forward and back. For consonants, it is moved back to make pharyngeal fricatives such as [ħ] and [ʕ] (found in certain dialects



of Arabic and other languages).

### 1.2.2 Manner of Articulation

Manner of articulation is a catch-all term that refers to multiple things. It can refer to the **degree of constriction** such as **stop**, **fricative**, or **approximant** (you may also sometimes see **affricate** used though this is not included on the official consonant section of the IPA chart). Figure 1.3 below shows the dynamic trajectory of each of these degrees of constrictions. For stop consonants, the active and passive articulator completely touch, blocking all air from exiting the mouth. For fricative consonants, the active articulator gets extremely close to the passive articulator in order to generate what is called **turbulent airflow**. This manifests as a hissing like sound and is the same phenomenon as when a car window is just slightly cracked open. For approximant consonants, the active articulator moves into the airstream towards the passive articulator, but not enough to generate turbulent airflow. Affricate consonants are a combination of stops and fricatives. So when these sounds are made, the active articulator and passive articulator completely touch for a brief moment before the active articulator pulls away just enough to create turbulent airflow.

Manner of articulation can also refer to where the air flows through the mouth. If it flows through the center of the mouth the term **central** is often used (though not listed on the IPA chart). If it flows through the sides of the mouth, the term **lateral** is used. Lateral sounds are made by curling the sides of your tongue up so that air is blocked from going down the center of your mouth and is instead forced around the sides. In English, we only have one lateral consonant: [l]. To get an idea of how the air is flowing through the side of your mouth we can do a quick activity. Make a [l] sound and hold it so your articulators stay in place. Now, breathe in. What you should feel is cold air coming in along the sides of your mouth. This is just the opposite of what is happening while you are making the [l] sound.

The final dimension that manner of articulation can refer to is **nasality**. This refers to whether or not the air flows only through the mouth (called an **oral** sound) or through both the nose and mouth (called a **nasal** sound). Physically, this involves your velum raising and lowering to block air from escaping through the nose. If we look at the IPA chart, we notice that

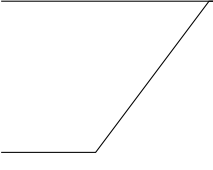
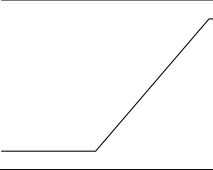
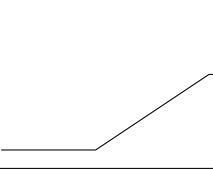
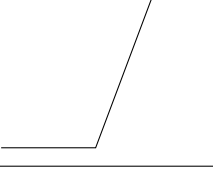
Type	Articulator Visualization		Description
Stop	upper/passive articulator		Lower/active articulator touches upper/passive articulator and blocks all air from escaping through the mouth
Fricative	upper/passive articulator		Lower/active articulator moves close enough to upper/passive articulator to generate turbulent airflow
Approximant	upper/passive articulator		Lower/active articulator moves into the airstream towards the upper/passive articulator, but not close enough to generate turbulent airflow
Affricate	upper/passive articulator		Lower/active articulator begins by moving to touch the upper/passive articulator but then retracts slightly away but remains close enough to generate turbulent airflow

Figure 1.3: Descriptions of the different degrees of constrictions used in speech

there is only one row for nasals. This suggests that they are all made in a similar way. Let's return to our discussion about degree of constriction from above. Can you guess what the degree of constriction for a nasal sound is? Make the [m] or [n] sound and think about what your active and passive articulators are doing. If you guessed "stop" as the degree of constriction then you are correct. Nasal sounds are all made by stopping air from exiting the mouth and lowering the velum so that it flows only through the nose. This is why you can hold a [m] sound out but not a [b] sound despite them both having the same active and passive articulators and degree of constriction.

Let's return now to the [Seeing Speech](#) consonant chart and look at nasal sounds and how they relate to oral stops. We will use the [m] and [b] example from the previous paragraph. First, click on the [b] and look at

what the articulators are doing. The lower lip comes up and touches the upper lip to block all air from exiting the mouth. Now, play the video again and notice that the velum raises up to block any air from going through the nose. Click on the [m] and do the same thing. Notice that the lower lip and upper lip are doing the exact same thing that they were doing during a [b] sound. Then, pay attention to the velum. Do you notice how it lowers to allow air to flow through the nose?

Before we move on, let's see how these three terms relate to the row headings of the IPA chart. The first row contains **plosive** sounds. All of these sounds have a stop degree of constriction, central airflow, and are oral. Colloquially, they are often just referred to as stops. The second row contains **nasal** sounds. All of these sounds have a stop degree of constriction, central airflow, and are nasal. Therefore, the only difference between nasals and plosives is the location of the velum and whether or not air is able to flow through the nose. Let's temporarily skip rows three and four.

The fifth row contains **fricative** sounds. All of these sounds have a fricative degree of constriction, central airflow, and are oral. Therefore, the only difference between fricatives and plosives is the degree of constriction. The sixth row contains **lateral fricative** sounds. As you may guess, all of these sounds have a fricative degree of constriction and are oral, but unlike the row above they have lateral as opposed to central airflow. The seventh row contains **approximant** sounds. All of these sounds have an approximant degree of constriction, central airflow, and are oral. Therefore, the only difference between plosives, fricatives, and approximants is the degree of constriction (or how close the articulators are to one another). The final row contains **lateral approximants** which once again are identical to the preceding row of **approximants** except for the fact that this row has lateral airflow as opposed to central airflow.

If we return to rows three and four we see rows for **trill** sound and **tap/flap** sounds. Trill sounds are made by jutting an articulator into the air stream in such a way that it naturally causes the airstream to open and close. There are only three trills known to exist in human language: the bilabial trill [ʙ], the alveolar trill [r], and the uvular trill [ʀ]. Tap sounds are essentially really fast stop sounds in that the two articulators touch, but only for a brief second. The alveolar tap [ɾ] is found in English words such as *butter* [ˈbʌt̬ə] or *better* [ˈbɛt̬ə]. Both trills and taps have central airflow and are oral. Descriptively, their degree of constriction

is their name: **trill** or **tap**. In actuality, it's a bit more complicated. In both cases, the oral cavity is completely cut blocked off. The difference between these sounds and stops is the amount of time in which this occurs. For a tap, the lower/active articulator just briefly makes contact with the upper/passive articulator. For a trill, the lower/active articulator is caused to make repeated brief contact with the upper/passive articulator due to the bernoulli effect (discussed in relation to voicing below). Figure 1.4 shows a schematization of both of these types of sounds.

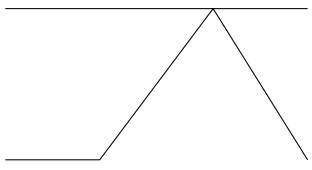
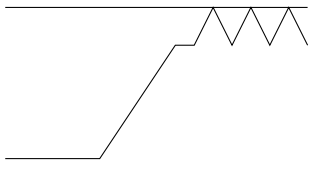
Type	Articulator Visualization		Description
Tap	upper/passive articulator		Lower/active articulator touches upper/passive articulator and immediately pulls away
Trill	upper/passive articulator		Lower/active articulator moves into the airstream with an appropriate tenseness. This causes it to be pulled/pushed away from the upper/passive articulator at semi-regular intervals

Figure 1.4: Descriptions of the different degrees of constrictions used in speech

To summarize, the rows of the IPA are labeled based on manner of articulation. Manner of articulation is broken up into three categories: degree of constriction, airflow location, and nasality. As the rows descend, the degree of constriction gets more and more open. If any degree of constriction allows for laterality, that is placed below the non-lateral row of the same degree of constriction. Nasality only appear in a single row where it is applicable.

### 1.2.3 Voicing

**Voicing**, at its most basic, refers to whether or not your vocal folds are vibrating while you are making a sound. In general, any type of sound can be voiced or voiceless, but it is usually the case that stops, fricatives, and affricates are they only types of sound that have a voicing **contrast**. When linguists use the term contrast, they mean that by changing a parameter

(voicing for our current case, but that's not the only parameter that can be used contrastively), the meaning of a word will change since there is a new, meaningful sound being made. Take the difference between [s] and [z] as an example. Both sounds have a fricative degree of constriction, central airflow, and are oral. Additionally, they both are made by bringing the tongue tip/blade (the active articulator) extremely close to the alveolar ridge (the passive articulator). The only way that they vary is based on whether or not the vocal folds are vibrating. But this is a meaningful difference since we can differentiate between words like *sit* [sit] and *zit* [zit] or *bus* [bʌs] and *buzz* [bʌz].

To get an idea of how only voicing changes between [s] and [z], try the following activity. Make an [s] sound and hold your hand over your vocal folds. Now switch to making a [z] sound. If you did it correctly, you should notice that the only change you can feel is the vibration of your vocal folds. The rest of your articulators should have stayed in place. You can try this exercise with other fricative pairs in English such as [f]/[v], [θ]/[ð], and [ʃ]/[ʒ].

In LIN 201 you were introduced to the basic articulatory property of voicing. It was based around the **bernoulli effect**. This is a general principle that states, given certain assumption, an increase in speed (of a fluid) occurs simultaneously with a decrease in pressure. When it comes to speech, the speeding up occurs when air tries to pass through the narrow space of the trachea that is the result of the vocal folds being placed closely together without a change in overall pressure. Since the speeding up occurs in between the vocal folds, this causes a decrease in pressure in that area and the vocal folds collapse in on each other. Pressure build up below the vocal folds causes them to be pushed back apart. By maintaining a satisfactory level of tenseness, the process will continue as the vocal folds spring back into their original position. Figure 1.5 schematizes this in an abstract way.

Generally, we can think of the process of voicing as a cycle that will only stop when one of the necessary conditions change:

1. Air flows between the vocal folds
2. Bernoulli effect pulls the vocal folds together
3. Pressure builds up beneath the closed folds

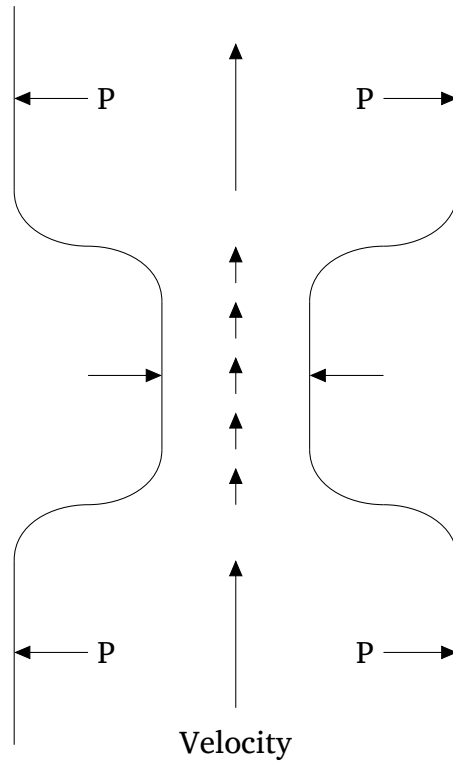


Figure 1.5: Schematization of the Bernoulli effect in speech production. The increased speed as air travels through the vocal folds causes a decrease in pressure which in turn causes them to collapse in on themselves. The steady flow of air increases pressure below the vocal folds and causes them to be pushed back open. This process repeats and creates vocal fold vibration a.k.a. “voicing”.

4. Vocal folds are pushed apart by the high pressure
5. Due to anatomical reasons, the vocal folds return to their original position
6. The cycle continues

You may remember from LIN 201 that there were three necessary conditions for voicing: the vocal folds are close together and protruding into the airstream (to speed up air flow), the vocal folds are appropriately tensed (so that they automatically return to their original position), and there is sufficient flow of air (to cause the speed up/pressure drop). If any of these conditions are not met, the cycle will cease to run.

### 1.2.4 Putting it all Together for Consonants

Now that we've reviewed the core descriptive terms, let's put it all together. You may remember that a lot of time in LIN 201 was spent giving articulatory descriptions of consonants. This may have seemed tedious, but we wanted to make sure that you had a strong understanding of how the various speech articulators worked in conjunction to create speech sounds. It is especially important when thinking about non-native sounds. If you know how the articulators work to make certain types of sounds in your language, you should be able to configure them in new ways to figure out how to make similar sounds in another language. We fit consonants into the following template: VOICING, PLACE, AIRFLOW, NASALITY, DEGREE OF CONSTRICTION. The options for airflow are *central* and *lateral*. The options for nasality are *oral* and *nasal*. Table 1.3 gives a few examples of consonant descriptions.

In addition to giving static definitions of consonants, we can also describe how the positions of the articulators change over time as they move from one consonant to the next. For example, we can describe how the articulators change in the word *helper* as they go from an [l] to a [p]: the vocal folds separate as it transitions from a voiced to a voiceless sound, the lower lip moves to touch the upper lip to prevent any air from escaping the mouth, the tongue tip moves away from the alveolar ridge and the sides of the tongue return to a position that allows for central airflow. You should be able to describe these types of movements for a transition between any two consonants.

	VOICING	PLACE	AIRFLOW	NASALITY	D.O.C
ŋ	voiced	retroflex	central	nasal	stop
χ	voiceless	uvular	central	oral	fricative
ʋ	voiced	labiodental	central	oral	approximant
ʈ	voiceless	alveolar	lateral	oral	fricative
dʒ	voiced	palatoalveolar	central	oral	affricate
c	voiceless	palatal	central	oral	stop

Table 1.3: Full articulatory descriptions for a variety of consonants

### 1.3 Articulation and Description of Vowels

Our review of the articulatory properties of vowels will be much shorter. In the second half of LIN 201 you may remember that it was revealed that vowels are organized by the IPA in terms of their acoustic properties, specifically values for the first two formants. Nonetheless, there are some loose connections between placement on the vowel chart and articulatory properties. **Tense/lax** roughly corresponds with advanced/retracted tongue root. In English this isn't *really* the case, but in other languages it is. **Vowel height** roughly corresponds with the size of the cavity between the tongue dorsum which can be related to its "height". **Vowel front-backness** roughly corresponds with which part of the tongue body is highest while the vowel is being made. **Roundness** is the articulatory property that is most direct: it refers to whether or not the lips are rounded while the vowel is being made. There is no "degree of constriction" for vowels, but they are in some sense making a constriction. It is less than what it is for an approximant. Recall that in consonants, the final term is the degree of constriction. You may remember that in LIN 201 we had the final term when describing vowels always just be "vowel". You can think of that as the vowel's degree of constriction.

This gives us a parallel way to describe vowels in a similar way that we described consonants. The template for vowels is: TENSENESS, HEIGHT, FRONTNESS, ROUNDING, VOWEL. The options for tenseness are *tense* and *lax*. The options for height are *high*, *mid*, and *low*. The options for frontness are *front*, *central*, and *back*. The options for rounding are *round* and *unround*. Our "articulatory" descriptions really fall apart when it comes to both diphthongs, as well as the mid-central vowels. For a much more in depth explanation on the articulation of vowels, see [Gick et al. \(2013\)](#).



## Lesson 2

# Review of Acoustic Phonetics

The review of acoustic phonetics will be brief for two reasons. One, not much is covered in LIN 201. Two, the remainder of this course will primarily focus on acoustic phonetics. Recall that linguistic phonetics is largely broken up into three subareas: articulatory phonetics, acoustic phonetics, and auditory phonetics. In this course, you will learn about ways that articulation is measured but will not get any hands on experience. Our primary focus will be acoustic measurements and you will get plenty of hands on experience in this area. The last couple lessons will focus on auditory phonetics and you will have a chance to run some simple speech perception experiments.

### 2.1 Spectrograms

Spectrograms are visual representations of sound waves that phoneticians and other speech scientists use to analyze the acoustic properties of speech. They are three dimensional graphs that represent time (x-axis), frequency (y-axis), and amplitude (z-axis). The z-axis is represented by the level of darkness in the spectrogram. So dark spots indicate frequencies in which there is a large concentration of energy. Above each spectrogram below is the corresponding sound wave. These graphs are two dimensional where the x-axis is once again time and the y-axis is (overall) amplitude.

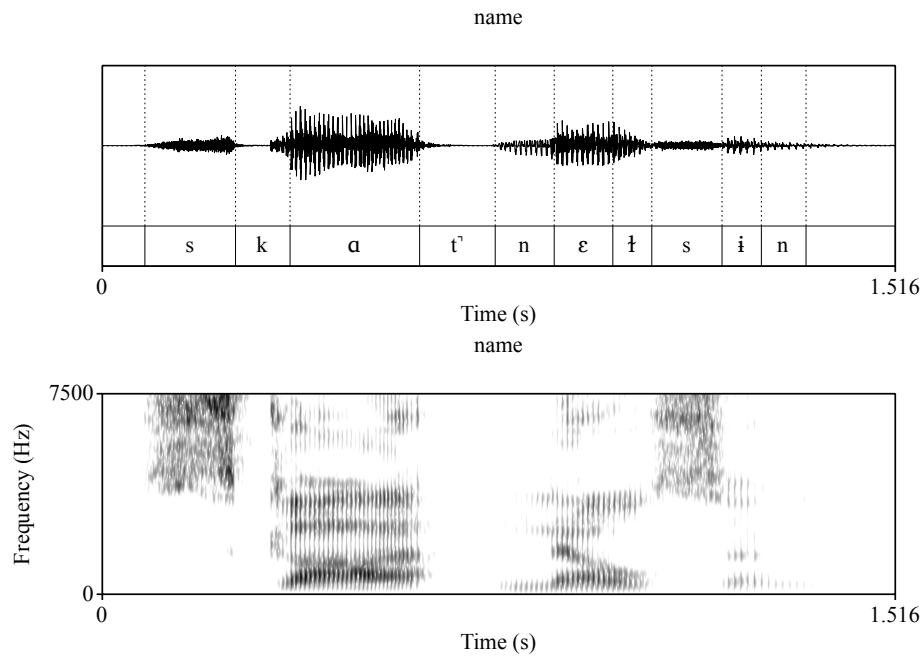


Figure 2.1: Waveform and spectrogram for [skat'nɛʃɪn].

On a spectrogram, there are three basic types of information we are interested in: resonance, silence, and noise. In LIN 201 you were introduced to the **source-filter theory**. We will cover this more in depth in this class, but you may remember that the basic idea is that your vocal folds act as a source of sound (via vibration) and this sound is then shaped by the configuration of the articulators in the vocal tract. Of course, this can't be the entire picture because we have voiceless sounds as well! One crude way we can split the difference is to say that vocal fold vibration leads to **resonance** while the turbulent airflow seen in voiceless fricatives, the second half of approximants, and with aspiration leads to **noise**. These are both technical terms that we will get into in more depth later, but you hopefully remember that resonance shows up on a spectrogram as concentrated dark bands of energy while noise has no apparent pattern. **Silence** is the third thing we can look for in a spectrogram and appears as white nothingness. Below, we review these terms in slightly more detail and show examples.

## 2.2 Resonance

The basic idea of resonance is that there are certain frequencies of sound that are amplified when passing through an object. It is usually not a *specific* frequency even though it is often talked about that way. In general, it is a band of frequencies but we refer to the central frequency as the location of the resonance. In general, everything has a resonant frequency that will cause it to vibrate. If you have 45 minutes to spare and a lot of patience, you can listen to this piece of sound art by Alvin Lucier called *I am Sitting in a Room*. It demonstrates the resonant frequencies of a room by continuously playing back the same recording into a room over and over again. Each cycle, the frequencies outside the resonant areas are filtered out until all that is left is the resonant frequencies of the room itself!

When it comes to speech, these resonances are called **formants**. They are caused by placing the vocal tract into different configurations. They most clearly show up in vowels. On a spectrogram, they appear as dark horizontal bands. There are many of them, but primarily we are interested in the first three: F1, F2, F3. The numbering of formants begins from the bottom and moves up. If you're wondering why we don't include the dark band at the very bottom of the frequency range, this is the voicing bar which we will return to in a moment.

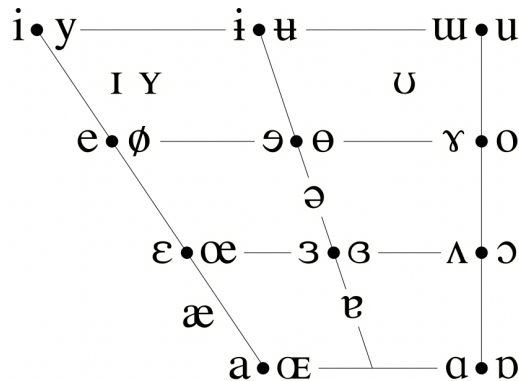


Figure 2.2: Official IPA vowel chart

For vowels, the IPA uses the values for F1 and F2 to chart their location. F2 is reversed on the x-axis and F1 is reversed on the y-axis. This means that vowels with a high F2 value are more front while vowels with a high

F1 values are more low. Figure 2.2 is the official IPA vowel chart. If we were to measure the F1 and F2 values for each of these vowels and plot those directly, we would get something that looks basically identical to this.

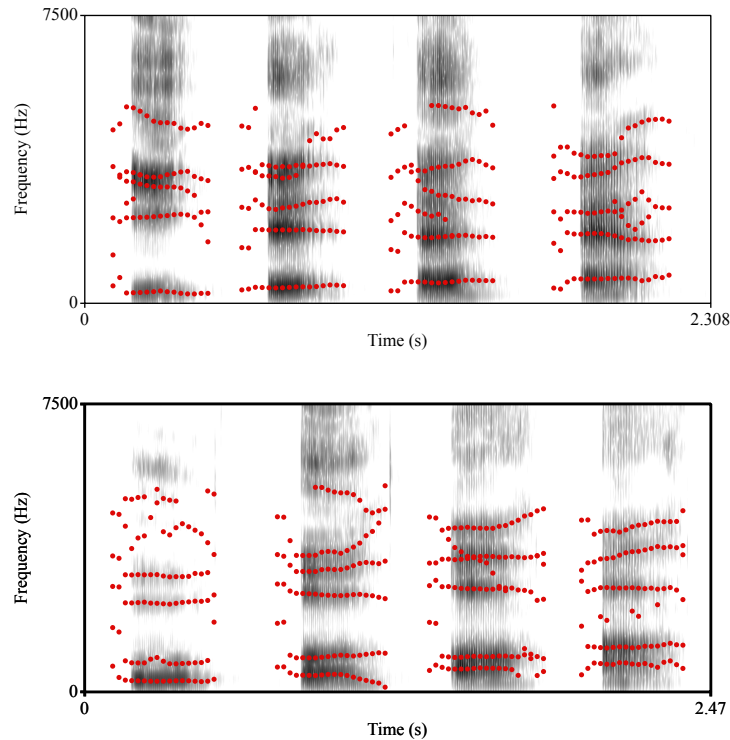


Figure 2.3: Spectrograms with automatic formant detection overlaid. Top row are the vowels [i], [ɪ], [ɛ], [æ]. Bottom row are the vowels [u], [ʊ], [ɔ], [ɑ].

We can look at the formant values for some American English vowels directly. Figure 2.3 includes spectrograms for four of the front and four of the back vowels and are overlaid with the location of the formants in red. The top row are the front vowels [i], [ɪ], [ɛ], and [æ]. Notice that we see a gradual rise in F1 as we go from left to right, indicating that the vowels are getting lower in the vowel space as expected. We also see F2 drop a little which may be unexpected since they are all “front vowels”, but if we look at the IPA chart we do notice a diagonal line along the

front dimension indicating that F2 should lower for the front vowels as F1 increases.

The bottom rows are the back vowels [u], [ʊ], [ɔ], [ɑ]. Since we're going from a high vowel to a low vowel, we should also expect to see F1 increase from left to right which we do. In general, the back vowels F2 is not correlated with F1, but you may notice that my vowel [ɑ] has a higher F2 than the other three "back" vowels. This is a property of my dialect and is likely due to having a lower [ɔ]. As you may remember some people have these vowels merged into a single vowel. My fronting is likely to ensure that the two are not merged.

While vowels have the strongest formants (resonance), many consonants also have formants as well. The reason that vowels are the most resonant is related to the fact that the oral cavity is least restricted in the production of vowels. But approximants also have visible formants in the spectrogram. The approximants [j] and [w] are often called **glides** or **semi-vowels**. This is because they are quite similar to the vowels [i] and [u] in their spectral make up. The formants in [j] and [w] are similar to what you will find in [i] and [u], but they typically are not as strong and the segment is shorter in duration.

