Lecture Notes in Speech Production, Speech Coding, and Speech Recognition

Mark Hasegawa-Johnson
University of Illinois at Urbana-Champaign

February 17, 2000

Chapter 2

Speech Production

2.1 Phonetic Classification of Speech Sounds

The smallest meaningful unit of speech is called a "phone." Phones which have different meanings in a given language are called the "phonemes" of that language. Phones which sound different but mean the same thing are called "allophones." Thus, for example, /l/ and /r/ are distinct phonemes in English, but in Japanese, they are both allophones of the phoneme /r/.

Phones are almost always 50-100ms in length. "Reduced" phones may sometimes be shorter (short implementations of /t/ or the reduced vowel /ə/). Stressed tense vowels may occasionally be longer.

2.1.1 Consonants

English consonants are most simply classified on a three dimensional grid. The three dimensions are manner, place of articulation, and voicing, as shown in the following table.

Voiced							
	Labial	Dental	Alveolar	Palatal	Retroflex	Velar	Laryngeal
Stop/Affricate	/b/		/d/	/ʤ/		/g/	
Fricative	/v/	/ð/	/z/	/ʒ/			
Nasal	/m/		/n/			/ŋ/	
Glide/Liquid	/w/		/l/	/j/	/r/		/h/
Unvoiced							
	Labial	Dental	Alveolar	Palatal	Retroflex	Velar	Laryngeal
Stop/Africate	/p/		/t/	/ʧ/		/k/	
Fricative	/f/	/θ/	/s/	/ʃ/			

Manner and Voicing Features

Linguists often divide each of these three dimensions into several binary distinctions, in order to better represent useful similarities and differences. These binary distinctions are called "distinctive features." Thus, for example, there are four manners (stop/affricate, fricative, nasal, glide/liquid), so manner can be coded using two distinctive features:

	Stop/Aff	Fricative	Nasal	Glide/Liquid
Sonorant (unimpeded air flow)	-	-	+	+
Continuant (incomplete vocal tract closure)	-	+	-	+

Voiced obstruents are marked as [+voice], and unvoiced obstruents are marked as [-voice]. Sonorants may be marked [+voice], but the feature is redundant, so it need not be marked. Thus for example a voiced

stop is [-sonorant,-continuant,+voice] (3 bits), while a nasal consonant is [+sonorant,-continuant] (2 bits).

Place Features

Many systems of features have been proposed to describe place of articulation, and there is no single system which has been accepted as uncontroversial by all linguists. Since there is so much controversy, let's just look at two acoustically-based features which are particularly useful to know about if you are learning to read spectrograms.

The first thing that most people notice when they see a spectrogram is the difference between [+strident] fricatives and affricates (/s,z,ʃ,ʒ,ʧ,ʤ/) and [-strident] fricatives (/f,v,θ,ð/). A strident fricative is very loud — at the frequency of the fricative's spectral peak, a strident is higher in amplitude than neighboring vowels are at the same frequency. A non-strident is very quiet, and the frication noise may not even be visible in a spectrum.

The second important and easily observed acoustic distinction is between [+compact] obstruents (/ʃ, ʒ, ʧ, ʤ, k, g/) and [-compact] obstruents (/f,v,θ,ð,s,z/). Compact obstruents are characterized by a large, low-bandwidth spectral peak (bandwidth $\approx$ 300Hz) near the center of the audible spectrum (1000 $<$ peak frequency $<$ 3500Hz). [-compact] obstruents are characterized by no peak at all (/f,v/) or by a peak which is extremely broad (bandwidth $\approx$ 2000Hz for the sounds /s,z,θ,ð/).

2.1.2 Vowels

English vowels are also classified in a three dimensional grid. The dimensions are tongue height, tongue fronting, and tongue root tension, as shown in the following table.