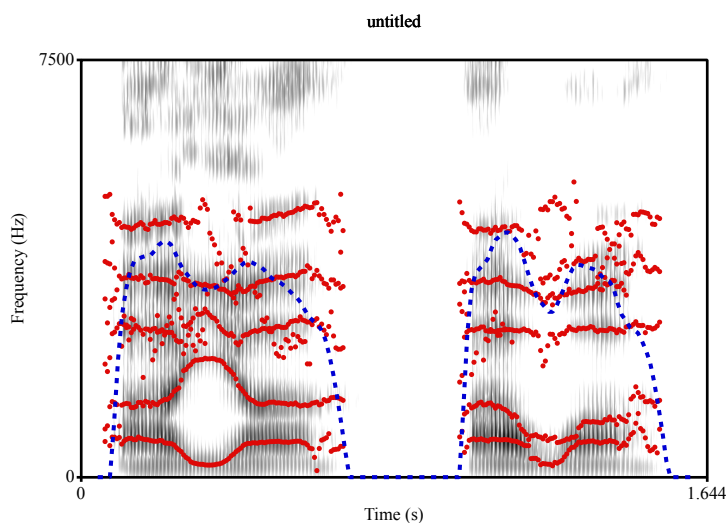


Figure 2.4: Spectrograms for [ajɑ] (left) and [awɑ] (right). Red dots are formant trajectories. Blue dashed line shows intensity profile.

Figure 2.4 shows spectrograms for [ajɑ] and [awɑ]. The red dotted

lines once again show the formant trajectories. These spectrograms also included a blue-dashed line showing the intensity profile. Notice that in both sequences, the intensity drops in the middle while the glide is being made, even though there are still formants. This is a characteristic of glide sounds. Also, notice that both glides match the formant profiles of their corresponding vowels: [j] on the left has a high F2 and low F1 just as [i] does while [w] on the right has a low F2 and low F1 just as [u] does.

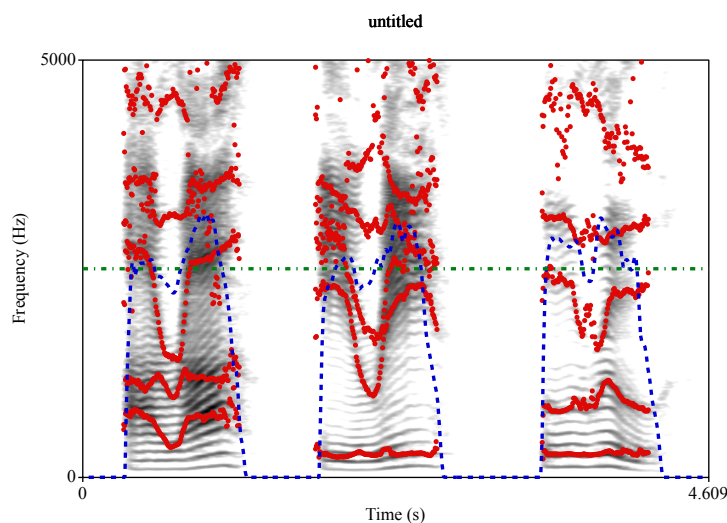


Figure 2.5: Spectrograms for [ɑɑ] (left), [i.i] (center) and [uru] (right). Red dots are formant trajectories. Blue dashed line shows intensity profile. Green horizontal line at 2500 Hz.

The English non-glide approximants [ɹ] and [ɻ] also have formant patterns that are noticeable. Starting with [ɹ], the main acoustic property of this sound is a lowered F3 (and to an extent a lower F2) as well. Figure 2.6 shows spectrograms for the English [ɹ] in between the low back vowel [ɑ] (left), the high front vowel [i] (center), and the high back vowel [u] (right). Once again, formant and intensity profiles are given in red and blue. All of these “words” were pronounced with stress on the second syllable. The intensity profile is therefore highest for the stressed vowel, second highest for the unstressed vowel, and lowest for the approximant. Again, remember that approximants have a more closed off oral cavity which reduces overall intensity, so this result is expected. The primary cue for English [ɹ] on a spectrogram is the lowering of F3 below 2500 Hz.

A green line indicating 2500 Hz is shown in Figure 2.6 as well to highlight the giant decrease in F3. Astute readers will also notice that F2 drops significantly as well, but this is only obvious in the middle spectrogram because [i] is the only vowel of the three that has a high F2 to begin with.

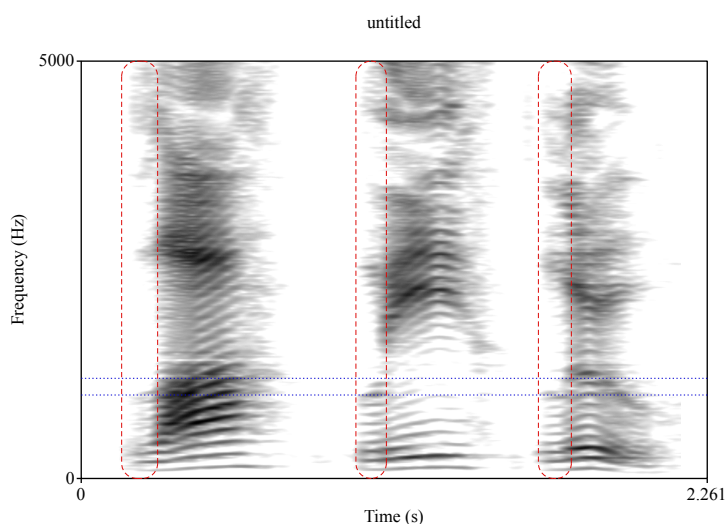


Figure 2.6: Spectrograms for [la] (left), [li] (center) and [lu] (right). Red dashed rectangles show the location of the onset [l] in each word. Blue lines indicate the range between 1000-1200 Hz.

The English onset [l] has a noticeable formant, usually around 1100 Hz. Figure 2.6 shows spectrograms for me pronouncing [la], [li], and [lu]. The location of the onset [l] is circled in red and the blue lines indicate a range of 1000-1200 Hz. Notice that for me, the formant shows up at the lower end around 1000 Hz. Coda [l] is less reliable to read on a spectrogram and we did not discuss in much in LIN 201. So we will leave it aside for now.

In leu of a figure, I will just remind you that nasal stops also usually have a visible formant around 2500 Hz. The acoustic properties of nasals and laterals are even more interesting than you were lead to believe in LIN 201. The articulatory aspect of both sounds lead to similar acoustic outcomes, albeit at different locations of the frequency spectrum. When articulating both of these types of sounds, a speaker creates a **side-branch** where air can flow into, but not escape from. This creates all sorts of interesting effects that we did not discuss in LIN 201. I'm hoping we have

enough time to discuss both of these types of sounds in more detail in this course. If you're interested, [Johnson \(2011\)](#) has a nice chapter that explains it in much more detail.

## 2.3 Silence

The idea of being able to “see” silence on the spectrogram is an interesting idea. Most of the recordings we work with, unless they were recorded in a special type of isolation booth, won't ever be completely silent. Instead, silence is usually viewed as when the overall sound level is below some low threshold. On a spectrogram, this appears as a large area of whiteness. Recall that on a spectrogram, the amount of darkness in a given area corresponds to how much energy there is at those frequencies. If there is no energy at any frequency then we will get whiteness in the spectrogram. The only type of sound that completely blocks all air from escaping (thus any sound) are the oral stops. Affricates also have this property because the first half of their articulation is just an oral stop.

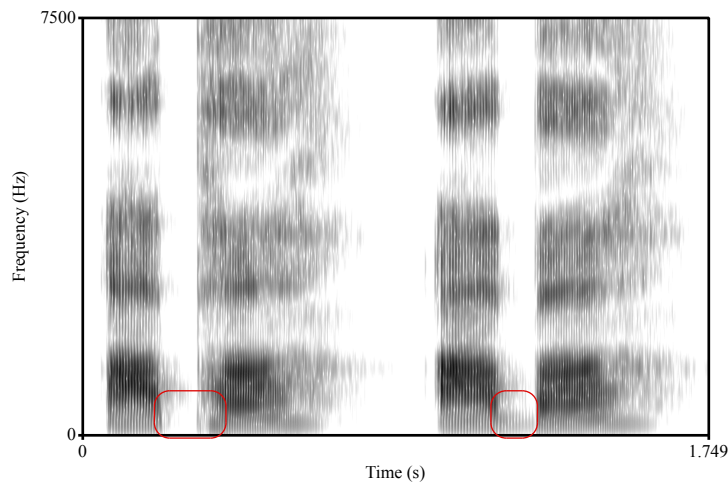


Figure 2.7: Spectrograms for [apa] and [aba]. Red circles indicate the lack or presence of a voicing bar.

There is one caveat of course, which is voiced stops in English between vowels are not completely “white” throughout the entire spectrogram dur-



ing the oral closure. You may remember from LIN 201 that voicing appears as a dark bar on in the low frequencies of the spectrogram. So we must take this into account when considering what is silence. Somewhat relatedly, the sound of your vocal fold vibrating emanates directly through your throat out into the world. This is how it is able to show up on the spectrogram even when no other sound is able to escape. Figure 2.7 shows spectrograms for [apɑ] and [ɑbɑ]. The red circles show the lack and presence of a voicing bar. Notice that in both cases the frequencies above this area are all white, indicating silence.

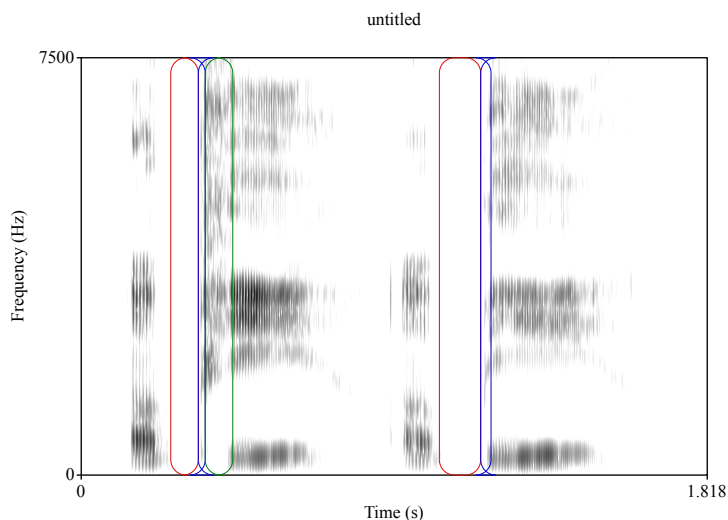


Figure 2.8: Spectrograms for [ə'pʰi] and [ə'bi]. Red circles indicate silence, blue circles indicate release burst, green circle indicates aspiration.

We end this section with a reminder that oral stops can be broken down into two parts: a closure and a release. The closure is the part where there is silence. There is always a release burst, but at the beginning of words there is variation between voiceless and voiced stops. Voiceless stops in English are aspirated, which means there is an additional period of noisy voiceless that appears on the spectrogram. Figure 2.8 shows the phrases *a pea* and *a bee*. The closure, release burst, and aspiration are circled in red, blue, and green. Notice, there is no voicing bar for the “voiced” [b]. You may recall from LIN 201 that word initial stops contrast in a property called **voice onset time** of VOT. We will do a deep dive into VOT later in this course.

## 2.4 Noise

The last acoustic information related to spectrograms we will look at is noise. Noise appears as (somewhat) equally dispersed energy throughout the spectrum. Therefore it is differentiated from silence by the fact that there is actually energy throughout the spectrum, and it is differentiated from resonance by the lack of clear formants and energy pooling around specific frequencies. Fricatives are the main sound type that has noise, but since the second half of an affricate is also like a fricative, affricates also have noise. Furthermore, the aspiration portion of voiceless stops also appears as noise on a spectrogram.

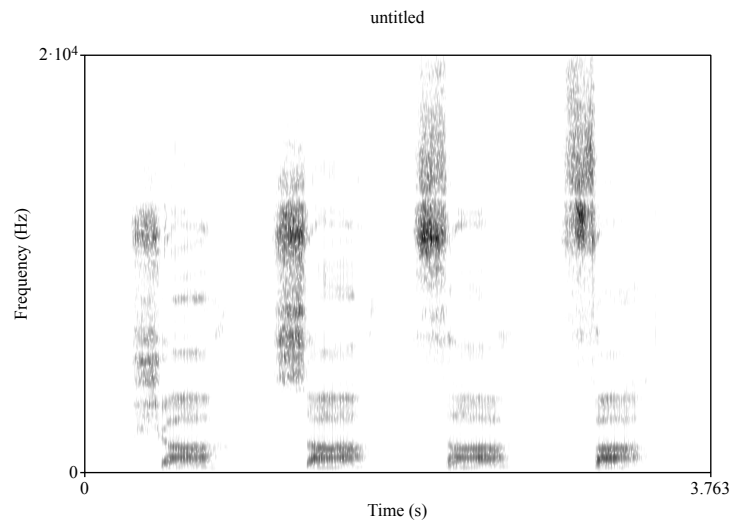


Figure 2.9: Spectrograms from left to right: [ʃa], [sɑ], [θɑ], [fa]

One thing we know about fricatives is that the location of their energy does provide cues to the place of articulation. Unlike vowel formants, we're more interested in a broad range of where the energy may peak. The location of the energy peak is directly related to how front in the mouth the constriction is made. Therefore, fricatives made more in the back of the vocal tract will have a lower peak while fricative made more in the front of the vocal tract will have a higher peak. This can be shown mathematically based on **tube models** of the vocal tract. These were covered briefly in LIN 201 and we will do a slightly deeper dive in this course. The general idea is that a constriction further back in the mouth causes there to be

a longer “tube” in front of it in the vocal tract which shapes the noise. The longer tube cause waves with longer wavelengths to be amplified and longer wavelength corresponds to a lower frequency. Hence, why the peak is lower for fricatives made further back in the mouth.

Figure 2.9 shows a spectrogram containing [ʃa], [sa], [θa], [fa]. As you should be able to say, energy is lowest on the left for [ʃa] and highest for [fa] on the right. That being said, there’s not much of a difference in the lowest peak for [θa] and [fa]. It turns out that in perception experiments these sounds often get confused. One reason for this is the small difference in “tube” sizes, but they also have much weaker energy across the board than the more strident fricatives [s] and [ʃ].



**Part II**  
**Main Lessons**



## Lesson 3

# f<sub>0</sub>-pitch-tone-intonation

In this lesson we will discuss the way in which vocal fold vibration is used to encode linguistic information. The first part of this lesson will be a little math heavy, but stick with it. It's perfectly ok to only have an *intuition* as to what the math tells us rather than fully understanding how it all works. Providing a formalization of the underlying system is useful because it helps us understand why things are the way they are. Furthermore, it can help us identify when things deviate away from expectations.

### 3.1 Sine Waves

You may remember **sine waves** from a trigonometry class. There are certain aspects of sine waves that we are interested in for speech. The first is the **period** of the wave which is the duration of a single cycle of the wave. You can measure from any point in the wave as long as you measure to the same point in the following cycle. In figure 3.1 there is only a single cycle, but we can still measure its duration from where it crosses the origin on the x-axis. The x-axis scale is in seconds, so the **period** of this wave is 4 seconds. **Frequency** of a sine wave can be measured based on the number of times it repeats per some unit of time. The unit used in speech analysis is seconds. We measure cycles per second which is more commonly referred to as **Hertz (Hz)**. In figure 3.1, the wave has an incredibly low frequency because it has only finished  $\frac{1}{4}$  of its cycle at the 1 second mark. This means the frequency is only 0.25 Hz which is well below what any human listener

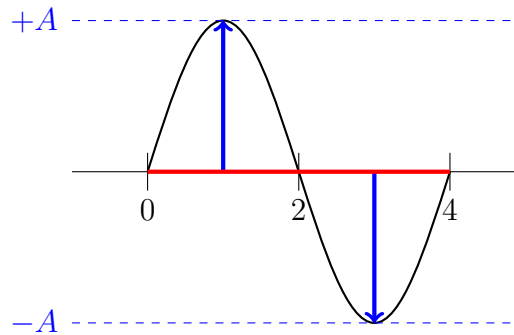


Figure 3.1: A simple sine wave with frequency  $\frac{1}{4}$  Hz and amplitude  $A$ .

can perceive.<sup>1</sup>

A second aspect of sine waves we are interested in is the **amplitude**. This is how much the wave deviates from its neutral position (the x-axis). Mathematically, sine waves move equally in both directions. So the amplitude is really the absolute value of its largest displacement. This ensures that the amplitude is always positive. In figure 3.1 the amplitude is represented with blue dashed lines and blue arrows. Mathematically, the x-origin can be at any value, but for our purposes we can just assume it's 0 and the amplitude is some value  $A$  higher or lower than 0.

One important thing to note is that **frequency** and **amplitude** are not related. That is, you can change the values independently without affecting the other. Figure 3.2 displays this property. The leftmost wave has the same frequency as the wave in figure 3.1 but twice the amplitude. The center wave has the same amplitude as the wave in figure 3.1 but twice the frequency. The rightmost wave has the same frequency as the center wave but twice the frequency. A practical example is this, imagine you were singing the note corresponding to middle C on a piano (256 Hz). A low amplitude would correspond to trying to sing it very quietly while a high amplitude would correspond to trying to sing it very loudly. In both cases, the fact that its frequency is 256 Hz does not change. You can play the same game for amplitude.

Why do these aspects of sine waves matter? We know that any complex

<sup>1</sup>Actually, it appears that this is lower than what any animal can perceive, as the [Guinness \(Book of\) World Records](#) lists the homing pigeon as the animal that can perceive the lowest frequency at 0.5 Hz



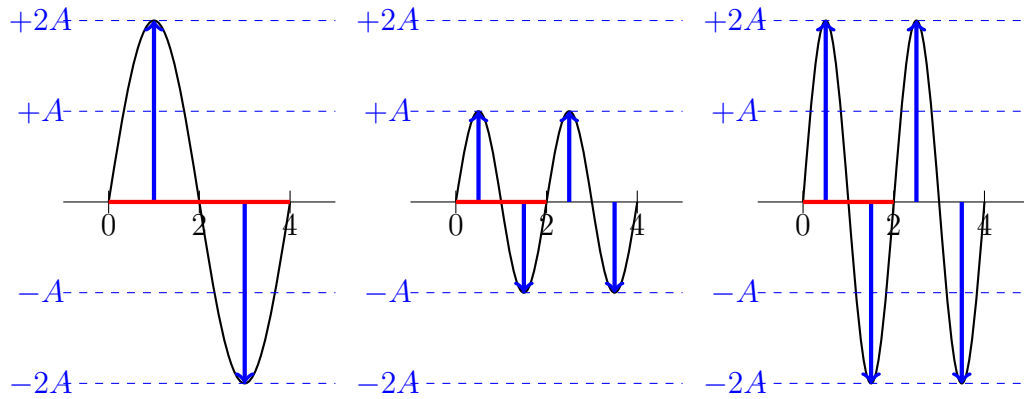


Figure 3.2: Sine

sound can be decomposed into the linear combination (sum) of simple sine waves. Pure sine waves don't really occur in nature but provide an intuitive way to talk about the complex waves that do occur. In this sense, sine waves do occur naturally, but only at a suitable level of abstraction: they are the component parts of the speech we produce and perceive every day.

There is a second aspect of this that is important for speech research. Given some assumptions, we can calculate a **fundamental frequency** (also referred to as **f<sub>0</sub>**) for a given complex wave. The fundamental frequency is the highest frequency which all the other waves share. Even if a sound is made up of a combination of many different frequencies, we will perceive it as the fundamental frequency. Furthermore, in speech research we can relate the fundamental frequency to the rate at which a speaker's vocal folds are vibrating while speaking a given utterance.

Before we completely dive into the speech aspect of fundamental frequency, let's try to get some intuition as to how amplitude and frequency interact when decomposing (and/or constructing) complex waves. Figures 3.3, 3.4, and 3.5 all contain sine waves with frequencies of 1 Hz, 2 Hz, 3 Hz, and 4 Hz in the first four rows. The fifth row is the combination of these waves. This is done by taking the value on the y-axis (amplitude) at every matching point along the x-axis and summing them all together. The outcome is a combination of each of the four sine waves.

In figure 3.3, each of the simple sine waves has the same amplitude, resulting in a complex wave that has equal properties of each wave. In

figure 3.4, the amplitude of the sine waves decrease as the frequency increases. This results in the the complex wave having properties more like the waves with lower frequency. For figure 3.5 it is the reverse: the amplitude of the sine wave increases as the frequency increases, resulting in a complex waves with more of the properties of the waves with higher frequencies. In all cases, the fundamental frequency is the same: 1 Hz. This is because it is the greatest shared frequency of the four waves. You should be able to see in the figures how the wave repeats at this frequency even though it's not a perfect sine wave.

## 3.2 Fundamental Frequency in Speech

Given we are now somewhat familiar with what fundamental frequency (i'll use  $f_0$  from here on out) is at an abstract level, let's discuss how it is used in speech to encode linguistic information. To start, let's compare some terminology. When we use the term  $f_0$  we are giving a numeric description of the signal. We might say something like, "the  $f_0$  at the midpoint of the vowel is 112 Hz," or "the highest  $f_0$  in this sentence is 200 Hz and occurs during the final word." This is a **quantitative** description. The corresponding **qualitative** description for this phenomenon is **pitch**. Pitch relates to how speakers perceive the  $f_0$  of a given utterance. Here we might say something like, "that sounded high pitched," or "can you say that again with a lower pitch." This is a non-numeric description and therefore is always relative in some sense.

The next term we can talk about is **tone**. Tone is also measured based on its  $f_0$  value. Similarly to pitch, there are high tones and low tones (and all sorts of in between), but unlike pitch it is more than just the relative frequency of a given utterance. Tone specifically refers to implementational targets. We won't get into specifics now because the next lesson will dive deeper into tone specifically, but its important to recognize that tone is another term used to describe  $f_0$ , but in a very specific way. Relatedly there is **intonation**, which can be thought of as the use of  $f_0$  over a larger domain than a single word/syllable. Of course, this isn't a perfect explanation, but it's a good enough starting point Intonation can be thought of as the melody over an entire phrase. You may remember from LIN 201 that we used values like M(id), H(igh), and L(ow) to describe intonational patterns. In the rest of this lesson we'll review the basics of

intonation and then look a little bit more at the acoustic and articulatory aspects of  $f_0$ .

### 3.3 Intonation

Intonation is the the tune or melody of an utterance and it can signal beliefs and expectations, turn taking, among other things. In LIN 201 you may remember learning about two basic intonational melodies used in American English: the **Declarative contour** (MHL) and the **yes/no question contour** (MH). These intonational melodies were defined by the placement of the various pitches at different targets within intonational phrases.

The declarative contour places a high marker on the strongest phrasal stress, the low marker immediately after that, and the mid marker at the beginning of the utterance. The contour starts mid, then jumps up to high once it reaches the marker, then jumps down to low immediately after and extends to the end of the sentence as needed. Figure 3.6 shows the pitch contour for the sentence “I won a monkey”. You can see that at the beginning of the utterance the pitch value is in the middle region before raising up to its highest point during the first syllable of “monkey”, and then dropping to its lowest value for the second syllable of “monkey”.

The yes/no contour places a high marker at the very end of the utterance and a mid marker at the very beginning of the utterance. The contour starts mid and then rises from mid to high once it reaches the strongest phrasal stress. Figure 3.7 shows the wave form and pitch contour for the questions, “did you see that monkey?” The strongest phrasal stress is on the first syllable of “monkey” and that’s exactly where we see it rise. Notice, this is the same spot where the declarative sentence rises to high as well.

There are many other possible intonation patterns that speakers can use. These are just a couple basic examples and serve as useful examples rather than exhaustive coverage of the phenomena. What I hope you takeaway from this section is a newfound understanding of what exactly is goin on in the articulatory and acoustic space when speakers vary their intonation.

In the articulatory space, speakers are varying the tenseness of their vocal folds to alter the frequency at which they vibrate. When the vocal

folds are more tense, they spring back into position more quickly, thus resulting in more cycles per second. This is why it is harder to sing higher notes: it literally puts more strain on your vocal folds.

The acoustic consequence of changing the tenseness of the vocal folds is a change in  $f_0$  or the fundamental frequency. Think now about all the different things we can signal with intonation, both linguistic and non-linguistic. We do this all by varying the rate at which our vocal folds vibrate. Throughout the rest of the course we will continue to make connections between articulations and their acoustic consequences. I hope this mini-review of intonation has provided a simple on ramp for the rest of our phonetic journey.

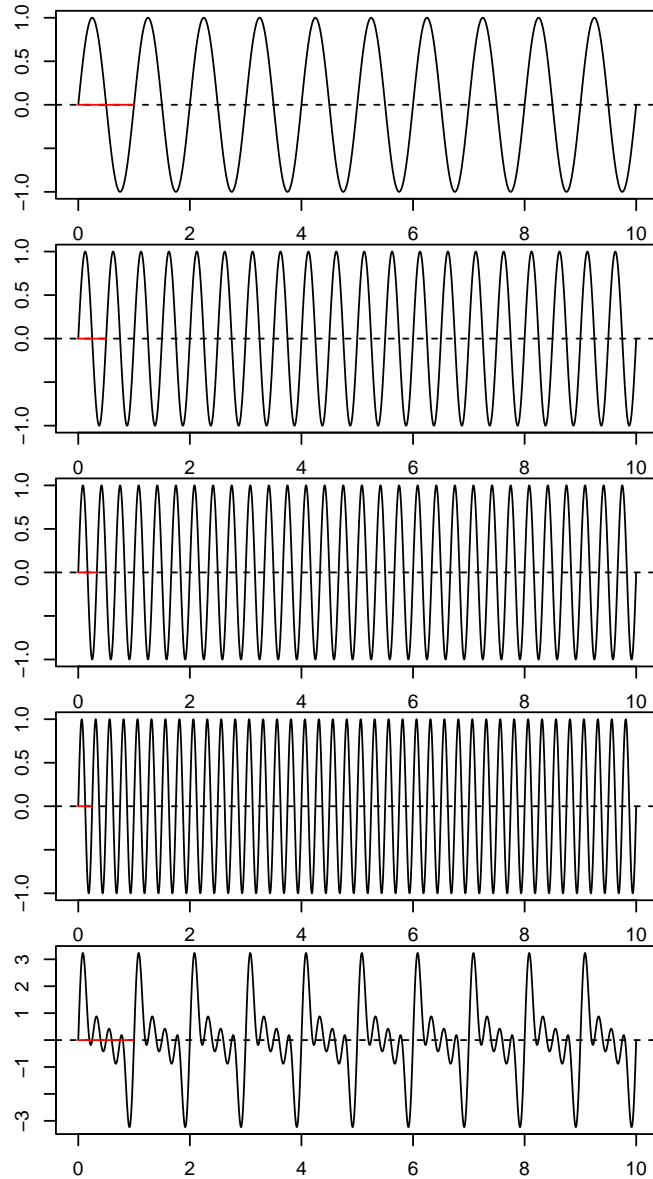


Figure 3.3: Rows 1-4 contain sine waves of increasing frequency with equal amplitude. Row 5 shows the combination of each wave which has the same frequency as the wave in row 1 but a higher amplitude. The red line indicates a single cycle of each wave.

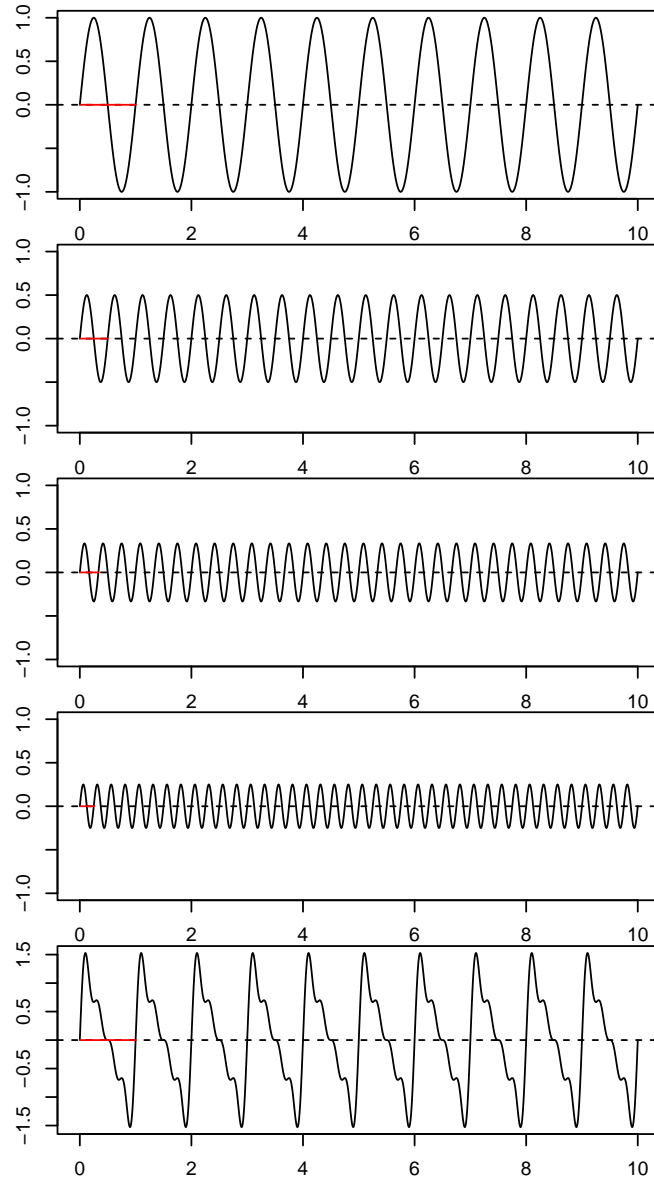


Figure 3.4: Rows 1-4 contain sine waves of increasing frequency with equal amplitude. Row 5 shows the combination of each wave which has the same frequency as the wave in row 1 but a higher amplitude. The red line indicates a single cycle of each wave.

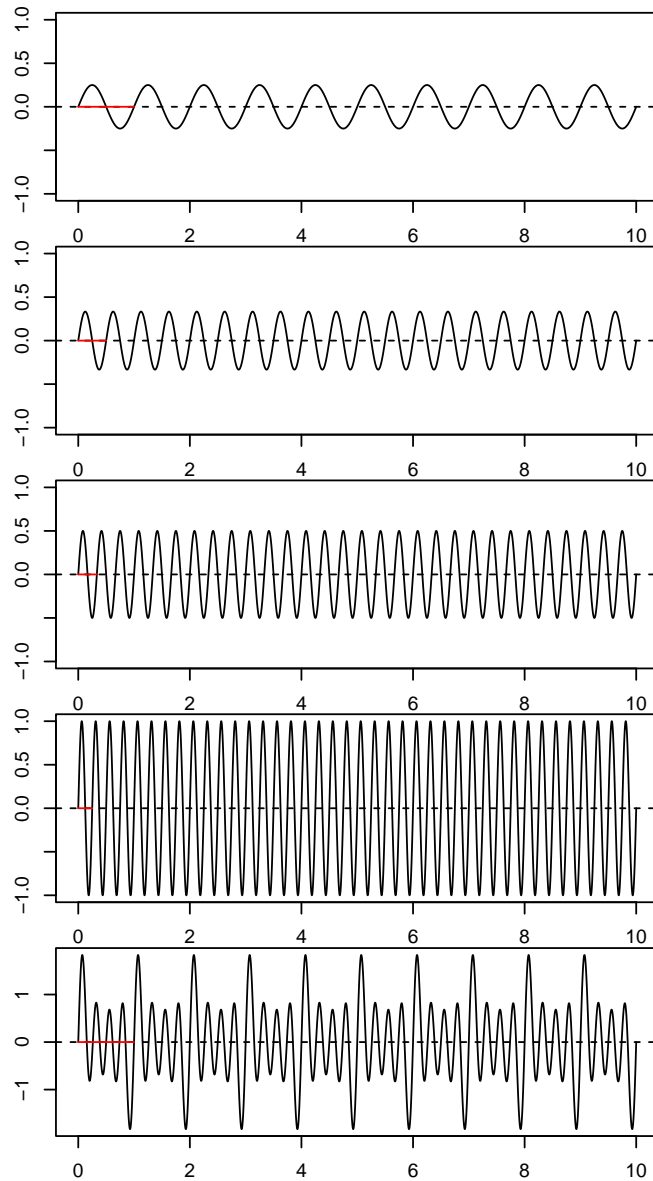


Figure 3.5: Rows 1-4 contain sine waves of increasing frequency with equal amplitude. Row 5 shows the combination of each wave which has the same frequency as the wave in row 1 but a higher amplitude. The red line indicates a single cycle of each wave.

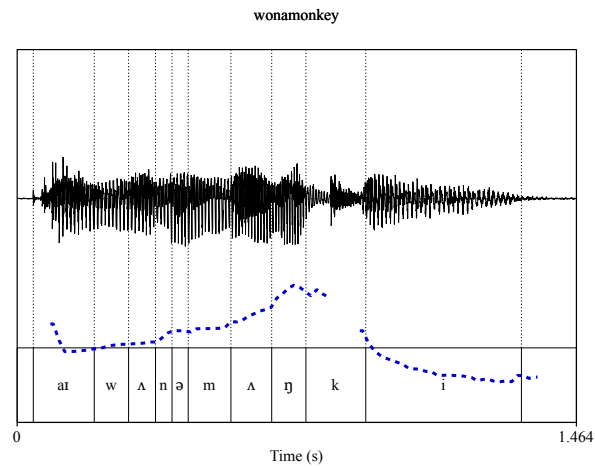


Figure 3.6: Wave form and pitch contour for the declarative phrase “I won a monkey”.

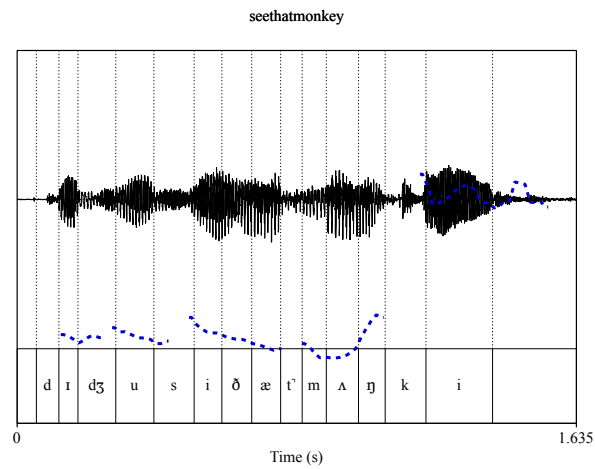


Figure 3.7: Wave form and pitch contour for the yes/no question “Did you see that monkey”.



# Lesson 4

## Lexical Tone + Contrast

In this lesson we will continue discussing f0, but this time focus on how it used for lexical contrast. Part of this lesson will therefore review the idea of contrast and how it relates to linguistic substance. There are many aspects within the speech signal that you will come to find out do not affect contrast and can therefore more freely vary. Diving into the acoustics lets us discover which aspects serve which purpose.

### 4.1 Contrast

We covered contrast a little bit in the review section, but let's dig into it a little bit deeper now. One way that we can tell if two sounds contrast in a language is by finding a **minimal pair**. A minimal pair is a set of two words that vary by only a single sound. As you will see later in this lesson, tone complicates things, but its the same general idea. In English we have the following sets of words:

[p<sup>h</sup>ɪt] ~ [bɪt]  
[p<sup>h</sup>ɪt] ~ [fɪt]  
[p<sup>h</sup>ɪt] ~ [k<sup>h</sup>ɪt]

The first pair are the words “pit” and “bit” which vary only in their first segment. We can actually be a little more fine-grained here and say that actually the only property in which these two sounds vary is voicing. The second pair is “pit” and “fit” which vary only in their degree of constriction: stop vs. fricative. Finally, we have “pit” and “kit” which vary

only in their place of articulation. On one level we now know that these segments are contrastive in the language, but we also know that English uses voicing, manner, and place in order to make contrasts.

Now, let's look at another set of "words":

[k <sup>h</sup> ɪs]	~	[hɪs]
[k <sup>h</sup> ɪs]	~	[χɪs]
[hɪs]	~	[χɪs]

The first row contains the English words "kiss" and "his". These form a minimal pair and therefore we know that the initial segments are contrastive in the language. But unlike the examples above, these sounds vary in both place of articulation and degree of constriction. What happens when we vary only one of those things at a time. The second row contains a "minimal pair" in the sense that only two sounds vary, but this is not really a minimal pair in English because the word [χɪs] does not exist. The third row is similar, only now we vary the place of articulation for the two fricatives. If you listen to both [hɪs] and [χɪs] being spoken, they probably both would sound like the English word "hiss" to you.<sup>1</sup> That doesn't mean you won't notice a difference in their acoustic properties, but rather the difference is not meaningful for linguistic contrast.

The minimal pair view of contrast requires us to stay at a relatively abstract level. It is useful when doing phonology because there we are interested in broad sound patterns, but in phonetics we are interested in detailed sound patterns. Therefore, our view of what is and what is not contrastive can be more fine-grained. As an example, let's consider the vowel [u] as it appears before the three voiceless stops in English. We can make pseudo minimal pairs:

[sut]	~	[sup]
[sut]	~	[suk]
[sup]	~	[sup]

The word [suk] is not a real English word, but it seems like it maybe could be. There aren't many [uk] sequences but they do show up in words like "puke" and "duke".

---

<sup>1</sup>You may also hear [χɪs] as "kiss" from time to time but that doesn't change the overall facts about it not being contrastive."

Up to now our focus has been on what *is* contrastive, but we also need to see what is *not* contrastive. Hopefully, everyone agrees that at some level of abstraction, the /u/ in these words is the same. But is it the same when it comes to their acoustic properties? It turns out it is not. The /u/ in “suit” is more front (meaning has a higher F2) than the /u/ in “soup” or [suk]. In figure 4.1, I plotted the F2 values at the midpoint of the vowel for five recordings each of the words above. You’ll see that there is no overlap between the F2 values for [sut] and the other two words.

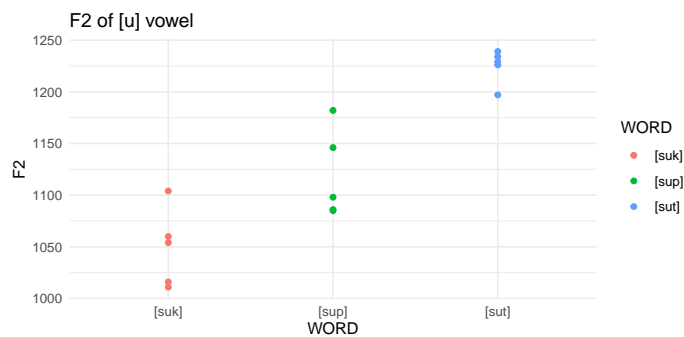


Figure 4.1: F2 values from five tokens each of the words [suk], [sup], and [sut].

Let’s think for a moment as to why the F2 is higher for the /u/ in “suit”. The only difference in these three words is the following consonant, therefore it’s a safe bet to assume that this is driving the difference. But does it make sense? If we think about the articulations involved, it becomes evident why this is the case. Since /u/ is a back vowel, it is articulated with the back of the tongue. For the consonant /t/, the articulation is made with the tongue tip. Even though these can move independently, they still have an effect on one another. The movement of the tongue tip towards the alveolar ridge in anticipation of the /t/ pulls the body of the tongue forward, therefore making the constriction closer to the front of the mouth. A closer constriction results in a shorter tube which in turns gives us a higher resonant frequency. We’ll return to tube models of the vocal tract in the next lesson.

The larger takeaway from this section is the following: contrast is important for figuring out the broad sound patterns of a language. Changes in major categories of linguistic sounds obviously results in drastic changes

in the acoustic properties of a sound. Fricatives and stops could not be more different in how they affect the acoustic signal: one results in energy dispersed throughout a large area of the spectrum (fricatives), the other results in energy being removed completely from a large area of the spectrum (stops). These are macro level changes and can alter the meaning of words. But we also looked at micro level changes like the way /u/ shows up in the acoustics before different types of stops. As phoneticians, we want to discover and capture this variation even though it does not result in phonological contrast.

## 4.2 Lexical Tone

In the previous section, contrast was defined such that two words varied in a single sound. But what about a language such as Mandarin that has the following words:

[ma <sup>55</sup> ]	‘mother’
[ma <sup>35</sup> ]	‘hemp’
[ma <sup>214</sup> ]	‘horse’
[ma <sup>51</sup> ]	‘scold’

All four of these words have the same segmental content, but vary in which tone they carry. That superscript numbers after the segmental material indicate the general shape of tonal contour. The numbering goes from 1-5 where 1 is a low tone and 5 is a high tone. So 55 refers to a high-level tone, 35 refers to a tone that starts of mid and goes high (mid rising), 214 refers to a tone that starts out low, goes a little lower, and then goes high(ish) (low rising), and 51 refers to a tone that starts out high and then goes low (high falling).

Figure 4.2 shows the pitch contour from Praat for these four words. The recordings come from [Ladefoged and Disner \(2012\)](#). As you can see, the correlation is not always 1-1. Especially for the low-rising tone, the end of the tone doesn’t really get to the “4” level, but it does maintain the expected profile of starting low(ish), going lower, and then rising.

Given this example of how tones can signify contrast, we really should update our definition of what contrast is. Rather than focusing on how a single “sound” can distinguish between two words, it’s probably better to be more neutral and say that single *property* varies between two sounds.

This can be a **segmental** level property such as place, manner, or voicing, but it can also be a **supersegmental** property such as tone. Supersegmental is the term used to describe aspects of a linguistic utterance “above” the level of the segment such as tone and stress.

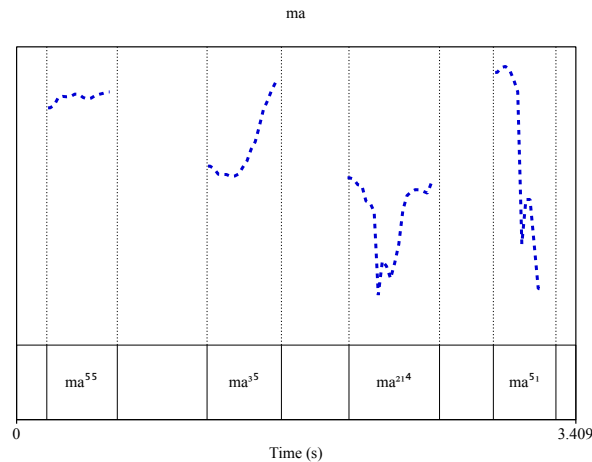


Figure 4.2: Pitch contours for the four Mandarin tones.

In some sense, tone is kind of like pitch where we think about it in relative terms. Typically, tones can be described by combining **H(igh)**, **M(id)**, and **L(ow)** tones in various ways. **Level tones** occur when the  $f_0$ /pitch of a tone stays constant throughout the entire tone bearing unit.<sup>2</sup> For example, the 55 high tone in Mandarin is a level tone because it stays high throughout the entire tone bearing unit. You may also see high tones represented as an acute accent over the vowel: [má].

In contrast to level tones are **contour tones** where the  $f_0$ /pitch of a tone changes over the course of the tone bearing unit. In general, contour tones come in two flavors: **rising tones** and **falling tones**. There are different combinations that can be made for each such a LH, LM, MH for rising and HM, HL, ML for falling tones. Another way you’ll see rising and falling tones represented is as [mǎ] and [mâ]. To fully understand this representation scheme you need to know that low level tones are repre-

<sup>2</sup>We won’t get too deep into the idea of a tone bearing unit here. This simply refers to what part of the word a tone attaches to. Typically, the tone bearing unit is a vowel, but specialists also refer to **moras**. If you are interested in learning more about this I can send you some readings.

sented with a grave accent [m̀à]. Therefore, [m̃ă] starts with a low tone and then goes high while [m̂â] starts with a high tone and then goes low. Of course, this is not a perfect scheme since there can be multiple types of rising tones. Plus what do you do when you have a mid tone? Well a mid tone is [m̄ā].

Believe it or not there are many other ways people use to represent tone. Personally, I like putting M,H,L over the vowel but I think this is my phonology bias sinking in. In general, it is fine to use whatever representation works best for you as long as you choose one and stick with it throughout all your work.

We'll end our discussion of tone by looking at one more language. The Ibibio language is spoken by the Ibibio people in Nigeria. Data come from the UCLA Phonetics Lab. In the following examples, we see bisyllabic words containing a combination of High and Low tones.

[ákpá]	[àkpá]
‘expanse of ocean’	‘first’
[ákpâ]	[àkpô]
‘square woven basket’	‘rubber tree’
[ákù]	[àkpà]
‘priest’	‘(small ant)’

Using just H & L we can represent the patterns in the following way:

HH	LH
HHL	LHL
HL	LL

Figures 4.3, 4.4, and 4.5 show the pitch contours for each of the words above. There is glottalization at the beginning of each word which causes the pitch tracker to add some high f0 points, you can ignore those. Focus on the parts labeled with L and H. You should notice that all the pitch contours roughly do what's expected. Many of the L tones have a downward slope, but notice that it's not quite as drastic as the actual falling tones in figure 4.4.

Unfortunately, just like with intonation, there is a lot more we could go into with tone. We really are just scratching the surface here. Realistically, you could have an entire 14 week course on just tone and intonation. This class requires a breadth rather than depth approach and therefore we only get to discuss the very basics.

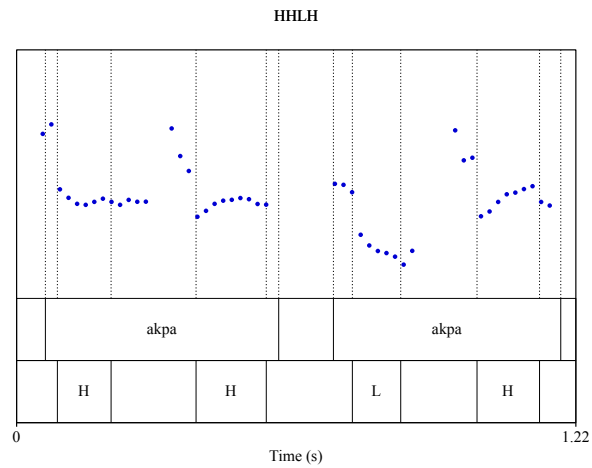


Figure 4.3: Pitch contours for HH and LH bisyllabic words in Ibibio.

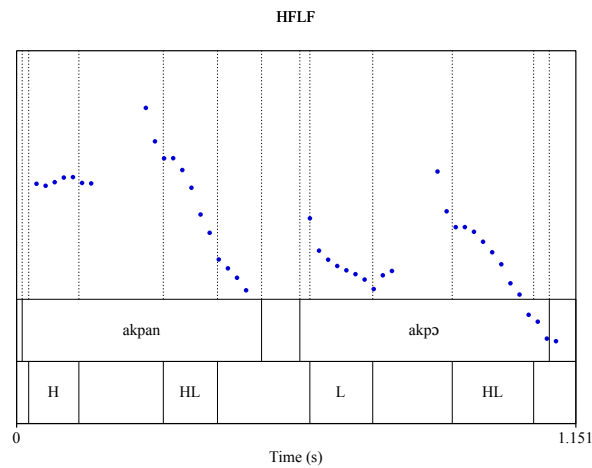


Figure 4.4: Pitch contours for HHL and LHL bisyllabic words in Ibibio.

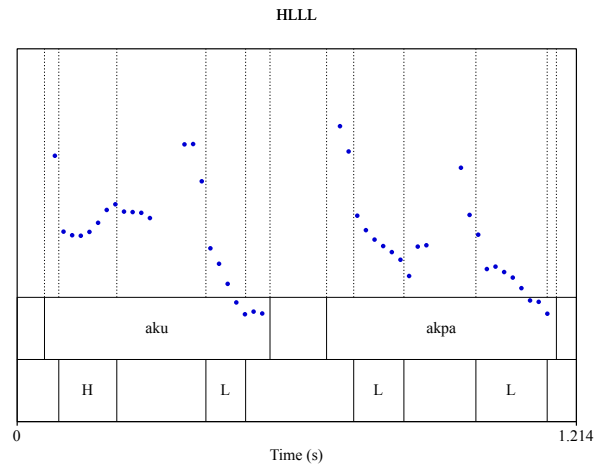


Figure 4.5: Pitch contours for HL and LL bisyllabic words in Ibibio.



## Lesson 5

# The Acoustic Theory of Speech Production

In this lesson we're going to go back to discussing sine waves and their theoretical implications for speech. Previously we have discussed sine waves and their relationship to fundamental frequency. This time, we are going to discuss their relationship to resonance. You will see that with some basic assumptions, we can reliably predict the formant patterns that we see in vowels.