Lax			Tense		
	High	Non-High	High	Mid	Low
Front	/ɪ/	/ɛ/	/i/	/e/	/æ/
Back	/ʊ/	/ʌ/	/u/	/o/	/a/

In addition to the above, all dialects of English contain a schwa (/ə/). Schwa is "reduced," meaning that it is a very short vowel, and it doesn't really have a characteristic tongue shape – the tongue just moves through from one consonant to the next.

Most linguists accept a distinctive feature notation for vowels which consists of the features high, low, front, tense, and reduced. /ə/ is [+reduced] (a one-bit representation!). All other vowels are [-reduced], with values of [high,low,front,tense] as indicated in the table above. Thus, for example, /u/ is [-reduced, +high, -low, -front, +tense] (5 bits).

In addition to the features listed above, many other languages make use of the feature [round]. A [+round] vowel (/u,ʊ,o/) is produced with rounded lips. Lip rounding lowers all of the formant frequencies. Since low F1 and low F2 are also the sign of a "non-low, back" vowel, lip rounding is used in English as a way of reinforcing the character of the non-low, back vowels.

Other languages use rounding distinctively. German, French, and all of the Scandinavian languages distinguish the vowels /y/ ([+round,+high,+front]) and /i/ ([-round,+high,+front]). Korean and Russian distinguish the vowels /ɯ/ ([-round,+high,+back]) and /u/ ([+round,+high,+back]).

2.2 A Source-Filter Model of Speech Production

The glottis has a much higher acoustic impedance than the vocal tract, and the vocal tract has a much higher acoustic impedance than open air, so it is often convenient to model speech production as the series connection of three roughly independent functions: a source function, a transfer function, and a radiation characteristic.

2.2.1 Source Function

The source of sound in the vocal tract can be either fluctuating pressure (caused by turbulence) or fluctuating flow (caused by vibration of the vocal folds). In either case, we usually assume a "high-impedance source," which means that changing the downstream impedance (by changing the shape of the vocal tract) does not change the source spectrum.

We need to be concerned with three different types of sources:

- **The voice source** $U_g(s)$ consists of puffs of air released through the glottis at a frequency of F_0 puffs per second. Each puff contains on the order of 1cm^3 of air.

  Vocal fold vibration is:

  - Periodic. "Pitch period" T_0, "Fundamental frequency" $F_0 = \frac{1}{T_0}$.
  - Smooth, like a half-wave-rectified cosine. Spectrum drops as $1/f^2$ above about 500Hz.

$$2\pi F_0 \approx \sqrt{\frac{K}{M}} \tag{2.1}$$

  K = stiffness of the vocal folds — easily adjusted for stress, singing, etc.
  M = mass of the vocal folds — characteristic of the speaker.

F_0 (Hz)	ave	min	max
men	125	80	200
women	225	150	350
children	300	200	500

- **The aspiration source** $P_h(s)$ is caused by turbulent jets of air from the glottis striking against the false vocal folds, about 2cm above the glottis. Turbulent fluctuations against an obstacle create local pockets of rapidly varying high and low pressure. The spectrum of the radiated sound spectrum is nearly flat from about 500Hz to 3000Hz, so we model it using a white noise source.

  There is always turbulence in speech. Synthetic vowels sound most natural when the amplitude of aspiration is about 10% of the amplitude of voicing. Turbulence becomes more important when the glottis is opened, as in breathy speech, or in the sound /h/.

- **The frication source** $P_f(s)$ is caused by turbulent jets of air from a vocal tract constriction somewhere between the glottis and the lips striking obstacles downstream from the constriction.

  Frication only occurs in sounds with an obstructed air flow: these sounds are called "obstruents." The obstruents in English include the fricatives (/f,v,θ,ð,s,z,ʃ,ʒ/), the stops (/p,b,t,d,k,g/), and the affricates (/ʧ,ʤ/).

  The opposite of "obstruent" is "sonorant," which has the approximate meaning "a sound you can sing." The sonorants in English are the vowels, the nasals (/n,m,ŋ/), the glides (/w,j,h/), and the liquids (/r,l/). All sonorants are excited by a mixture of voicing and aspiration; aspiration is most important in /h/, and voicing is most important in all other sonorants.