### 5.1 Filters and Resonators

What is a filter? Well, depending on your background, you might give a different answer. If you're a data scientist, you might say that it's something you use to smooth out your data and handle outliers. If you're a visual artist, you might say that it's something you use to change the way your picture/film looks. If you work in audio, you might say it's something you use to remove certain frequencies from the audio signal. This last case is the one is the most relevant to us, but before we go more in depth I'd like to discuss an even more general case of filtering.

Imagine you are at the beach. A common beach toy is a sand sifter. Sometimes it's shaped like a handheld shovel, sometimes it's a big circle, but in both cases there are holes that allow sand to flow out the bottom while capturing rocks and other items that are too big to pass through the holes. You can even get sifters of various sizes and pass the sand through

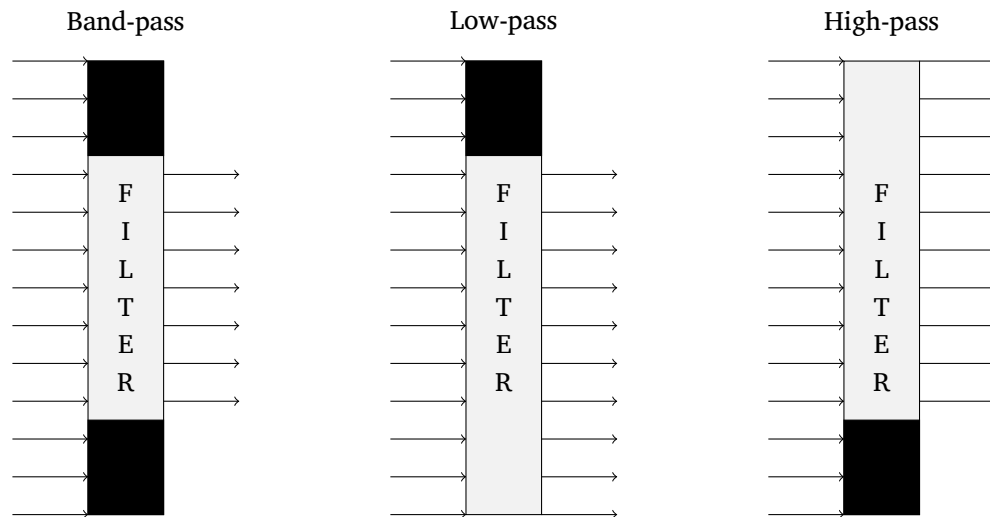


Figure 5.1: Visualization of band-pass (left), low-pass (center), and high-pass (right) filters

bigger ones and then smaller ones so that you're only left with the finest grained pieces. This is exactly what a filter does.

A **filter** is something that takes some type of input (say, sand from a beach) and allows some part of it through (the actual sand) and blocks others (the rocks, shells, etc...). In the audio world, a filter takes an audio input and allows certain frequencies to pass through but blocks (or reduces) others. A **band-pass** filter is a type of filter that allows a specific band of frequencies to pass through unaffected while either completely blocking, or **attenuating** frequencies outside the defined range. Attenuation means to alter the strength at which the non-band frequencies pass through. There are special cases of band-pass filters called **high-pass** filters where the band ranges from some frequency to infinity, and **low-pass** filters where the band ranges from some frequency to 0. Figures 5.1 and 5.2 show visual examples of these different types of filters. In figure 5.1 there is no attenuation, while in figure 5.2 there is linear attenuation of the frequencies not in the critical band.

Another special type of filter is a **resonator**. A resonator takes a very narrow range of frequencies and amplifies them instead of reduces them. Alternatively, it has a very narrow band and full attenuation of the frequencies outside the critical band. As you may remember from 201 and

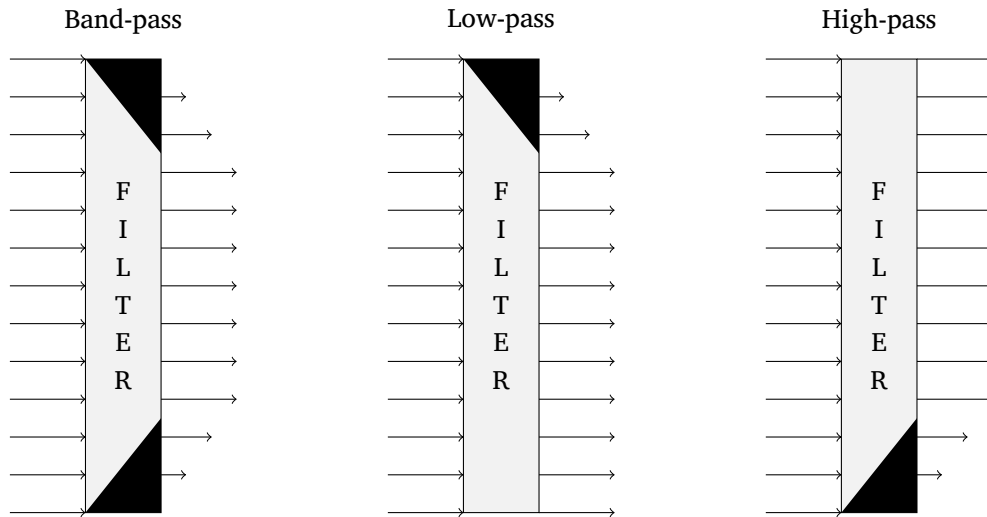


Figure 5.2: Visualization of attenuated band-pass (left), low-pass (center), and high-pass (right) filters.

the review lessons, **resonance** is an important aspect for speech analysis. The reason we care about filters and resonators is because the vocal tract as a resonator for certain frequencies, especially vowels. Figure 5.3 shows two narrow band-pass filters that act as resonators that amplify certain frequencies and suppress all others.

## 5.2 Tube Models

Now that we have an idea about filters and resonators, let's move on to discuss how objects work as natural resonators. The object that we'll use is a tube because we can somewhat realistically break the vocal tract up into a series of tubes. Therefore, if we understand how tubes act as resonators, we can then see how that relates to our vocal tract acting as a resonator.

In an open tube, there are multiple frequencies where the corresponding sine wave creates what is called a **standing wave**. A standing wave is one in which the location of maximum displacement does not change when the wave reflects. You can think of these as the waves that “fit” the length of the tube such that there is a point of minimal displacement at the open end of the tube. This creates **anti-nodes** at the point of maximal

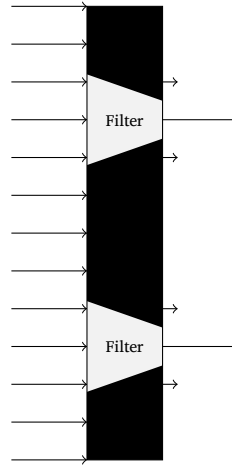


Figure 5.3: Visualization of multiple narrow band filters acting as resonators

change and **nodes** at the point of minimal change. One reason for this is because the air pressure at the open end of the tube is fixed to the atmospheric pressure around it, while the pressure within the tube changes due to the sound moving through it.

Figure 5.4 shows the three lowest frequency standing waves for a tube open at both ends. Notice that at both endpoints, the wave crosses the 0 point (represented with a dashed line). There is also a mathematical relationship between the frequency/wavelength that will create a standing wave and the length of the tube. Another term for the different standing waves are the **harmonics** which are measured in Hz. The harmonics  $f_k$  for a tube open at both ends is  $\frac{c}{2l}$  where  $c$  is the speed of sound (approximately 35,000 cm/s) and  $l$  is the length of the tube. We can also figure out the length of the tube with the formula  $\frac{kc}{2\lambda_k}$  where  $c$  is the same and  $k$  is the  $k^{th}$  harmonic.

We are also interested in a tube open at only one end. In fact, we can think of the vocal tract as a tube closed at one end (the larynx) and open at the other end (the lips). Just like with a tube open at both ends, there is a mathematical relationship between the frequency/wavelength that will create a standing wave and the length of the tube. Tubes that are open (and, in fact, closed) at both ends are called half-wave rectifiers, but tubes open at only one end are called quarter-wave rectifiers. Figure 5.5 shows

the three lowest frequency standing wave for a tube open at one end and closed at the other. The harmonics  $f_k$  for a tube open at one end and closed at the other is  $\frac{(2n-1)c}{4l}$ .  $c$  and  $l$  are the same as before. Similarly, we can figure out the length of the tube with the formula  $l = \frac{(2n-1)c}{4f_k}$ . You will use this in Lab 3 to try to determine the length of your vocal tract.

A classic result from the acoustic theory of speech production, is that we can determine the expected F1, F2, and F3 values of schwa if we know the length of the vocal tract. This is because schwa is said to have a *neutral* articulation where the tongue is mostly in a rest position. Suppose the vocal tract length is 17.5 cm. Then we can directly determine the first three formants by just plugging the right values into the formula,

$$F1 = \frac{(2 \cdot 1 - 1) \cdot 35,000}{4 \cdot 17.5} = \frac{35,000}{70} = 500$$

$$F2 = \frac{(2 \cdot 2 - 1) \cdot 35,000}{4 \cdot 17.5} = \frac{105,000}{70} = 1500$$

$$F3 = \frac{(2 \cdot 3 - 1) \cdot 35,000}{4 \cdot 17.5} = \frac{175,000}{70} = 2500$$

If we take 17.5 cm to be an average vocal tract length, this means that schwa should on average have an F1 around 500 and an F2 around 1500.

## 5.3 Vowels as multiple tubes

We end this lesson with a discussion on how the vocal tract can be modeled in most cases as a series of multiple tubes. Low vowels can be modeled as two tubes that are closed at one end and open at the other. There is a back tube that is thought to be smaller in diameter, that is closed at the larynx and open at the junction with a larger front tube. The front tube is closed at the junction and open at the lips. Both tubes will resonate depending on their length. This will affect which tube corresponds to F1 and which corresponds to F2. A **nonogram** is shown in figure 5.6. The low vowel [a] involves a constriction around the 8cm mark which is where we expect F1 to be its highest (thus being the lowest vowel) and F2 being its lowest (thus being the most back vowel) just as it appears in the vowel chart.

High vowels are a little bit more complicated because it also involves a front and back tube with an additional small intermediate tube that creates

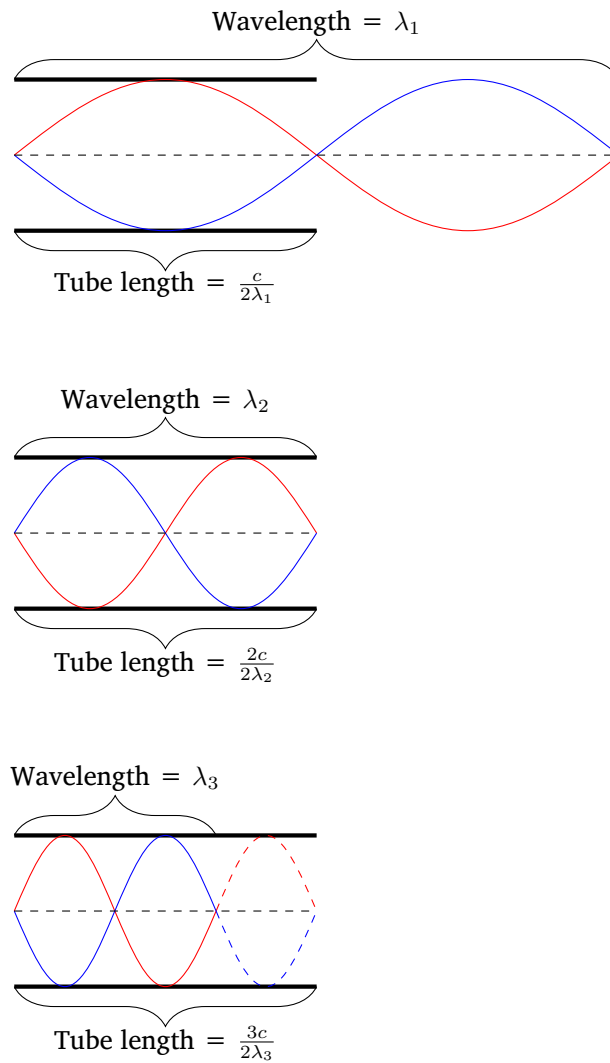


Figure 5.4: Three three lowest frequency standing waves for an open tube

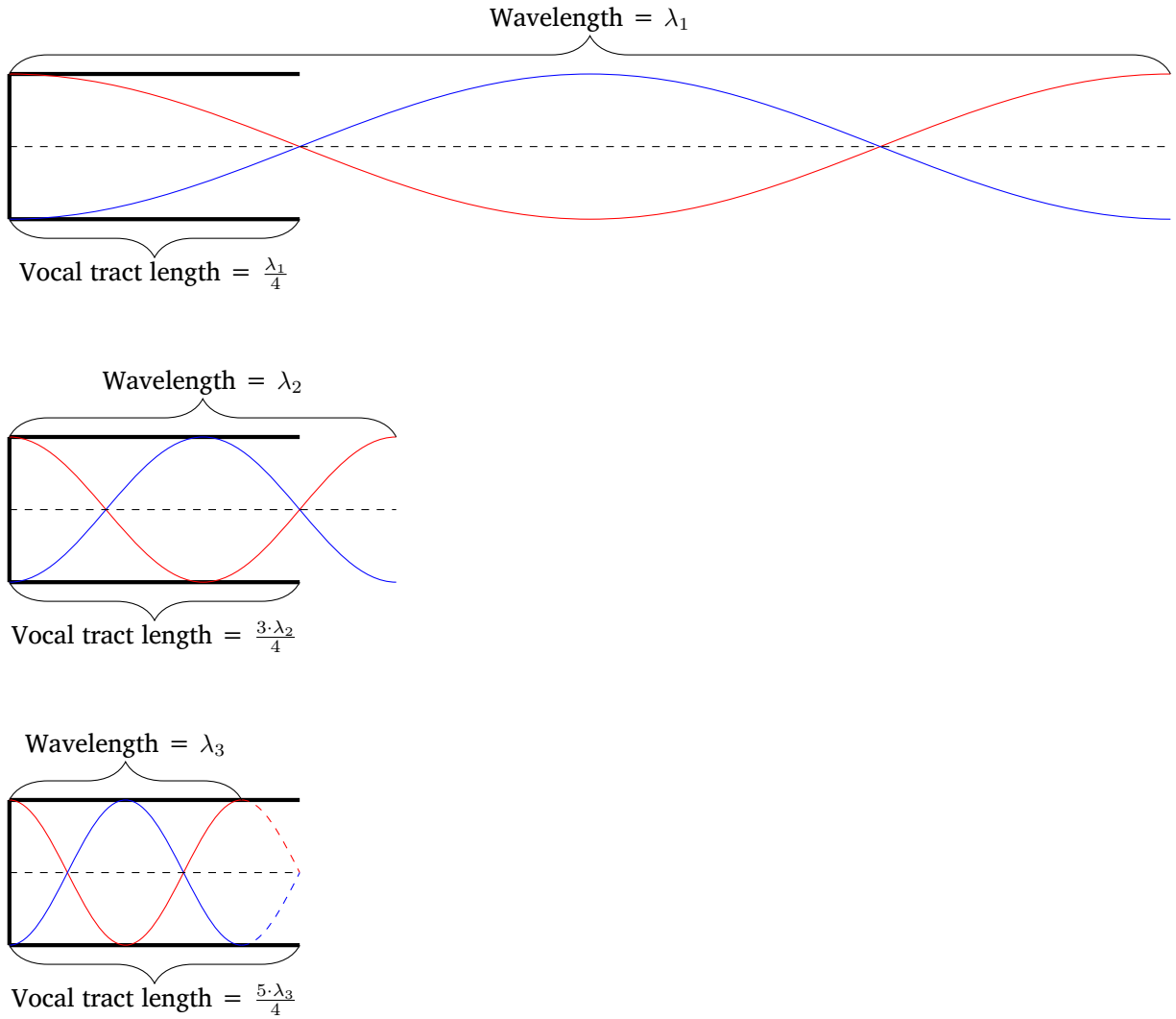


Figure 5.5: The three lowest frequency pressure waves that meeting the criteria for resonance.

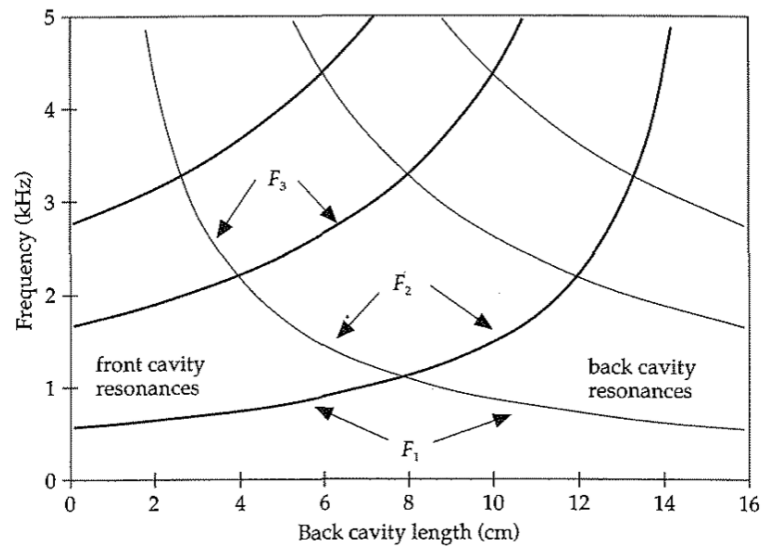


Figure 5.6: Nomogram for low vowel resonances. Figure taken from [Johnson \(2011\)](#).

what is called a **helmholtz resonator**. This type of resonator is the same thing as what happens when you blow across the top of a pop bottle. You can calculate the frequency at which it will resonate based on the back and intermediate tube cross-sectional areas and their lengths. The back tube for high vowels is closed at both ends and the front vowel is closed at one end and open at the other. The nomogram for these types of vowels is shown in figure 5.7 Notice, the F1 stays relatively low. This is directly influenced by the helmholtz resonator of the intermediate tube.

The large takeaway from this lesson is that we can model vowels in speech directly based on viewing the vocal tract as a series of tubes. Because we know what types of frequencies will resonate in tubes, we can predict which frequencies will resonate when we configure the vocal tract in specific ways. Further more, we can think of the act of producing speech as producing a source (the vibration of the vocal folds) that is then filtered (by changing the configuration of the vocal tract). Recall that resonators are a type of narrow band filters that amplify certain frequencies and dampen others.

For a bonus insight into tube models of the vocal tract, watch [this video](#).



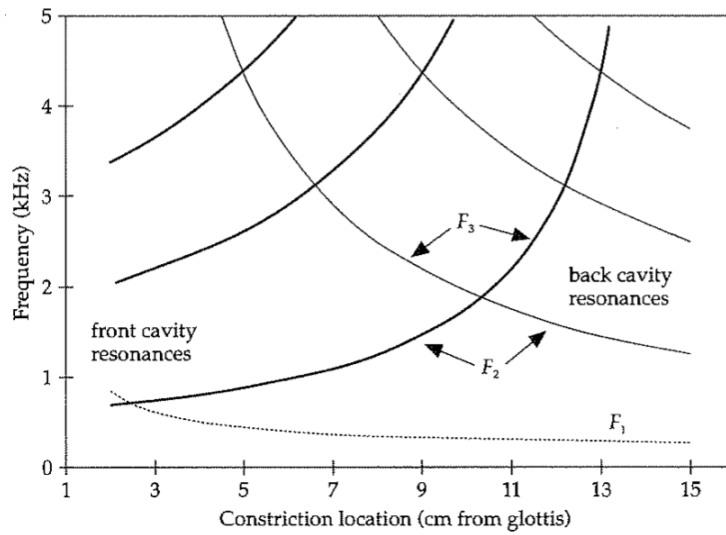


Figure 5.7: Nomogram for front vowel resonances. Figure taken from [Johnson \(2011\)](#).



# Lesson 6

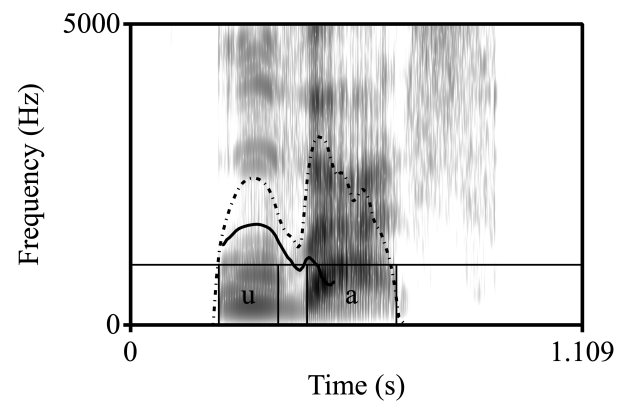
## Acoustic Correlates of Stress

Previously in this course we talked about intonation, which is closely related to stress. In this lecture we will focus more on lexical stress and its acoustic consequences.

### A background on stress

**Stress** is a property of syllables, but largely we think about the vowel as carrying the acoustic information that signifies stress. This is a simplifying assumption, but it works pretty well for analysis. Of course, there are syllabic consonants in English and even languages such as Imdlawn Tashlhiyt Berber that has full words with no vowels, but our focus will be on how stress shows up on vowels.

Stress is a relative property pertaining to syllables. There are typically four parts of the acoustic signal that signify stress: intensity, duration,  $f_0$ , and  $F1/F2$ . Stressed syllables are usually louder (higher intensity), longer (greater duration), higher pitched (higher  $f_0$ ), and more peripheral ( $F1/F2$  pushed towards the edges). Some of these properties are often thought to be related to one another. For example, a longer duration means the articulators have more time to reach their goal, thus resulting in the  $F1/F2$  being more peripheral. It should also be noted that languages use this properties in various ways. Some languages like Japanese and Swedish rely a lot on  $f_0$  and have been called **pitch-accent** languages for this reason.



# Lesson 7

## Voicing

In this chapter we will discuss three acoustic properties that correlate with contrastive voicing in obstruents. **Obstruent** is a commonly used name for the group of consonants including stops, fricatives, and affricates. Recall that in the IPA chart, these are the only sounds that contrast for voicing. We already know that one way to tell if a segment is voiced on a spectrogram is to look for the “voicing bar” of low frequency energy. Today, we are going to focus on three other cues for voicing: voice onset time (VOT), duration of preceding vowel, and percentage of segment voiced.

### 7.1 Voice Onset Time

So far, we have been calling the difference between, say, /p/ and /b/ in American English a difference in *voicing*. This is only a half truth. In the first part of this lesson we will focus on word-initial stops. Recall also that the articulatory correlate of voicing is vocal fold vibration.<sup>1</sup> The acoustic consequence of vocal fold vibration is low frequency energy, colloquially called the “voicing bar”, to show up on a spectrogram.

It turns out that for most American English speakers, the voicing bar is completely absent from the closure portion of the stop while word-initial “voiced” stops are being articulated. You may remember from LIN 201 or a previous class that the voicing difference in American English has more to do with aspiration than it does with voicing. **Aspiration** is the

---

<sup>1</sup>More specifically, it requires properly tensing the vocal folds and producing a consistent airstream in order to cause the vocal folds to vibrate via the bernoulli effect.

noisy voicelessness that appears after a voiceless stop in certain positions. Hopefully you remember that one of those positions is the word-initial position.

In the 1960's researchers discovered that there was a single acoustic cue that reliably encoded the voicing contrast in word-initial stops in various languages. This acoustic cue is called **voice onset time** and it is the difference in time between the onset of voicing and the time of the release burst. A single formula can be used to calculate VOT if you know the time when voicing starts ( $t_{voi}$ ) and the time of the release burst ( $t_{bur}$ ):

$$\text{VOT} = t_{voi} - t_{bur}$$

In general, there are three types of VOT values we can get. If voicing starts at a later time than the burst, we get *positive* VOT. If voicing starts at an earlier time than the burst, we get *negative* VOT. If voicing and burst happen at the same time we get 0 VOT.

0 VOT is basically impossible. In general, the continuum of VOT is broken up into three regions: negative VOT, short lag positive VOT, and long lag positive VOT. Negative VOT corresponds to what is sometimes called **pre-voicing** or **full voiced stops**. This is when there is voicing that occurs during the stop closure. Short lag positive VOT typically corresponds with **voiceless unaspirated stops**. This is essentially the 0 VOT category. Finally, Long lag positive VOT typically corresponds with **voiced aspirated stops**.

Different languages use these categories in different ways. In English, our “voiced” stops at the beginning of words are actually voiceless unaspirated stops and our “voiceless” stops at the beginning of words are actually voiceless aspirated stops. This points to the fact that the contrast for a [voice] *phonological* feature is not always phonetically realized as a contrast in voicing per se. As mentioned above, the acoustic contrast really has to do with the aspiration and not the voicing. Figures 7.1 and 7.2 below show the waveform and spectrograms for the English words “peach” and “beach”. They have been marked with vertical lines showing the burst point, as well as the onset of voicing. In figure 7.1, the initial stop is a /p/ which is phonetically realized as an aspirated voiceless stop. If we measure the time between the release burst and the onset of voicing, we get a VOT time of 71ms. In figure 7.2, the initial stop is a /b/ which is phonetically realized as an unaspirated voiceless stop. Notice, there is no voicing

Negative VOT	Short Lag VOT	Long Lag VOT	Example Language
✓	✓		Spanish
✓		✓	Swedish
	✓	✓	English
✓	✓	✓	Thai

Table 7.1: Different VOT category combinations and example languages that uses that combination to form contrasts in stops.

bar on the spectrogram during the closure portion preceding the release. If we measure the time between the release burst and the onset of voicing, we get a VOT time of 10ms.

In general, there are four types of languages we could come up with that use the VOT categories to form a contrast in their language. Table 7.1 summarizes the four possible category combinations and lists an example language for each category. Spanish contrasts its stops using a negative VOT for its voiced stops and a short lag VOT for its voiceless stops. Sometimes languages that use this contrast are called **voicing** or even **true voicing** languages. As discussed above, English contrasts its stops using short lag VOT for its “voiced” stops and long lag VOT for its “voiceless” stops. Languages like English are often called **aspirating** languages which contrasts with the voicing languages like Spanish. Languages that contrast their stops with negative VOT for voiced stops and long lag VOT for their voiceless stops are rare, but not unheard of. Swedish is one language that does this. Finally, we can imagine a language that doesn’t just have a single contrast in stops, but in fact uses all three. Thai is a language that has negative VOT, short lag VOT, and long lag VOT stops. In fact, Thai was one of the languages studied in the original VOT paper.

We end this section with a reminder that VOT is not a perfect measurement. This isn’t to say that we should just get rid of it, but we should be aware of its limitations. For example, Hindi has a four-way laryngeal contrast:

koʈ    *fort*  
 k<sup>h</sup>oʈ    *flaw*  
 goʈ    *border*  
 g<sup>fi</sup>oʈ    *grind*

Specifically, the language has what may be called a “voiced aspirated”

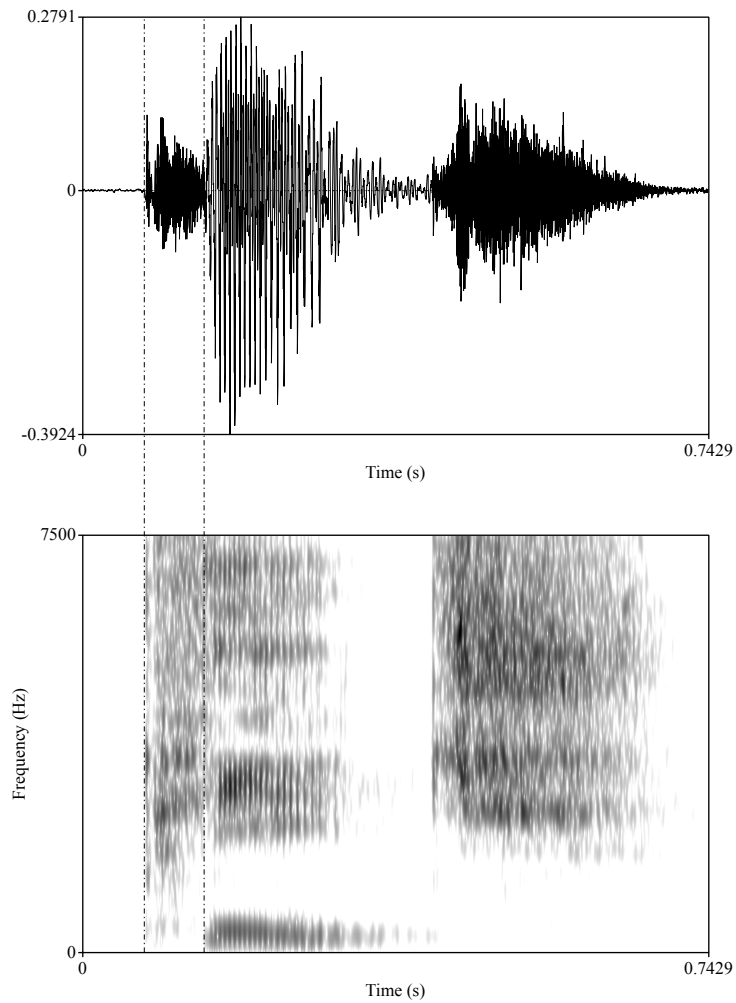


Figure 7.1: Spectrogram and waveform for the English word “peach”. The release burst occurs at the 73ms mark and the onset of voicing occurs at the 144ms mark. The overall VOT for the onset /p/ is therefore is 71ms.



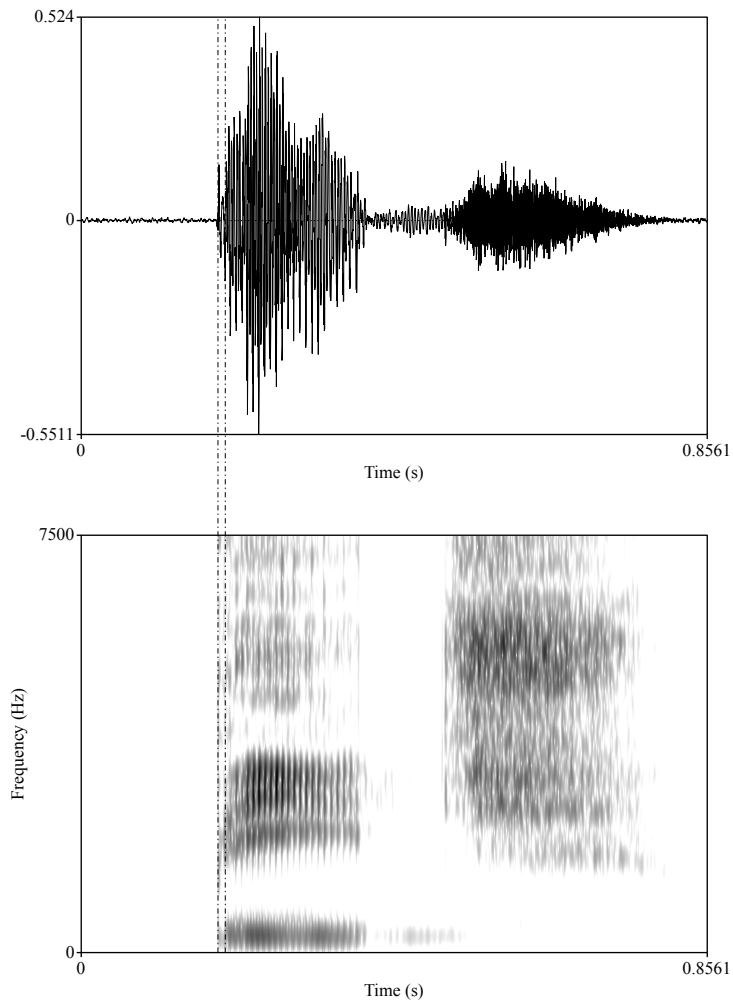


Figure 7.2: Spectrogram and waveform for the English word “beach”. The release burst occurs at the 187ms mark and the onset of voicing occurs at the 197ms mark. The overall VOT for the onset /b/ is therefore is 10ms.

stop. This is unable to fit into the three-way distinction of VOT as defined above since voicing and aspiration occur simultaneously. It has also been shown that there is a lot of variation within the long lag VOT category. [Cho and Ladefoged \(1999\)](#) discusses the aspiration difference and [Abramson and Whalen \(2017\)](#) is a 50 year anniversary retrospective article that celebrates the successes, while also discussing the shortcomings, of VOT as first defined in [Lisker and Abramson \(1964\)](#)

## 7.2 Other cues for Voicing

VOT works fine for stops in initial position, but what about other obstruents in other non-word initial position? There are two other cues we will briefly discuss in relation to fricatives at the end of words. Note, these cues can be used in non-word final position, but its easiest to discuss in that position. One of the two cues is the length of the preceding vowel. In English, voiced sounds have longer preceding vowel durations than voiceless sounds. Another cue is how long voicing persists going into the sound. These cues are especially crucial for voiced fricatives at the end of words.

Producing voiced fricatives is actually very difficult and because of this, voiced fricatives are relatively rare cross-linguistically. The reason they are difficult to produce is because you have two competing goals. To make a fricative you need to produce high speed turbulent airflow. This is easy when the vocal folds are spread and not impeding the airflow in any way. But when making a voiced sound, the vocal folds block and reduce the speed of the airflow, thus making it hard to maintain the high speed needed to maintain turbulence. For this reason, especially at the end of words, fricative often appear “voiceless”. Due to the lack of voicing, we need to look at other cues to see in what way “voicing” appears.

Figures [7.3](#) and [7.4](#) show spectrograms for the English words *peace* and *peas*. They include vertical lines showing the start and end point of the vowel, as well as the start and end point of the fricative, and in the middle of the fricative interval a line indicating where voicing ends. For *peace*, the vowel starts around 251ms and goes until 422ms giving it a duration of 171ms. In *peas*, the vowel starts around 219ms and goes until 519ms giving it a duration of 300ms. This matches the characterization above that vowels are longer before voiced sounds than voiceless sounds.

Now, if we look at the duration of voicing into the fricative, we also

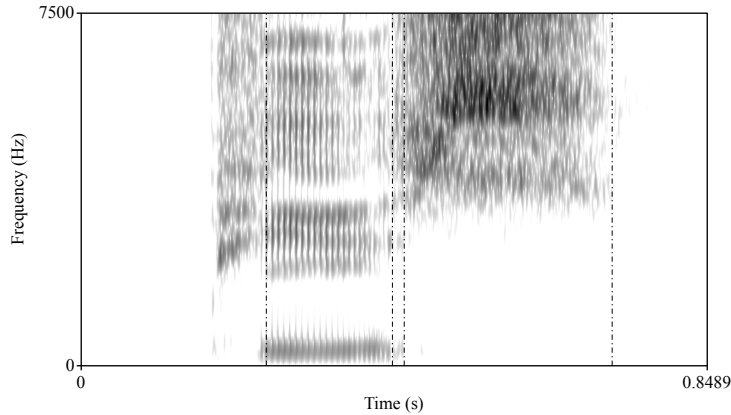


Figure 7.3: Spectrogram for the English word “peace”. Vertical lines indicate, in order, start point of vowel, end point of vowel/start point of fricative, offset of voicing, end of fricative.

see that voicing persists longer for the voiced /z/ than the voiceless /s/. In *peace*, the vowel ends around 422ms and the voicing persists until 438ms. So there is around 16ms of voicing at the start of /s/. In *peas*, the vowel ends around 519ms and the voicing persists until 580ms. Here, there is around 61ms of voicing at the start of /z/. In both cases, there is no apparent voicing in the acoustic signal for the majority of the fricative, but it is much more persistent for the sound that we call voiced than the one we call voiceless. Instead of raw values, sometimes people calculate ratios. If we were to do this, the difference would become even more apparent.

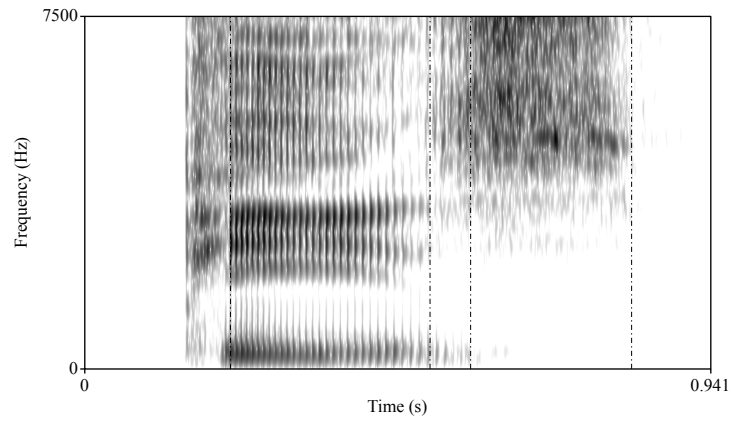


Figure 7.4: Spectrogram for the English word “peas”. Vertical lines indicate, in order, start point of vowel, end point of vowel/start point of fricative, offset of voicing, end of fricative.

## Lesson 8

# Acoustic Correlates of Place of Articulation

Recall that consonant sounds can roughly be described based on their place of articulation, manner of articulation, and voicing. Manner of articulation further splits into degree of constriction, airflow location, and nasalization. So far in this course we have talked about the acoustic correlates of voicing and the acoustic correlates of various degrees of constriction. We will not discuss the acoustic properties of airflow location and nasalization in this course, but we will discuss ways to measure those properties outside of acoustic analysis in the last week of classes.<sup>1</sup> That leaves us with degree of constriction as the final consonant parameter that we need to talk about. In this lesson we will focus on the acoustic properties of stops and fricatives that provide cues for place of articulation.

### 8.1 Formant Transitions & Stop Bursts

Imagine you are trying to produce a stop in between two vowels (e.g., speaking the word *about*). This would require you to move the necessary articulators close together and then release them. Therefore, it would start with a period of shutting where your articulators are moving into place, then have a period of closure where no air escapes through the mouth,

---

<sup>1</sup>The acoustic analysis relies on tube models. In fact, laterals and nasals are acoustically quite similar. If you are interested in learning more, [Johnson \(2011\)](#) has a nice chapter you can read.

then finally a period of release where the articulators are moving away from each other. During the transitional periods, the shape of the vocal tract is changing, thus changing the frequencies that resonate. This causes the formants at the edge of stops to have characteristic transition patterns.

In the 1950's<sup>2</sup>, it was discovered that different places of articulation had different formant transition patterns. Experiments using synthetic stimuli showed that listeners could accurately predict the place of articulation after only hearing formant transitions. If we take a moment to think about it, it makes a lot of sense that we need to use something beside the stop articulation itself to figure out the place of articulation. You should remember at this point that the acoustic correlate of a stop is silence. This is a very robust cue for the *manner of articulation*, but gives us absolutely nothing in regards to the place of articulation. But furthermore, this highlights the *dynamic* nature of speech production and perception. While it is true that we have a static target we are trying to achieve when making a stop (bringing a pair of articulators into complete contact), the actual implementation of reaching the target unfolds over time. There's no reason why we shouldn't use this knowledge to our advantage in both analysis and everyday speech perception.

Our focus will be on the first three formants. Largely F2 is the formant that is most informative during formant transitions. In general F1 decreases going into a stop for all places of articulation. As for F3, we really only care about its behavior for stops with a velar place of articulation. So let's break down the transition pattern for bilabial, alveolar, and velar stops. When a bilabial stop is being made, all the formants of a vowel drop rapidly going into the stop and rise rapidly coming out of it. When a velar stop is being made, F2 and F3 come together (sometimes called the "velar pinch") during the closure portion. That means they start apart and come together going into a velar stop, and start together and spread apart coming out of a velar stop. When an alveolar stop is being made, F2 moves to around 1800-2000 Hz. This means that sometimes it rises (for back vowels), sometimes it lowers (for very front vowels), and sometimes it stays pretty level.<sup>3</sup> Returning to F2, in general the value at the onset/off-

---

<sup>2</sup>Delattre et al. (1955)

<sup>3</sup>When trying to identify what the place of articulation for a stop is on a spectrogram, it's useful to look for all the formants lowering, then look for whether or not F2 and F3 come together, and if neither of those occur then you can be fairly confident it's an alveolar stop.

set of the vowel (the part “connected” to the stop) should be highest for a velar stop, lowest for a bilabial stop, and in the middle for an alveolar stop.

Figure 8.1 shows the sequence [aga]. Pay close attention to F2 and F3. While they don’t completely come together in the spectrogram, it should be clear that they are moving towards one another in a pinching fashion. Imagine continuing the trajectory just a little bit into the stop portion. If that were the case then the two formants would clearly come together. Figure 8.2 shows the sequence [ibi]. Here, we see that all the formants drop drastically where the closure of the stop occurs. In this case, we even see F4 drop a little bit. Finally, figures 8.3 and 8.4 show the sequences [ada] and [idi]. In the case of [ada], F2 rises because it is a back vowel and needs to get to that 1800-2000 Hz range. On the other hand, the F2 in [idi] lowers since it starts off higher than the target.

This part ends with figure 8.5 which shows the synthetic formant transitions originally used to test the hypothesis that these transitions were an acoustic correlate of place of articulation. As you can see, the F1 pattern is the same all throughout the different places of articulation. It rises coming out of the stop (all of these transitions are for a CV sequence). In the top row we see that F2 also is drops going into the stop (coming out) regardless of where it starts. In the bottom row we see F2 rises going into the stop regardless of where it starts. In the middle row we see that there is a consistent target right below 2 kHz (2000 Hz) which causes F2 to either rise or fall depending on where it begins. As a quick reminder, its important to remember that while we have an idealized understanding of what *should* happen, it’s not what does always happen. You may find yourself looking at formant transitions and being dead certain it’s one thing when it turns out to be something completely different.<sup>4</sup>

As a transition into talking about place of articulation in fricatives, let’s talk about one other aspect of stop acoustics that may cue place of articulation: the burst. A burst is a very quick bit of “noise”. In passing I have mentioned that we can use tube models to predict the general location of the noise in fricatives. Well it turns out we can do the same thing for stop bursts. If we model the area in front of the constriction as a single tube, then we can use our general knowledge of tube models to say that the loca-

---

<sup>4</sup>Say for example when you’re teaching this class how to read spectrograms without listening to them.

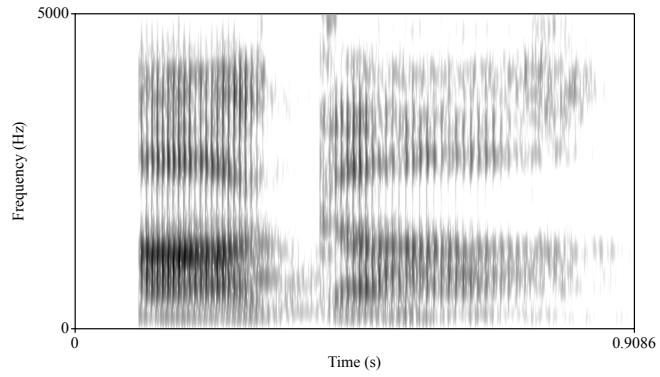


Figure 8.1: Spectrogram for [aga].

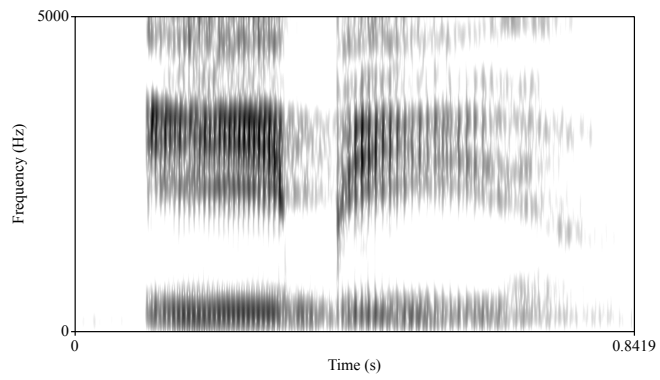


Figure 8.2: Spectrogram for [ibi].



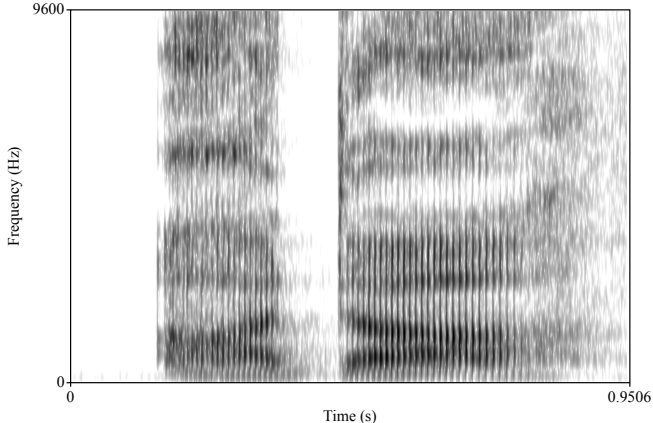


Figure 8.3: Spectrogram for [ada].

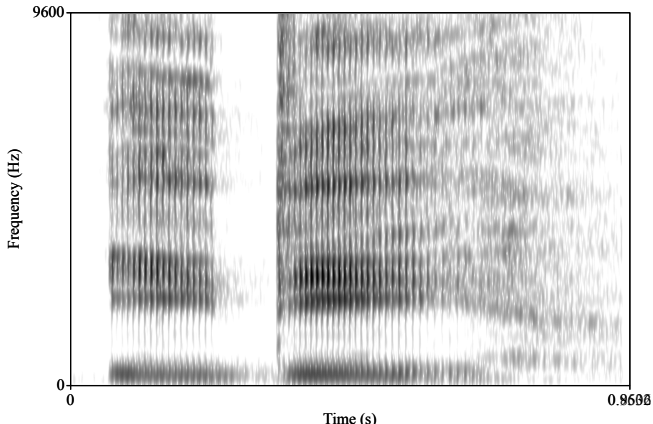


Figure 8.4: Spectrogram for [idi].

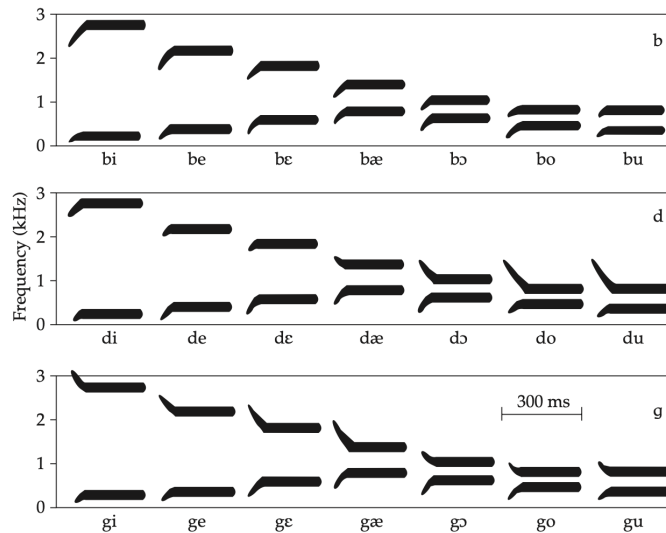


Figure 8.5: Synthetic formant transitions used to cue place of articulation as shown in [Johnson \(2011\)](#) and adapted from [Delattre et al. \(1955\)](#).

tion of the energy for a velar stop should be lower than it is for an alveolar stop. Why? Well, remember that if a constriction is made further back in the vocal tract, then there is a longer “tube” in front of it. Consequently, *lower* waves will cause it to resonate. What about bilabial stops? When these are made, there is no “tube” in front of it. In this case, we expect to see no shaping of the source signal and therefore expect a very even distribution of the noise at all frequencies. In the next section we’ll discuss fricatives more carefully and look into what are called spectral moments.

## 8.2 Spectral Moments in Fricatives

As was the case with stop, formant transitions, we can also look at the F2 locus at the onset/offset of the vowel to get an idea of the place of articulation. A general rule is that as the F2 at the onset of the vowel becomes higher in frequency, the further back the articulation being made is. Figure 8.6 shows this using the English words, *fie*, *thigh*, *sigh*, and *shy*. The arrows in the figure point to the F2 value when the vowel starts and list the frequency in Hz at that point. As you can see, F2 increases as we

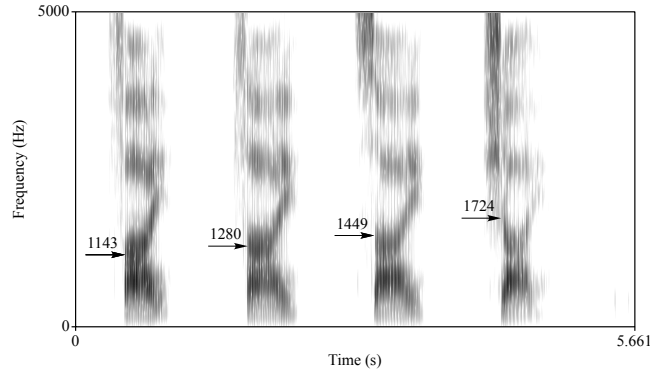


Figure 8.6: Spectrograms for English words *fie*, *thigh*, *sigh*, and *shy*. Arrows mark the onset of F2. Adapted from Ladefoged and Johnson (2014).

go further back in the mouth.

Next, we turn to the tube model analysis. This was mentioned in the acoustic review chapter, and I repeat the figure below as figure 8.7. Hopefully the “tube length” story is becoming familiar and you see the directly relation between constrictions being made further back in the mouth and which frequencies have more energy. The more back the constriction, the lower the noise concentration. You may notice that in both the F2 loci above and in the location of the energy, [f] and [θ] are very similar. Next week when discuss speech perception we will see that this leads to a lot of perceptual confusion.

We end this section on place of articulation in fricatives with a mention of what are sometimes called spectral moments. These are the **central of gravity (weighted average)**, **standard deviation**, **skewness**, and **kurtosis**. These terms come from probability theory and are ways to describe distributions, but they have been found to be useful measures for fricatives and other consonant sounds in language. The central of gravity provides an idea of where the energy is concentrated. This is one dimension that is closely related to what we have been talking about so far with the tube analysis. The other dimension is skewness which refers to whether or not the energy is skewed towards the higher or lower frequencies. The standard deviation tells us how spread out the energy is and kurtosis tells us how strong of a peak the sound has. Fricatives like [s] and [ʃ] typically have stronger peaks than [f] and [θ].

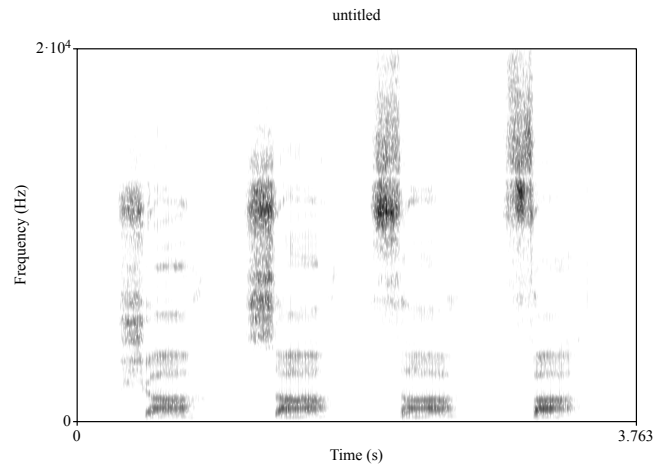


Figure 8.7: Spectrograms from left to right: [ʃa], [sa], [θa], [fa]

When taken as a whole, these four measurements provide a good way to distinguish place of articulation. Below, I provide figures of various **Beta** distributions to give examples of skewness and kurtosis. Figure 8.8 has three distributions: one with no skew (purple), one with positive skew (red), and one with negative skew (blue) Figure 8.9 has three distributions: one with noticeable peak and the highest kurtosis (purple), one that is relatively flat without a well defined peak and the lowest kurtosis (blue), and one in between the two (red).

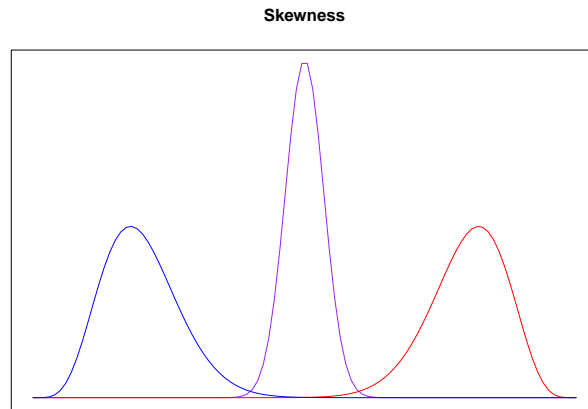


Figure 8.8: Three beta distributions with different levels of skewness. The purple distribution has a skew value of 0, the red distribution is positively skewed, and the blue distribution is negatively skewed.

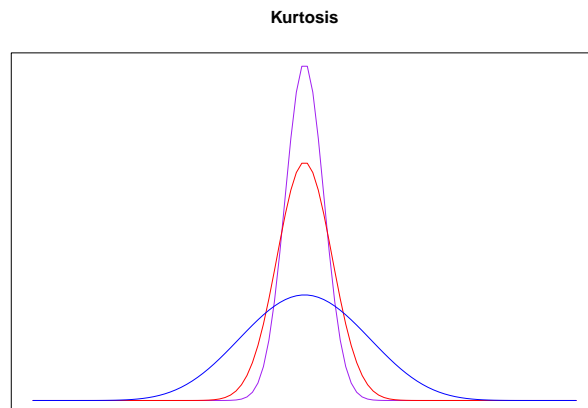


Figure 8.9: Three beta distributions with different levels of kurtosis. The purple distribution has the highest kurtosis value because it has the strongest peak. The blue distribution has the lowest kurtosis value because it has the weakest peak (most flat). The red distribution has a value in the middle.



## Lesson 9

# Introduction to Speech Perception

Recall that the field of linguistic phonetics can be split up into three subcategories: articulatory phonetics, acoustic phonetics, and auditory phonetics. Articulatory phonetics is the study of how the different speech organs move while creating speech. This was the primary focus of LIN 201 at a broad descriptive level. In the last week of this course, we will discuss various articulatory measurements that can be made. So far in this course, we have focused on acoustic phonetics which studies the properties of sound as it exists between a speaker and a hearer. This has taken place both at the descriptive level (theoretical foundations such as sine waves and tube models), as well as at the measurement level where we have measured various acoustic properties. This week we will do a mini unit on auditory phonetics with a primary focus on speech perception. Auditory phonetics also includes describing how the ear manipulates the acoustic signal on its way to the brain, but we will not have time to get into that.

Today's lesson will cover three broad categories that researchers in speech perception are interested in: categorical perception, perceptual illusions, and speaker normalization.

## 9.1 Categorical Perception

## 9.2 Perceptual Illusions

## 9.3 Normalization

In a previous lesson, we discussed how we can predict vocal tract length based on formant values. This highlights one of the issues for understanding how we recognize speech and how the lack of invariance of the acoustic/auditory signal may not mean there is a lack of invariance in whatever percept we use to make a decision about what we heard. For example, if schwa is a neutral vowel where the articulators are completely at rest, then the formant values will vary from person to person based solely on the length of their vocal tract. Therefore, the formant space can vary drastically based solely on physiological makeup.

One primary difference between children and adults is that adult's bodies are all around larger. This includes the vocal tract. Therefore, we expect the formant space for children to naturally be higher since a shorter vocal tract results in a shorter tube model which results in shorter standing waves. As a reminder, a shorter wavelength corresponds to a higher frequency. Figure 9.1 and 9.2 come from Chung et al. (2012). They measured the F1 and F2 vowels of speakers from five different languages in both children and adults and compared the results.

In figure 9.1 we see that the children formant space is shifted down (higher F1) and to the left (higher F2) than the corresponding adult speakers. Figure 9.2 shows the same vowel spaces, but after doing a **transformation**. A transformation is a way that we can normalize the scale across speakers in order to account for speaker-specific variation. Notice that after the transformation, children and adults of the same language have much more similar vowel spaces. This suggests that when we perceive vowels, we are doing some type of transformation. I'm sure all of you have listened to a child speak and not had much trouble understanding the sounds they're making (even if the semantic/lexical content makes no sense).

If we transform the speech signal in some way, it's likely that this happens for all speech sounds and not just vowels, but here we will focus on vowels. Part of the reason for this is because vowels have been known to



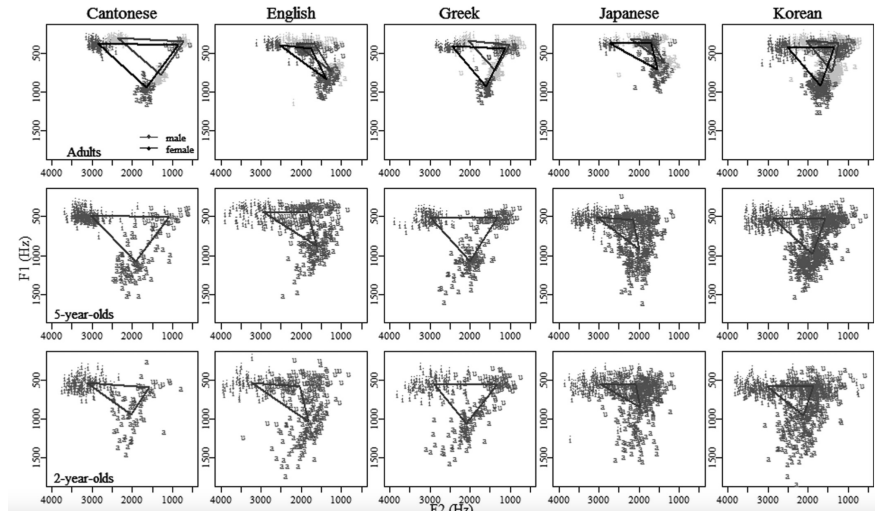


Figure 9.1: Vowel data from (Chung et al., 2012) in its raw form, divided by age and language of speaker.

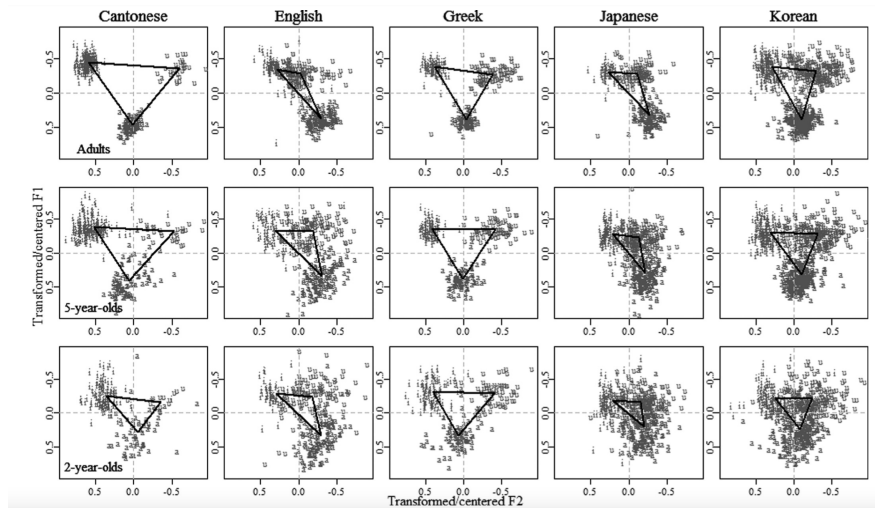


Figure 9.2: Vowel data from (Chung et al., 2012) in its normalized form, divided by age and language of speaker.

encode lots of different types of information within their acoustic/auditory properties. At the very least, they encode three types of information: phonemic information (what category is the speaker producing), anatomical/physiological information (what's the size of the tube, etc.), and sociolinguistic information (age, race, gender, ...). The ultimate goal of speech perception is to obtain the phonemic information, but we can't do that without using the second and third pieces of information. But we can use that information advantageously to recover the phonemic information.

There are various ways to think about vowel normalization. One goal of speech perception research may be to uncover the way in which we normalize the incoming acoustic signal. When researchers model normalization, there are four goals to keep in mind: minimize inter-speaker variation due to physiological or anatomical differences, preserve inter-speaker variation in vowel quality due to dialectal or social factors and preserve features in a sound change, maintain phonological differences, and model cognitive processes that allow listeners to understand different speakers. Typically, vowel normalization procedures are based on three traits: the speaker, the vowel, and the formant. Each of these traits can be based on either intrinsic or extrinsic properties. **Intrinsic** properties are based on only a single speaker, single vowel, or single formant while **extrinsic** properties are based on all the speaker, all the vowels, and all the formants of a given data set.

If you would like to learn more about all the ways in which phoneticians and other speech perception researchers have tried to normalize vowel data, I strongly suggest [this site](#) by Marissa Barlaz. It explains all of the different types of normalization using the R statistical programming language, but you don't need to know anything about programming to read/look at the different normalization procedures. Below, I will focus on one type of normalization as a working example: Lobanov/Z-score normalization.<sup>1</sup>

Z-scoring comes from statistics and it is a way to center a data set around the mean. Suppose we had a list  $X$  of 5 observations.

$$X = [12, 17, 14, 16, 16]$$

Now, suppose we found the mean  $\mu = 15$  and the standard deviation  $\sigma = 1.79$ . The Z-scored value for each value in  $X$  can be calculated with the

---

<sup>1</sup>Lobanov (1971)

following formula:

$$z = \frac{x - \mu}{\sigma}$$

This gives us a new version of  $X$ .

$$X_z = [-1.68, 1.12, -0.56, 0.56, 0.56]$$

Our new version of the list is now the same values but scaled to the mean and standard deviation. Numbers that were closer to the mean are now closer to 0 and numbers that were less than the mean are negative while numbers that were higher than the mean are positive. Suppose I take a new list now called  $Y$  where the values are

$$Y = [25.81, 32.80, 28.60, 31.40, 31.40]$$

Well, if we were to do the same Z-score procedure, we would get the following list:

$$Y_z = [-1.68, 1.12, -0.56, 0.56, 0.56]$$

It's exactly the same transformed list as the transformed  $X$  list even though it started out as all different values. This is the same idea with vowel normalization. Think about list  $Y$  as the child formant values and list  $X$  as the adult formant values. Even though when we look at the raw values the child values are higher than the adult values, once we do some normalization they look exactly the same!

Z-score normalization for vowels works in the following way. For each speaker, find the mean and standard deviation for both F1 and F2 for all the vowels you have data for. This is the  $\mu$  and  $\sigma$  from above. Then for each individual token, do the z-transformation for both F1 and F2 with the correct values. You now have a normalized vowel space. We end this section by showing this process in action with a famous data set of American English vowels (Hillenbrand et al., 1995). Figure 10.1 shows the raw data and figure 9.4 shows the normalized data.

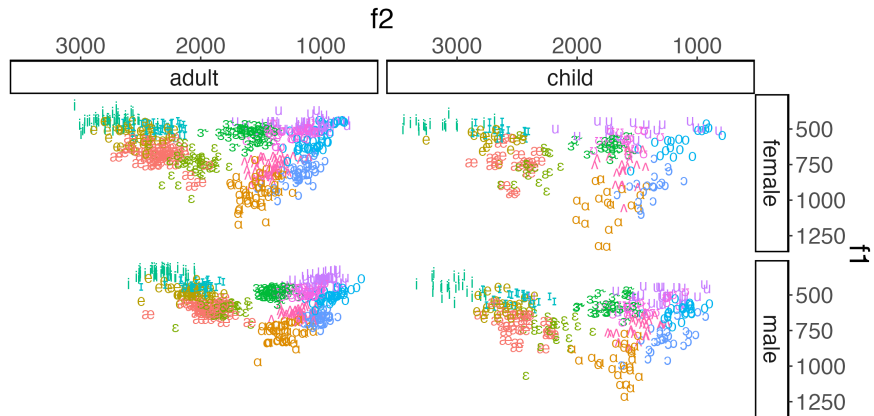


Figure 9.3: Vowel data from (Hillenbrand et al., 1995) in its raw form, divided by age and sex of speaker.

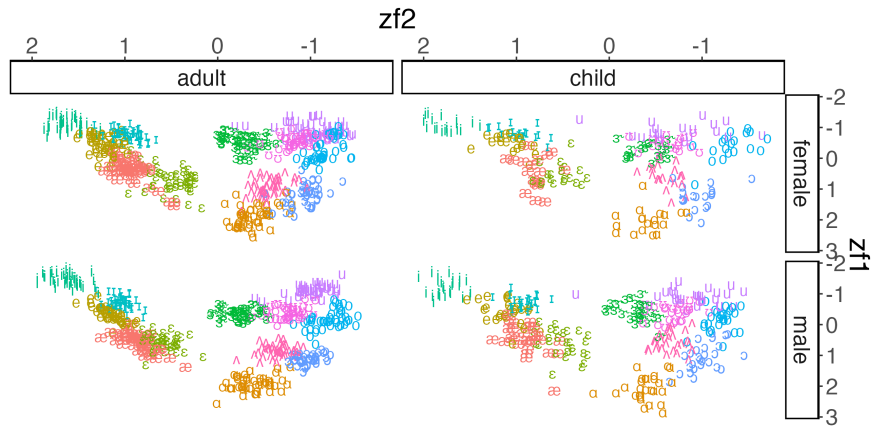


Figure 9.4: Vowel data from (Hillenbrand et al., 1995) in its normal form, divided by age and sex of speaker.

# Lesson 10

## Theories of Speech Perception

This last lesson will briefly discuss two major theories of speech perception: the auditorist account and the gesturist account. The major dividing line in these two theories has to do with what type of information we actually perceive. For auditorists, all perception stems from the auditory signal (the acoustic signal after it has been transformed by the ear). For gesturalists, perception involves “hearing” the distal gesture, possibly by a specialized system designed specifically for perceiving speech. The discussion will focus on the English articulation of /ɪ/ as well as perceptual confusability.

### 10.1 Objects of Perception

This class is not really about *theories* per se, but experimentation doesn't happen in a vacuum and is driven by theoretical work. Therefore, it's worthwhile to think a little bit about what we're actually perceiving when we perceive speech. There is a classic formulation of the “speech chain”<sup>1</sup> that goes:

Speaker's Brain → Gestural Representation → Acoustic Signal → ??? → Listener's Brain

The original formulation has nothing where the box is, but this box is exactly what we're interested in when we think about speech perception.

On one account, we might put *auditory representations* in the box. An auditory representation is very similar to an acoustic representation, except it has been transformed (more on this later) by the inner workings of

---

<sup>1</sup>Denes and Pinson (1933)

the ear. The other account would be that just like going from the brain to the acoustics for the speaker, going from acoustics to the brain also has an intermediate gestural representation step. One of the reasons that researchers have proposed that we perceive gestures and not an auditory representation is the lack of invariance problem discussed in the previous class. While it is true that there is not invariance when it comes to the acoustic signal, there is invariance in the gestures. The acoustics surrounding a velar stop are notoriously variable, but if we think about the articulation invariance emerges. In each case we are taking the tongue dorsum and pressing it up against the velar. Just because the rest of the articulators are affecting it, doesn't mean that the target gesture has changed.

Ok, so should we just accept that we perceive gestures? The reason researchers believe we perceive only the auditory signal is because carefully crafted experiments have shown that listeners respond to decidedly non-speech sounds the same way that they respond to speech sounds. Non-speech sounds have no gesture involved, therefore suggesting that it can't be that we're perceiving gestures. So how do we rectify these two facts? I'm not sure, and I'm not sure the field is ready to accept one over another. In the next couple sections, we'll look at a couple findings and see how they might relate to this debate.

## 10.2 Articulation of English /ɹ/

It is well known that there is variability in the way that English speakers articulate an /ɹ/ sound. And unlike variation from coarticulation such as an /u/ vowel before stops with different places of articulation, the different /ɹ/ articulations appear to be completely different gestural targets. One type of /ɹ/ is the retroflex version where the tongue tip actually curls back and the underside of the tongue moves towards the passive articulator. The second type of /ɹ/ is called the bunched version where your tongue just kind of bunches together. What's interesting is that both of these articulations result in basically the same acoustic form, which is most strongly associated with a lowering of F3. It is not possible for listeners to tell whether or not the an individual token of /ɹ/ was produced with a specific type of articulation, therefore suggesting that perception is primarily based on the acoustic signal.

	p	t	k	f	θ	s	ʃ	b	d	g	v	ʋ	z	ʒ	m	n
p	150	38	88	7	13											
t	30	193	28	1												
k	86	45	138	4	1		1									1
f	4	3	5	199	46	4		1				1				1
θ	11	6	4	85	114	10					2					
s		2	1	5	38	170	10			2						
ʃ		3	3			3	267									
b				7	4			235	4		34	27	1			
d								189	74	48		4	8	11		
g								74	161			4	8	25		
v				3	1			19		2	177	29	4	1		
ʋ								7		10	64	105	18			
z								17	23	4	22	132	26			
ʒ								2	3		1	1	9	191		1
m								1							201	6
n															8	240

Figure 10.1: 0 dB S/N confusion matrix from (Miller and Nicely, 1955).

### 10.3 Perceptual Confusion

Imagine I played you a bunch of words. Chances are you might mishear some of them, but can we predict which sounds you might mishear? The answer is probably yes. In the 1950's, two scientists played a bunch of CV syllables (alongside noise to prevent perfect response) to participants and asked them to report which consonant they heard.<sup>2</sup> This resulted in what is called a confusion matrix. Each row represents the sound that was played, and each column represents the number of times that sound was reported. We can see then that the sound /p/ was reported to be a /p/ 150 out of the 296 times it was played. What's interesting here is that we can see which sounds were most likely to be reported instead of /p/. Notice it was primarily /t/ and /k/ suggesting a perceptual similarity among voiceless stops.

Given a table like this, we can mathematically determine a similarity score. This is done in the following way. First, we calculate a percentage correct score which is equal to the number of correctly guessed tokens for a given sound, divided by the number of incorrectly guess tokens. So for /p/ this value would be  $150/296 = 0.51$  For /k/ it would be  $138/276 = 0.5$  Likewise, we can calculate the percentage of times /k/ was guessed for /p/ and vice versa. This gives us a four by four table which is shown as table 10.1.

The next step is to calculate the similarity value. This is done by adding

<sup>2</sup>Miller and Nicely (1955)

	“p”	“k”
[p]	$150/296 = 0.51$	$88/296 = 0.30$
[k]	$86/276 = 0.31$	$138/276 = 0.50$

Table 10.1: Proportion of “p” and “k” responses when hearing [p] and [k]

the percentages where an incorrect choice was made (top right and bottom left cells) and dividing this by the sum of the correct choices (top left and bottom right cells). So for our running example of /p/ and /k/ this would be:

$$S_{pk} = \frac{0.3 + 0.31}{0.51 + 0.5} = 0.60$$

The actual final step is to take the negative natural log. We won’t do that here, but this allows for similarity to scale exponentially (which may be wanted for independent reasons).



# **Part III**

## **Labs**



# Lesson 11

## Lab 0

This lab is unlike the labs that will follow because it is strictly designed to have you practice using Praat before having to do the main labs that will count towards your grade. That does not mean this will not count towards your grade. If you fail to turn this lab in, it will affect your participation grade. You must type up your responses, but this does not to be in the “lab report” style outlined in the syllabus.

There is an assignment called **Lab 0** on Brightspace where you will submit your work. A successful submission will include: a typed up (preferably pdf) of all questions below, 2 text grid files (*Lab0-a* and *Lab0-b*), and 1 wav file (*Lab0-c*).

### Task 1

1. Download the Lab 0 files from the course website and import them into Praat.
2. *Lab0-a*. {wav, TextGrid}
  - \* Intervals for each individual word have already been created for you. Don't forget to save your TextGrid after annotating it.
  - (a) Using IPA symbols, provide a transcription on the “word” tier of the TextGrid.
  - (b) What is the duration of each word?

- (c) Why might determining the exact start point of word 4 be problematic?
- (d) Is the final stop in word 2 released or unreleased? Explain.

3. *Lab0-b.wav*

- (a) Create a TextGrid with two tiers: word and segment.
- (b) On the word tier, mark the boundaries for each word and provide an IPA transcription.
- (c) On the segment tier, try to mark the boundaries for each segment and provide an IPA transcription. Use your previous knowledge of phonetics to do the best you can.
- (d) What are the F1 and F2 values at the midpoint of each vowel?
- (e) Save the TextGrid as *Lab0-b.TextGrid*

## Task 2

1. Record yourself saying 4 words that match the following patterns:
  - (a) A word that starts with a voiceless fricative and contains a low vowel.
  - (b) A word that ends with a bilabial nasal and starts with an affricate.
  - (c) A word that has a CCCV syllable structure.
  - (d) A word contains a major diphthong and a velarized lateral approximant.
2. Please list your words in your typed up pdf.
3. Save the file as *Lab0-c.wav*

# Lesson 12

## Lab 1

In this lab you will be measuring  $f_0$  as it is used to signify intonation in American English. For this lab, you must write it up in the “lab report” style outlined in the syllabus. There are two parts to this lab, but you can write them up as a single report.

There is an assignment called **Lab 1** on Brightspace where you will submit your work. A successful submission will include a typed up lab report (preferably pdf), and a zip file containing all the recordings you made. If you would like help with the latter portion, please contact me.

### Task 1

1. Come up with 10 unique sentences that can be read as both a declarative statement and a yes/no question. For example, “We were invited to a party.” and “We were invited to a party?”.
2. Record both versions of the sentence one time each (20 total read sentences)
3. Measure the highest pitch value in each sentence. It should appear in the correct spot given the MHL and MH pitch contours we discussed in class.
4. In your lab report, include the mean and standard deviation for the H’s in the declarative sentences and the H’s in the yes/no questions.

Discuss anything you find interesting. Figures/plots are helpful as well.

## Task 2

1. Come up with a sentence of the form: Pronoun - AuxVerb - Verb - Determiner - Adjective - Noun. An example is: You might buy the old couch.
2. Record yourself saying your sentence 20 times total which will be 5 times each with emphasis on the Pronoun, Verb, Adjective, and Noun.
  - (a) YOU might buy the old couch
  - (b) You might BUY the old couch
  - (c) You might buy the OLD couch
  - (d) You might buy the old COUCH
3. Measure the highest  $f_0$  value in each recording of the emphasized word.
4. In your lab report, include the mean and standard deviation for these values, summarized by emphasized word. Also include the mean and standard deviation for all of these values lumped together. Compare these values to what you measured in Task 1. Discuss anything you find interesting. Figures/plots are helpful as well

# Lesson 13

## Lab 2

In this lab you will be measuring  $f_0$  as it used to signify lexical contrast in San Juan Quiahije Eastern Chatino. This is a language spoken by a small number of speakers in a mountainous area of Oaxaca, Mexico. The data come from [Cruz and Woodbury \(2014\)](#). For this lab, you must write it up in the “lab report” style outlined in the syllabus.

There is an assignment called **Lab 2** on Brightspace where you will submit your work. A successful submission will include a typed up lab report (preferably pdf), and a spreadsheet with all your measurements.

### Task 1

1. Download the Lab2 folder from the course website. This contains 26 unique words from SJQ Chatino.
2. For each word, measure the  $f_0$  value during the vowel portion of the word at five points:
  - (a) The starting point
  - (b) The 25% point
  - (c) The mid point
  - (d) The 75% point
  - (e) The ending point

3. Determine how many lexical tones there are in SJQ Chatino by comparing the  $f_0$  values for each word (i.e., the pitch contour).
4. In your lab report, discuss your procedure for determining how many lexical tones you think there are. Be sure to use the quantitative measures you make to support your conclusions.
5. Include figures drawn with Praat that show all the different lexical tones.



# Lesson 14

## Lab 3

In this lab you will be measuring F1 and F2 as they are used to signify vowel quality in your own speech. Part 1 will look at English monophthongs, part 2 will look at English diphthongs, and part 3 will look at schwa in order to calculate the length of your vocal tract. For this lab, you must write it up in the “lab report” style outlined in the syllabus.

There is an assignment called **Lab 3** on Brightspace where you will submit your work. A successful submission will include a typed up lab report (preferably pdf), a spreadsheet with all your measurements, and a zip file containing your audio recordings.

### Task 1

1. Record yourself saying the following words, 10 times each.
  - (a) beat, bit, bet, bat, but, boot, put, bought, bot, bird
2. Do not say the words in isolation, but rather use a carry phrase.
  - (a) The carrier phrase you should use is “I will say ... this time.”
3. Randomize the order so you don’t say the same target word twice in a row.
  - (a) E.g. “I will say beet this time.”, “I will say bird this time”, “I will say put this time,”, etc...

4. Measure the F1 and F2 at the midpoint of the vowel in the target word for each recording.
5. Report the mean and standard deviation for each vowel (there is a unique vowel in each word for me, but some of you may merge two of them)
6. Plot the mean value for each vowel, making sure that the axes are reversed so it looks like the standard IPA vowel chart.
7. In your lab report, discuss anything interesting you notice about your vowel space.

## Task 2

1. Record yourself saying the following words, 10 times each.
  - (a) bait, boat, bite, bide, bowed, boid (it's a surname)
2. Do not say the words in isolation, but rather use a carry phrase.
  - (a) The carrier phrase you should use is "I will say ... this time."
3. Randomize the order so you don't say the same target word twice in a row.
  - (a) E.g. "I will say bait this time.", "I will say bowed this time", "I will say bide this time,", etc...
4. Measure the F1 and F2 vowel at five points:
  - (a) The starting point
  - (b) The 25% point
  - (c) The mid point
  - (d) The 75% point
  - (e) The ending point
5. In your lab report, analyze and plot the F1 and F2 trajectory for each of these vowels as they would appear on a vowel chart (axes reversed).

6. Discuss anything interesting you notice

### Task 3

1. Come up with a list of 10 disyllabic words that have stress on the second syllable and the first vowel is schwa.
  - (a) E.g. - *about* (you can use this as one)
2. Record yourself saying each word 1 time.
3. Use the carrier phrase from above.
4. Measure the F1, F2, and F3 at the midpoint of the schwa vowel in each word.
5. Find the mean for each formant.
6. Use the formula from the course notes to calculate your vocal tract length using all three formants
7. Discuss anything interesting you notice



# Lesson 15

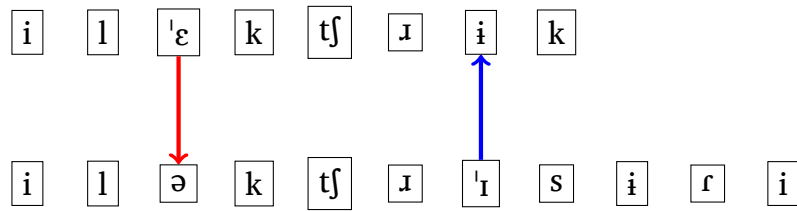
## Lab 4

In this lab you will be measuring intensity, duration, f0, F1, and F2 to see how they are used to signify stress in American English. For this lab, you must write it up in the “lab report” style outlined in the syllabus.

There is an assignment called **Lab 4** on Brightspace where you will submit your work. A successful submission will include a typed up lab report (preferably pdf), a spreadsheet with all your measurements, and a zip file containing your audio recordings.

### Task 1

1. In English, the affixes -ic, -ity, and -ion shift the location of the stress on a root word. For example, electric has stress on the second syllable, but electricity has stress on the third syllable.
2. Come up with 5 words for each affix that can take that affix.
3. Record each word with and without the affix 3 times. Use a carrier phrase.
4. There are two pairs of vowels you are interested for each word, the stressed vowel without the affix and its unstressed counterpart with the affix (red) and the unstressed vowel without the affix and its stressed counterpart with the affix (blue).



5. For each of these pairs, measure the total duration of each vowel, the average intensity over the duration of the vowel, the max intensity over the duration of the vowel, the average f0 over the duration of the vowel, the max f0 over the duration of the vowel, the F1 at the midpoint of the vowel, and the F2 at the midpoint of the vowel.
6. Since you have three tokens of each words, report the mean and standard deviation for each vowel in each pair.
7. Also report the overall mean and standard deviation of stressed vs. unstressed vowels.
8. Plotting instruction...tbd
9. Include anything interesting/noteworthy you find in your write up. Do any of these properties seem to be more important than others for signifying stress in American English?

# Lesson 16

## Lab 5

In this lab you will be measuring VOT and percentage of voicing during the segment to get an idea of how voicing is acoustically realized in American English. For this lab, you must write it up in the “lab report” style outlined in the syllabus.

There is an assignment called **Lab 5** on Brightspace where you will submit your work. A successful submission will include a typed up lab report (preferably pdf), a spreadsheet with all your measurements, and a zip file containing your audio recordings.

### Task 1

1. You will be measuring the voice onset time of stops in American English.
2. For each stop (p,t,k,b,d,g), find 2 monosyllabic words that begin with the stop and contain the vowels (i, u, æ, ɑ).
  - (a) So two words with a [pi] sequence, two with a [pu] sequence, ..., two with a [gɑ] sequence.
  - (b) This should result in 48 total words
3. Record each word in isolation (no carrier phrase)
4. Make sure you randomize the words and pause in between each one
5. Report the mean and standard deviation of VOT for each stop

6. Include anything interesting/noteworthy you find in your write up. Be sure to compare the difference in voiced/voiceless as well as the difference in place of articulation.

## Task 2

1. You will be measuring the percentage of voicing during the segment for English fricatives at the end of words.
2. For each fricatives (f,v,θ,ð,s,z,ʃ,ʒ) find 2 monosyllabic words that end with the fricative and contain the vowels (i, u, æ, ɑ). If you can't find 2 real words, it is okay to use a nonce word. A nonce word is a made up word that sounds like it could be in the language. You may have to use exclusively nonce words for one of these sounds.
3. Record each word in isolation (no carrier phrase)
4. Make sure you randomize the words and pause in between each one
5. Report the mean and standard deviation of preceding vowel duration for each fricative
6. Include anything interesting/noteworthy you find in your write up. Be sure to compare the difference in voiced/voiceless as well as the difference in place of articulation.
7. Discuss any thoughts about voicing in stops vs. fricatives as well as any difference you can think of between beginning and end of word voicing.



# Lesson 17

## Lab 6

In this lab you will be measuring F2 locus equations for stops, and various spectral measurements for fricatives. You will reuse the recordings you made in Lab 5. For this lab, you must write it up in the “lab report” style outlined in the syllabus.

There is an assignment called **Lab 6** on Brightspace where you will submit your work. A successful submission will include a typed up lab report (preferably pdf), a spreadsheet with all your measurements.

### Task 1

1. Use the recordings of only the voiced stops you made for Task 1 in Lab 5.
2. Measure two values for each token:
  - (a) F2 at the onset of the vowel
  - (b) F2 at the midpoint of the vowel
3. For each stop, calculate a slope
  - (a) The formula for calculating slope is  $\text{slope} = \frac{N*\Sigma(xy) - \Sigma(x)\Sigma(y)}{N*\Sigma(x^2) - \Sigma(x)^2}$ .
  - (b) If that reads as a bunch of gibberish to you , you can use [this](#) google sheet to automatically calculate the slope for you. Just plug in your 8 values for each specific stop, and write down what slope it gives you in cell D1.

4. Report the slope for each stop and discuss what you notice. A steeper slope (corresponding to a higher absolute values) corresponds to more coarticulation. Which place of articulation has the most coarticulation?
5. Additionally, report the mean value for  $F2_{\text{onset}}$ . Do the relative values for each place of articulation match what you expected given the lecture. Discuss

## Task 2

1. Use the recordings of all the fricatives you made for Task 2 in Lab 5.
2. Create a spectrum object at the midpoint of each fricative token.
3. Using the spectrum object, measure four values:
  - (a) Center of Gravity
  - (b) Standard Deviation
  - (c) Skewness
  - (d) Kurtosis
4. Report the mean and standard deviation of each value grouped by segment.
5. Discuss and compare these values by place of articulation, voicing, and [f,v,θ,ð] vs. [s,z,ʃ,ʒ].
6. Include anything else you find interesting.

# Lesson 18

## Lab 7

In this lab you will run a speech perception experiment on yourself and one other person. The specific nature of the experiment will be revealed after every student submits their data. For this lab, you must write it up in the “lab report” style outlined in the syllabus.

There is an assignment called **Lab 7** on Brightspace where you will submit your work. A successful submission will include a typed up lab report (preferably pdf).

### Task 1

1. Download the Lab 7 experiment file from the course website.
2. Open Praat and click Open > Read from file... and open experiment7.txt.
3. You will see three experiment files in the object window. Make sure they are all highlighted and then click Run.
4. This will start the experiment. There are three parts. Some notes:
  - (a) Parts 1 and 3 are the same, don't worry.
  - (b) When you click on an answer, there's no sensory feedback, but your click was captured, don't worry.
5. When you are finished, you can click out of the experiment window.
6. DO NOT EXIT PRAAT BEFORE SAVING YOUR DATA

7. You must repeat this for all three Parts:
  - (a) Click on one of the parts in the object window
  - (b) Click on `Extract Results`
  - (c) A new object will appear, select it and click `Collect to Table`.
  - (d) A final new `Table` object will appear. Click on it, then click `Save > Save as comma-separated file`. Then change the file name to be the initials of the person who was just tested (you or your other person), the part of the experiment, and then `.csv`. So if I ran the experiment on myself my file would be called `SN1.csv`. (or `SN2.csv`, `SN3.csv`).
8. Send me all the data files by Friday, August 11th at 9:30am.
9. Once I get all the data, I will aggregate it and share the results and more info that you can use in your write up.

# Lesson 19

## Lab 8

In this lab you will run a speech perception experiment on yourself and one other person. The specific nature of the experiment will be revealed after every student submits their data. For this lab, you must write it up in the “lab report” style outlined in the syllabus.

There is an assignment called **Lab 8** on Brightspace where you will submit your work. A successful submission will include a typed up lab report (preferably pdf).

### Task 1

1. Download the Lab 8 experiment file from the course website.
2. Open Praat and click Open > Read from file... and open experiment8.txt.
3. You will see one experiment files in the object window. Make sure it is highlighted and then click Run.
4. This will start the experiment. There is only one part. Some notes:
  - (a) Parts 1 and 3 are the same, don't worry.
  - (b) When you click on an answer, there's no sensory feedback, but your click was captured, don't worry.
5. When you are finished, you can click out of the experiment window.
6. DO NOT EXIT PRAAT BEFORE SAVING YOUR DATA

7. Click on the experiment in the object window
8. Click on Extract Results
9. A new object will appear, select it and click Collect to Table.
10. A final new Table object will appear. Click on it, then click Save > Save as comma-separated file. Then change the file name to be the initials of the person who was just tested (you or your other person and then .csv. So if I ran the experiment on myself my file would be called SN1.csv.
11. Send me the data files by Sunday, August 13th at 9:30am.
12. Once I get all the data, I will aggregate it and share the results and more info that you can use in your write up.

# Bibliography

- Abramson, A. S. and Whalen, D. H. (2017). Voice onset time (vot) at 50: Theoretical and practical issues in measuring voicing distinctions. *Journal of phonetics*, 63:75–86.
- Cho, T. and Ladefoged, P. (1999). Variation and universals in vot: evidence from 18 languages. *Journal of phonetics*, 27(2):207–229.
- Chung, H., Kong, E. J., Edwards, J., Weismer, G., Fourakis, M., and Hwang, Y. (2012). Cross-linguistic studies of children’s and adults’ vowel spaces. *The Journal of the Acoustical Society of America*, 131(1):442–454.
- Cruz, E. and Woodbury, A. C. (2014). Finding a way into a family of tone languages: The story and methods of the Chatino language documentation project. *Language Documentation & Conservation*, 8:490–524.
- Delattre, P. C., Liberman, A. M., and Cooper, F. S. (1955). Acoustic loci and transitional cues for consonants. *The journal of the acoustical society of America*, 27(4):769–773.
- Denes, P. B. and Pinson, E. (1933). *The speech chain*. Macmillan.
- Gick, B., Wilson, I., and Derrick, D. (2013). *Articulatory phonetics*. John Wiley & Sons.
- Hillenbrand, J., Getty, L. A., Clark, M. J., and Wheeler, K. (1995). Acoustic characteristics of american english vowels. *The Journal of the Acoustical society of America*, 97(5):3099–3111.
- Johnson, K. (2011). *Acoustic and Auditory Phonetics*. John Wiley & Sons.
- Ladefoged, P. and Disner, S. F. (2012). *Vowels and consonants*. John Wiley & Sons.

- Ladefoged, P. and Johnson, K. (2014). *A course in phonetics*. Cengage learning.
- Lisker, L. and Abramson, A. S. (1964). A cross-language study of voicing in initial stops: Acoustical measurements. *Word*, 20(3):384–422.
- Lobanov, B. M. (1971). Classification of russian vowels spoken by different speakers. *The Journal of the Acoustical Society of America*, 49(2B):606–608.
- Miller, G. A. and Nicely, P. E. (1955). An analysis of perceptual confusions among some english consonants. *The Journal of the Acoustical Society of America*, 27(2):338–352.



# Appendices



# Appendix A

## Introduction to Praat



### A.1 What is Praat?

Praat (pronounced [pɾɑt]) is a free, open-source software created by Paul Boersma and David Weenink at the University of Amsterdam. It has many functions, but is primarily used to analyze speech. In this course, we will be using it to learn about different speech sounds and how we can analyze them manually by hand or automatically with the help of a computer. You can download Praat by going to <http://www.praat.org> and following the links for your specific platform.

---

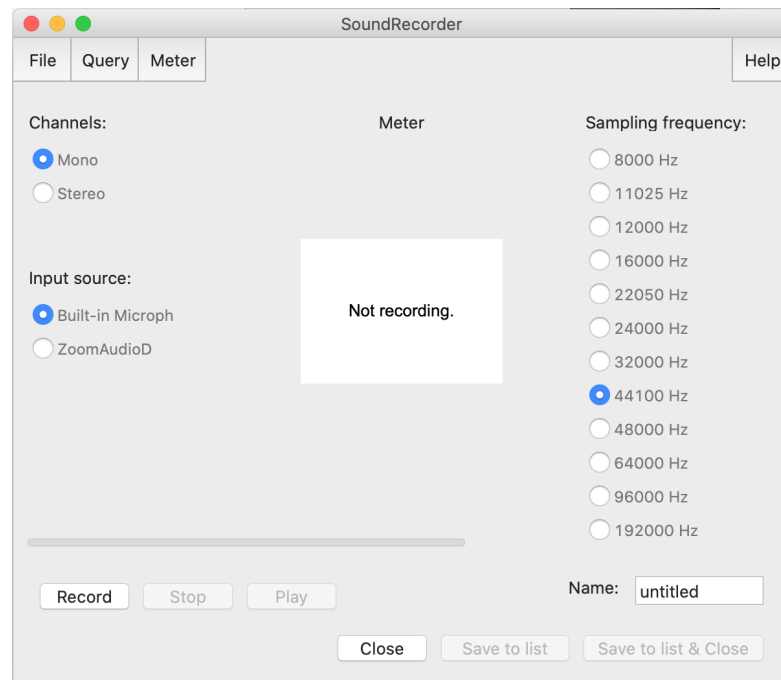
**\*\*\* Activity 1 \*\*\***

Download and install Praat. If you run into any problems, send me an e-mail at [scott.nelson@stonybrook.edu](mailto:scott.nelson@stonybrook.edu) or message me on Slack and I will help you troubleshoot.

---

## A.2 Opening Praat/Sound Files

When you first open Praat you should have two windows pop up. One should be called PRAAT OBJECTS and the other should be PRAAT PICTURES. You can close out of PRAAT PICTURES as we won't be using that for anything that we do. In order to do anything with the PRAAT OBJECTS window, we need an object to work with. So let's create one now. At the top of the Objects window you should be able to click New → Record mono Sound. . . . This should cause a new window called SOUNDRECORDER to pop up.



You may have to change your Input source, but otherwise you should be ready to record at this point. To record, simply click *Record* and speak into whatever input source you chose. If your computer does not have a built in microphone and you are unable to record, that is ok. I will show you how to import sound files momentarily. Once you are done speaking, click *Stop*. You can listen back to the recording you just made by clicking *Play*. If you are satisfied with the recording you made, name your file in the text box next to *Name:* and then click *Save to list & Close*. If you are unsatisfied with your recording, you can click *Record* again. This will essentially record over your original recording and start the cycle anew.

In the *Objects* window you should now see an object named “1. Sound <name of sound>”. When this sound is selected, it should bring up a bunch of buttons you can click on the right side of the *Objects* window. We’ll ignore those for now. Instead, let’s save the recording we just made. At the top of the *objects* window click *Save → Save as WAV file...* and then save the file onto your computer. Now that we’ve saved the file, let’s see how we can import sound files. The first step will be to select the sound in the *Objects* window. After you’ve done this, there is a button at the bottom of the window labeled *Remove*. Once you clicked the *remove* button, the sound file should disappear. Let’s add it back to the *Objects* window now by going to the top and clicking *Open → Read from file...* Find the file that you just saved to your computer and select it then click *Open*. You should see it reappear in the *Objects* window as “2. Sound <name of sound>”. Don’t worry about the numbering. Praat adds 1 to every new object number added to the window regardless if it has been there before. This resets every time you open the program.

If you were unable to record a sound file directly into your computer, but had other ways of recording sound (e.g. - with your phone), you can upload those into Praat this same way. I’d suggest saving it as a .WAV file when transferring from source to source as Praat can be a little picky with audio file types.

---

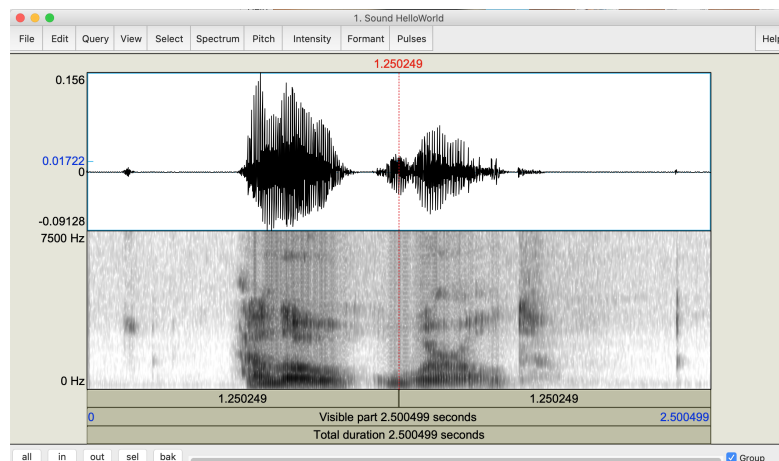
### \*\*\* Activity 2 \*\*\*

Make sure you understand how to record new sounds in Praat. Record yourself saying the phrase “Hello World” and save it to your computer as *HelloWorld.wav*. If you recorded it within Praat, exit out of Praat, reopen it and load the saved file into your *Object* window. If you did not

record within Praat, open Praat now and load the saved file into the Object window.

### A.3 Text Grids

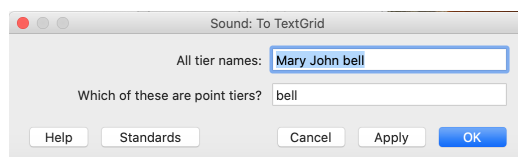
Now that we have a sound to work with, let's see what it looks like. If you highlight your sound file in the Objects window, the buttons you saw earlier should reappear. Let's take a look at what a sound object looks like in Praat. Click on *View & Edit* now. A window called **SOUND <NAME OF SOUND>** should pop up now. This is a spectrogram. It should look something like the figure below.



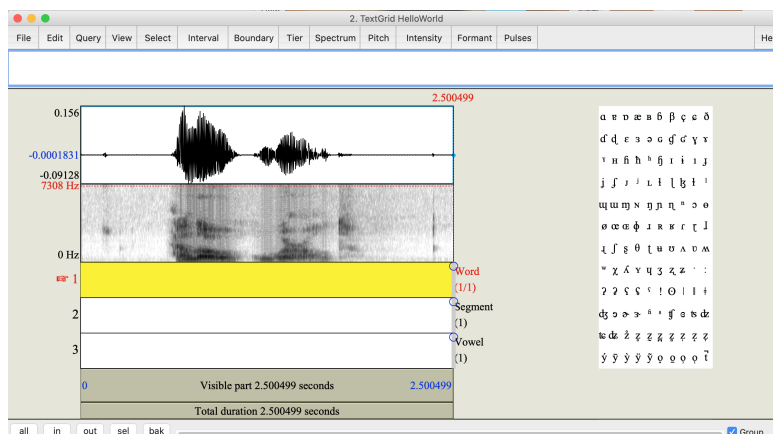
A spectrogram is a visual representation of sound. The x-axis is time, the y-axis is frequency, and the z-axis (represented by how dark a spot is) is amplitude. A higher amplitude means that there is a lot of energy at that specific point. If you click anywhere on the spectrogram, you will be presented with two values: the time at which you clicked (in seconds) and the frequency value at which you clicked. These are useful for making measurements by hand.

Another useful thing to do is divide the speech signal up into different segments. We can do this by creating what is called a TextGrid. Let's make one of those now. First you need to close out of the Sound Viewer

window. This should bring you back to the Objects window. Make sure your sound is highlighted. On the right side of the Objects window click `Annotate → To TextGrid...` You should have a window called `SOUND: TO TEXTGRID` pop up that looks like the one below.



You can delete all the values in both cells. We will leave the cell for `Which of these are point tiers?` blank. In the cell for `All tier names:` we can supply a list of different tiers we want to create. This is done by putting spaces in between each tier name. Let's add two tiers. The first tier will be called "Word", the second tier will be called "Segment", and the third tier will be called "Vowel". Once you have done that click `OK`. You should now have two objects in the Objects window. Your original sound and "TextGrid <name of sound>". Highlight both in the Objects window. When you click `View & Edit` now your `SoundViewer` window should look slightly different than before.



The window now includes the three tiers we just created underneath the spectrogram, and a set of IPA symbols on the right. The IPA symbols can be used to label different segments on the tiers. Let's make a few

segments now. If you click on the first tier (Word), it should turn yellow. This tier is for segmenting the speech sound into different words. If you click on the spectrogram at the start of a word you should see a line appear on all the tiers. If you click Enter on your keyboard this will create a boundary marker on the selected tier. You can then do the same thing for the end of the word. Once you're done segmenting all the words, be sure to label them with English spelling or IPA. Once you have different segments created on a tier, clicking on a specific area will highlight only that interval and you will be able to enter text there.

---

### \*\*\* Activity 3 \*\*\*

Create a TextGrid for “Sound HelloWorld” with one tier labeled “Word”. Create an interval on the word tier for each word of the phrase. Use the IPA letters on the right to label each segment with the correct IPA transcription.

---

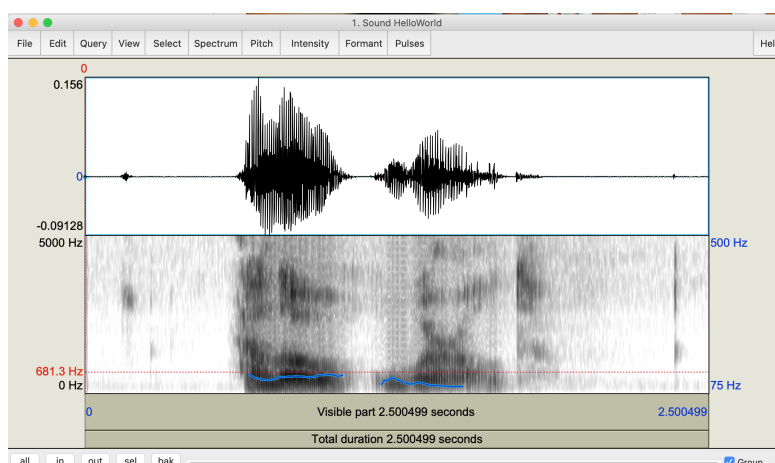
## A.4 Measuring Things

There are different parts of the speech signal that phoneticians and other speech researchers use for analysis. I will mention two of them here today: pitch and formants. Let's start by looking at this in the sound viewer. For this, we do not need the TextGrids, so you can open just the sound in the SoundViewer.

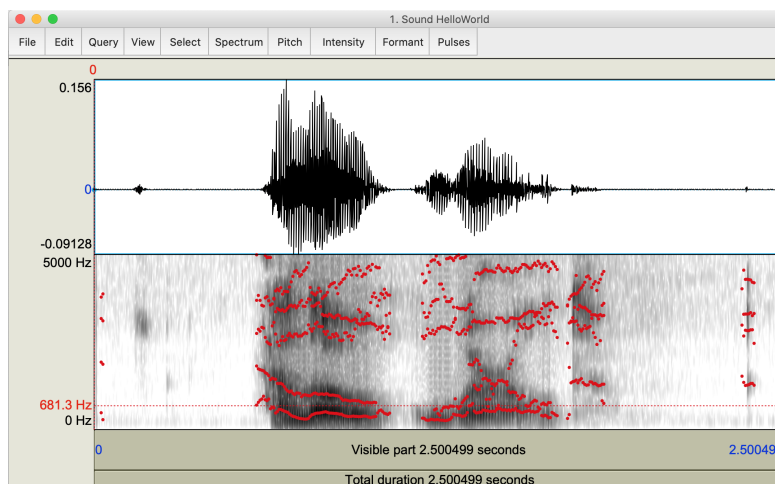
At the top of the window you can click Pitch → Show pitch. This should do two things to the spectrogram. First, it should add a blue line overtop the spectrogram. This is the pitch path. Second, it should add a blue numbers (500) to the top right side of the spectrogram. This is the range for the pitch values. Notice now when you click on the spectrogram it shows a value in red on the left, but also at the bottom right it shows a value in blue (given there is a pitch at the certain time selected). This allows us to manually get pitch values at any time we'd like.

If you click Pitch → Show pitch again, the blue line and numbers will go away. Let's show the formants now. If you click Formant → Show formants you will see a bunch of red dots appear. Unlike pitch, there are no new numbers on the right hand side as formants are really just





measurements of the normal energy distribution on the spectrogram. Even though they are dots, you should notice trend lines appearing. While there is one single pitch value, there are multiple formants within the speech signal. We are primarily interested in the bottom two which are called F1 and F2. If you click on one of the lowest of the red lines at any given point on the spectrogram, the red number on the left is the F1 value at that point. If you click on the second lowest point, then that is the F2 value.



We can also get values for pitch or formants without using the Sound-Viewer. Let's go back to the Objects window to see how to do this. The

first thing we have to do is create a Pitch object and a Formant object. To create the Pitch object, highlight the sound you want to analyze and click *Analyse periodicity - → To Pitch...* This will open up a window name **SOUND: TO PITCH**. You can just click **OK** as we don't need to adjust any of the settings. You should now see a new object in the Objects window called "Pitch <name of sound>".

Let's do the same thing for formants. Highlight the sound then click *Analyse spectrum - → To Formant (burg) ...* This will open a window called **SOUND: TO FORMANT (BURG METHOD)**. We can once again avoid changing any of the parameters and just click **OK**. You should now have four objects in the Objects window with the newest being called "Formant <name of sound>".

Now that we have the Pitch object and the Formant object, we can query values at a specified time. To do this, highlight either of these objects and click *Query → Get value at time...* Once you've done this, you can specify the time (in seconds) that you would like to get a value at. If you are querying formant values, you will also have to choose which formant you would like information about. Once you enter these parameters, click **OK** and a new text window will pop up with the values given the entered parameters.

---

**\*\*\* Activity 4 \*\*\***

Open up "Sound HelloWorld" and analyze its pitch. Try to measure its pitch at five different points. Record the time and pitch value for each one. Now, go back to the Objects window and create a Pitch Object for "Sound HelloWorld". Using the Pitch object, query the pitch values at the times you recorded. Do they match up with the values you got from the SoundViewer?

---

## A.5 Other Resources

In this course we will go more in depth using Praat, but If you'd like to learn even more about it, I suggest looking at [Will Styler's manual](#).

Praat also has a built in scripting language. We may use this a little bit throughout the course. Scripting allows for automatic analysis of speech

and is very useful when conducting research. If you have a coding background and are interested in learning more about this, there is a tutorial within Praat. In the Objects window click Help → Praat Intro and then scroll down under the General section and click on Scripting. There are additional third party tutorials [here](#) and [here](#).



# Appendix B

## Syllabus

LIN 350 – Experimental Phonetics

Summer 2023

---

Instructor: Scott Nelson

Office: SBS N235

✉ [scott.nelson@stonybrook.edu](mailto:scott.nelson@stonybrook.edu)

Class Time: T/Th 9:30am – 12:55pm

Class Location: SBS N310

Office Hours: by appointment

### Office Hours/Contact

I will be holding both in person and virtual office hours via Zoom by appointment. If you would like to talk with me outside of class, please do not hesitate to send me an e-mail so we can work out a time to meet up. You may access my “virtual office” using the meeting link provided here: <https://stonybrook.zoom.us/j/4091340647>. *Note: Zoom does not require you to download their software to join a meeting.*

If you are contacting me via e-mail, please put [LIN 350] in the subject line. I will respond within two hours from 9am-5pm on Monday-Friday. All other times I will respond sporadically, but at least within half a day.

You may also communicate with me via our class Slack channel (details below). In theory I should be able to respond to Slack posts quicker. Please

limit this type of communication to quick or general questions. Any communication about class performance or grades needs to be done via e-mail.

## Materials and Resources

**Brightspace** - The official course Brightspace page will be used to host readings, accept assignments, and post grades. A supplemental course website will be used for everything else.

**Class website** - A website for the class can be found at [this link](#). Additional class material will be posted here including a course log and links to Zoom/Google Forms/etc.... Please check this page regularly as you are responsible for all information posted there for this class.

**Slack** - You can join the class Slack channel [here](#). Slack is a messaging app designed to be used for workplace communication. You can find an introduction on how to use it [here](#). It is also good for class communication as it allows for easier and less formal communication between myself and the students. This is good for quick questions or off topic thoughts you want to share with me or other students. It also allows for sharing files/pictures/etc... so it can be used for all different types of communication. I recommend downloading the app for either your phone or desktop.

**Textbook** - There is no required textbook for this class. Readings will come from a variety of sources and will be posted on the class Brightspace page.

## Course Description

Introduction to common experimental methods for studying the sounds used in human language. Topics include basic speech acoustics, acoustic analysis, oral and nasal airflow, static palatography, linguography and electroglottography, as well as design of perception experiments. Students will learn the physical processes affecting each experimental variable and common methods of analyzing each kind of data. Students will get hands-on experience with each analysis method and will use two or more types

of data to explore a hypothesis about sound structure in English or some other language of interest. Students will learn how to use software for making measurements and analyzing data. Students will learn to assess the validity of claims about language based on their understanding of the scientific method as applied to speech. The course will give students a solid foundation for further courses in laboratory skills relevant to assessment of normal and disordered speech and for pursuing research, either as undergraduate researchers, or in the early stages of graduate work.

## Grading/Structure of Course

Every class meeting will be broken up into three sections: **Lecture**, **Praat**, and **Lab**. The **Praat** and **Lab** portions will be interactive and will require students to bring a laptop that is able to run Praat to class. If you do not have access to a laptop, please consult with the [Laptop Loaner Program](#).

A.	Final Project	40%
B.	Reflection Essay	10%
C.	Lab Reports (8)	40%
D.	General Participation	10%

### A. Final Project

Each student will be required to do an acoustic analysis on a topic of their choosing. The analysis must involve a comparison of some type between two speech communities. This can include speakers of different languages, speakers of the same language but different dialects, or disordered and non-disordered speakers. The main goal of this project is for students to apply the technical skills learned in this class to new data. A secondary goal is to provide practice for students presenting their work to others. There are three parts to the final project: a short proposal, an in-class presentation, and a written report.

You must submit a project topic proposal to me for approval by **Friday, August 4 at 5pm**. The proposal can be a short paragraph outlining the analysis you plan to perform. The goal of the proposal is to make sure

you have a plan in place. I may request changes to your plan if necessary. Each student will also be required to give an approximately 10-15 minute presentation on their topic. After each presentation, there will be a 3-5 minute question period where students will ask the presenter further questions about their project. Presentations will take place on **August 15** and **August 17**. The final report should be 3-5 pages (single spaced, 1in margin, 12pt font) and is due **Friday, August 18 at 5pm**.

## **B. Reflection Essay**

Alongside the final draft of your project report you will turn in a reflection essay. This will be a one page document (single spaced, 1in margin, 12pt font) where you reflect on your final project and the class as a whole. It should minimally include what you consider to be the strengths and weaknesses of your final project and what you may have done differently if you had more time. Other than that, the topic of your reflection essay is open ended. Some possible directions to take it include: what you liked and didn't like about the course, how you will use what you learned in this course moving forward, or general suggestions for the structure of this course in the future.

## **C. Lab Reports**

In weeks 2-5, each class period will end with a lab. These labs are designed to give students hands on experience measuring and analyzing acoustic data. Each lab will be completed individually and involves writing up a short lab report. The lab report is due by the start of the class period one week after the assignment of the lab (i.e. - if the lab is on a Tuesday, it is due by the start of the following Tuesday's class).

Every lab report should contain the following four sections: *Introduction*, *Material and methods*, *Results*, *Discussion and conclusion*. The introduction provides a brief overview of what is being measured and why it is of interest. The materials and methods sections should give an explicit description of how things were measured. This includes both descriptions of materials used (software, microphone) as well as techniques used (e.g. - measured



the vowel at its midpoint). Someone should be able to replicate your process exactly after reading this section. The results section contains the results. This should be a combination of both prose and figures. The discussion and conclusion section summarizes all of the results and discusses interesting and significant findings.

## **D. General Participation**

Students are expected to be active participants in class. If you are unable to make a class, please try to let me know ahead of time.

## **Student Accessibility Support Center Statement**

If you have a physical, psychological, medical, or learning disability that may impact your course work, please contact the Student Accessibility Support Center, Stony Brook Union Suite 107, (631) 632-6748, or at [sasc@stony-brook.edu](mailto:sasc@stony-brook.edu). They will determine with you what accommodations are necessary and appropriate. All information and documentation is confidential.

## **Academic Integrity Statement**

Each student must pursue his or her academic goals honestly and be personally accountable for all submitted work. Representing another person's work as your own is always wrong. Faculty is required to report any suspected instances of academic dishonesty to the Academic Judiciary. Faculty in the Health Sciences Center (School of Health Professions, Nursing, Social Welfare, Dental Medicine) and School of Medicine are required to follow their school-specific procedures. For more comprehensive information on academic integrity, including categories of academic dishonesty please refer to the academic judiciary website at [https://www.stonybrook.edu/commcms/academic\\_integrity/index.html](https://www.stonybrook.edu/commcms/academic_integrity/index.html).

## **Critical Incident Management**

Stony Brook University expects students to respect the rights, privileges, and property of other people. Faculty are required to report to the Office of Student Conduct and Community Standards any disruptive behavior that interrupts their ability to teach, compromises the safety of the learning environment, or inhibits students' ability to learn. Faculty in the HSC Schools and the School of Medicine are required to follow their school-specific procedures. Further information about most academic matters can be found in the Undergraduate Bulletin, the Undergraduate Class Schedule, and the Faculty-Employee Handbook.

# Appendix C

## Example Lab Report

Below is an example lab report for Lab 2.

---

### Lab 2

Scott Nelson

#### **Introduction**

Lexical tone is a way in which languages use  $f_0$  to contrast between different lexical items. One language that has lexical tones is San Juan Quiahije Eastern Chatino which is a language spoken by a small number of speakers in a mountainous area of Oaxaca, Mexico. In this lab, the task is to figure out how many, and which, lexical tones the language uses. The results suggest that there are approximately nine different lexical tones used by speakers of the language.

#### **Materials and Methods**

Recordings of 26 monosyllabic words in San Juan Quiahije Eastern Chatino were provided. Praat (Mac; version 6.3.01) was used to automatically analyze the  $f_0$  contour of each word. A custom Praat script extracted the  $f_0$  value at five points in each vowel: 0%, 25%, 50%, 75%, and 100% of its duration. The Praat pitch tracking algorithm was used with the following settings: Time Step (s) = 0; Pitch Floor (Hz) = 50; Pitch Ceiling (Hz)

Tone Pattern	Words
H	10,14,18,19
HL	15,24
HM	3,8,11
L	1,5,6,13,26
LM	2,22
M	7,9
M0	12,17,23
MH	21,4
ML	16,20,25

Table C.1: Proposed tone grouping for SJQ Chatino

= 600. After a first pass, the data were hand inspected to look for any unusual results. In some instances, the end points of the vowel had to be moved in order to cover time points where the Praat pitch algorithm captured values. Additionally, word 6 was manually adjusted to remove the end breathy portion of the vowel because it was registering with a super low f0 and the 25% point of word26 was manually calculated by hand measuring a single cycle because this point was not being picked up by the pitch tracker. These fixes resulted in reasonable f0 values for each word at all five points.

### Results

Figure C.1 shows the proposed tone groupings based on the f0 measurements at the five time points. There are nine total groups, including three level tones: L, M, H; three rising tones: MH, M0 (0 = super high), LM; and three falling tones: HL, HM, ML. The specific words in each group are shown in table C.1.

The mean and standard deviation for each individual tone target was taken at the onset and offset depending on where the specific tone pattern started and finished. For example, H at onset was the average of H, HL, and HM at the 0% point. L at offset was the average of HL, L, and ML at the 100% point. Table C.2 summarizes these values.

### Discussion and Conclusion

The results strongly suggest that there are nine different lexical tones used

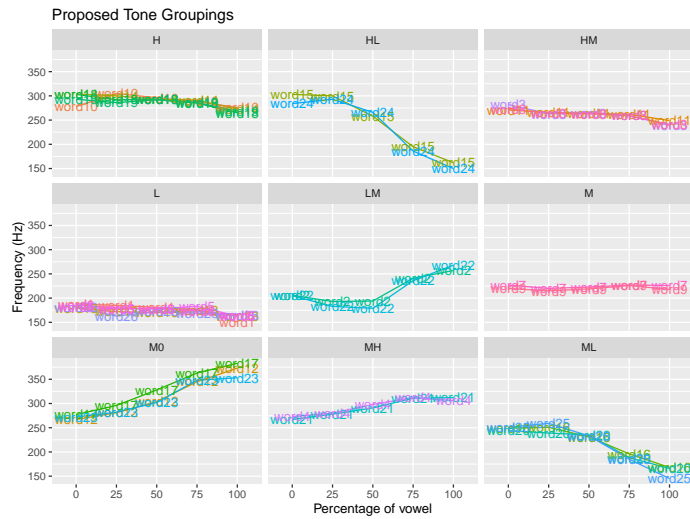


Figure C.1: Proposed tone groupings for SJQ Chatino

Tone	Onset		Offset	
	Mean	SD	Mean	SD
0	NA	NA	369.67	15.28
H	287.89	13.21	283.503	20.19
M	254.7	19.9	242.57	17.27
L	189.43	11.4	159.70	7.94

Table C.2: Mean and standard deviation of 0 (super high), H, M, and L tones at the onset (0%) and offset (100%) of the vowel

in SJQ Chatino. This was determined first by visual inspection of the tonal contours at the 0,25,50,75, and 100% points. The L, M, H, and 0 points were then averaged at the proposed onset and offsets. The mean values all go in the direction one would expect: the mean for L is lower than the mean for M which is lower than the mean for H which is lower than the mean for 0. Furthermore, the mean value for all offsets was lower than the mean value for the corresponding onsets. This follows the general pattern that  $f_0$  drifts lower throughout an utterance. Figure C.2 provides an example word spectrogram and actual tone contour for all nine proposed lexical tones.

In this lab, it was determined that SJQ Chicano has nine lexical tones. This was based off an analysis of the tonal contours. Future work would require working with speakers of the language to determine whether or not they contrast all nine of these proposed tone groupings. For example, the HM and H categories are very close in the  $f_0$  space at all points. HM is slightly lower overall, but they follow a similar shape. It may be that the observed variation is not perceivable and/or not meaningful. The type of analysis is ultimately only one tool that can be used for discovering tonal contrasts in a language.

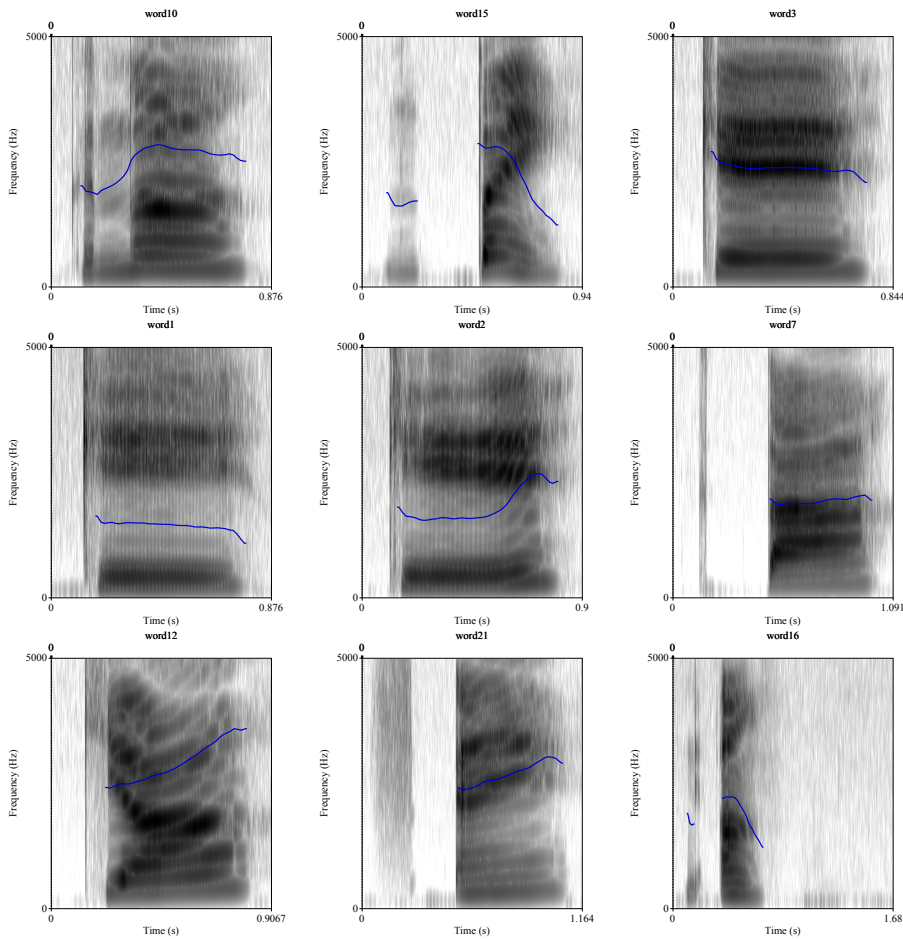


Figure C.2: Spectrograms and full tone contour of all 9 tone groups. From top to bottom within each row it is H, HL, HM, then L, LM, M, and finally MO, MH, and ML.

This would not go in the normal lab report, but I have included the Praat script so you can get an idea of how I automated the analysis.

```
# set the directory where all the files are #
directory$ = "/Users/snelson/Downloads/Lab2/"

# create a file to store the data and add a header line #
dataFile$ = directory$ + "output.csv"
writeFileLine: dataFile$, "WORD,POINT,VALUE"

# create string list of textgrids and wav files
sound_strings = Create Strings as file list: "sound_list", directory$ + "*.wav"
tg_strings = Create Strings as file list: "tg_list", directory$ + "*.TextGrid"

# get number of textgrid files
selectObject: tg_strings
nFiles = Get number of strings

# loop through textgrid files

for i to nFiles

## import sound file ##
selectObject: sound_strings
sound_fileName$ = Get string: i
Read from file: directory$ + sound_fileName$
objectName$ = replace$(sound_fileName$,".wav","",0)

## import textgrid file
selectObject: tg_strings
tg_fileName$ = Get string: i
Read from file: directory$ + tg_fileName$

## get number of intervals on transcription tier (tier 1) ##
selectObject: "TextGrid 'objectName'"
vowelinterval = Get number of intervals: 1

## create pitch object ##
```



```

selectObject: "Sound 'objectName$"
To Pitch: 0, 50, 600

## get values for pitch range over vowel ##
for j to vowelinterval
selectObject: "TextGrid 'objectName$"
label$ = Get label of interval: 1, j

if label$ == "V"
start = Get starting point: 1,j
end = Get end point: 1,j
midpoint = (start + end)/2
twentyfive = start + (end-start)/4
seventyfive = start + (3*(end-start))/4

selectObject: "Pitch 'objectName$"

f0_0 = Get value at time: start, "Hertz", "linear"
f0_25 = Get value at time: twentyfive, "Hertz", "linear"
f0_50 = Get value at time: midpoint, "Hertz", "linear"
f0_75 = Get value at time: seventyfive, "Hertz", "linear"
f0_100 = Get value at time: end, "Hertz", "linear"

appendFileLine: dataFile$, objectName$ + "," + "f0_000" + "," + fixed$(f0_0,0)
appendFileLine: dataFile$, objectName$ + "," + "f0_025" + "," + fixed$(f0_25,0)
appendFileLine: dataFile$, objectName$ + "," + "f0_050" + "," + fixed$(f0_50,0)
appendFileLine: dataFile$, objectName$ + "," + "f0_075" + "," + fixed$(f0_75,0)
appendFileLine: dataFile$, objectName$ + "," + "f0_100" + "," + fixed$(f0_100,0)

endif

endfor

endfor

```