The biggest difference between aspiration and frication is that aspiration is generated close to the glottis, so the transfer function for aspiration is the same as the transfer function for voicing. Frication is generated near the lips or tongue, so the transfer function for frication may be much different from the voicing transfer function.

Stops and affricates are produced using a sequence of all three sources. First, the vocal tract is completely closed, so that there is no sound. When the stop is first released, there is a burst of frication, which lasts 5-10ms for a stop, or about 50ms for an affricate. What happens next depends on the syllable position and voicing class:

- If the stop or affricate is voiced, voicing begins at about the same time frication begins. When frication ends, voicing becomes the dominant source.

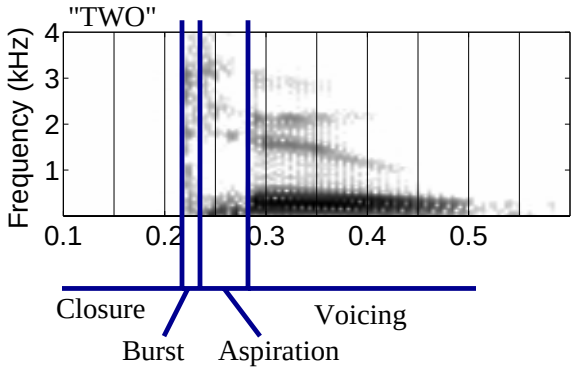


Figure 2.1: Several sources are active at the release of a stop. A period of silence is followed by a frication burst lasting 5-10ms. If the stop is unvoiced and syllable-initial, frication is followed by 30-80ms of aspiration. When the vocal folds are brought close enough together to begin vibrating, voicing begins.

- In an unvoiced stop which is also syllable-initial (e.g. "cool," "two"), voicing is delayed for 25-80ms following the end of frication. During this time, the aspiration source $P_h(s)$ is the only active source.
- In an unvoiced affricate, or an unvoiced stop which is not syllable-initial (e.g. "school," "look out!"), voicing begins at about the time that frication stops.

In review, let's look at which sources are active for each type of speech sound. The following table lists the approximate relative amplitudes of various sources during different types of speech sound, assuming that the amplitude of the primary source is 1.0.

	$U_g(s)$	$P_h(s)$	$P_f(s)$
/h/	0-0.5	1.0	0
Other Sonorant	1.0	0.1-0.3	0
Voiced Fricative, Voiced Stop Burst	0-1	0	0-1
Unvoiced Fricative, Unvoiced Stop Burst	0	0	1.0

2.2.2 Transfer Function

The transfer function is the ratio between flow or pressure at the lips, and flow or pressure at the source. As noted above, there may be two different transfer functions active at the same time: one between the glottis and the lips (which works for both voicing and aspiration), and one between the frication source and the lips.

$$T_g(s) = \frac{U_m(s)}{U_g(s)} = \frac{P_m(s)}{P_h(s)} \tag{2.2}$$

$$T_f(s) = \frac{P_m(s)}{P_f(s)} \tag{2.3}$$

Glottal Transfer Function

$T_g(s)$ is an all-pole function for most speech sounds, and even sounds which have measurable transfer function zeros (such as nasal consonants /m,n,ŋ/ and liquids /r,l/) can be synthesized reasonably well using an all-pole approximation. For this reason, we will assume that $T_g(s)$ has the form:

$$T_g(s) = \frac{1}{\prod_{n=1}^{\infty}(1 - s/s_n)(1 - s/s_n^*)} \tag{2.4}$$

$$s_n = -\pi B_n + j2\pi F_n \tag{2.5}$$

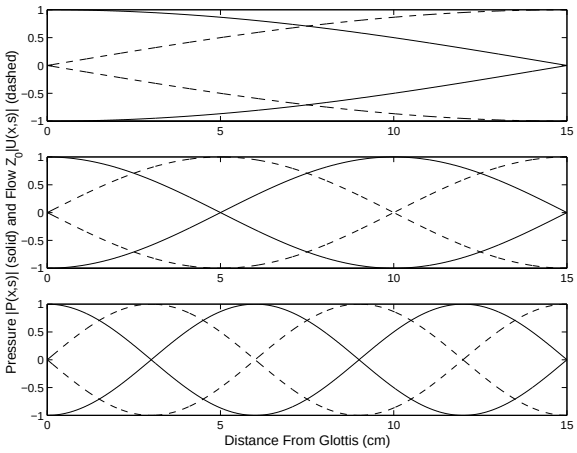


Figure 2.2: Standing wave patterns of a hard-walled acoustic resonator of uniform area, closed at $x = 0$. The amplitude of the pressure standing wave, $|P(x,s)|$, is shown using a solid line, normalized to unity magnitude. The amplitude of the flow standing wave, $Z_0(x)|U(x,s)|$, normalized to unity magnitude, is shown using a dashed line.

The pole frequencies F_n are called the "formant frequencies," and the B_n are called the "formant bandwidths." The formant frequencies are the resonant frequencies of the entire vocal tract, from glottis to lips.

Suppose, for example, that the vocal tract is modeled as a hard-walled tube of constant area, closed at one end (the glottis end) and open at the other end (the mouth end). This system will resonate (forever, if there are no losses!) at any frequency for which the standing wave pattern has zero flow at the closed end, and zero pressure at the open end. The first three standing wave patterns which qualify are shown in figure 2.2. Notice that the resonant wavelengths are such that the tube length, L, is an odd multiple of a quarter wavelength:

$$F_n = \frac{(2n-1)c}{4L} \tag{2.6}$$

This "constant-area tube" is a reasonable model of the vowel /ə/. For men ($L \approx 17.5$cm), the formants of /ə/ are at approximately 500Hz, 1500Hz, 2500Hz, and so on. For women ($L \approx 15$cm), the formants are at approximately 580Hz, 1750Hz, 2900Hz, etc.

Perturbation Theory

Suppose that we perturb the area of the tube in figure 2.2 over a short region centered at position x – say the region $x-\epsilon/2$ to $x+\epsilon/2$. Perturbing the area function $A(x)$ has the effect of modifying the stored energy W_n associated with each of the resonant frequencies F_n. Conveniently, there is a theorem in electrical network theory [22] which says that for small perturbations of $A(x)$,

$$\frac{\partial F_n}{\partial A(x)} \propto -\frac{\partial W_n}{\partial A(x)} \tag{2.7}$$

The stored energy is composed of kinetic energy and potential energy:

$$W_n(x) = T_n(x) + V_n(x) \tag{2.8}$$

The kinetic energy in a section of length ϵ can be calculated from Newton's law; it is

$$T_n(x) = \frac{1}{4}|U(x, j2\pi F_n)|^2 \frac{\rho\epsilon}{A(x)} \tag{2.9}$$

The potential energy is related to the local compression of air according to the ideal gas law; it works out to be

$$V_n(x) = \frac{1}{4}|P(x, j2\pi F_n)|^2 \frac{\epsilon A(x)}{\rho c^2} \tag{2.10}$$

Constriction Location	x	Examples	F_1	F_2	F_3
Labial	15	/u,ʊ,o,b,p,m,f,v/	⇓ 200-300	⇓ 600-1000	⇓ 2000-2200
Dental	13.5	/θ,ð/	⇓ 200-300	↓ 1300-1400	↔ 2500-2700
Alveolar	12.5	/s,z,t,d,n/	⇓ 200-300	↑ 1600-1800	⇑ 2900-3200
Palatal	11.5	/i,ɪ,j,ʃ,ʒ/	⇓ 200-300	⇑ 2000-2300	↑ 2700-2900
Velar	7-11	/e,ɛ,k,g,ŋ/	↓ 300-500	↑ 1500-2200	⇓ 1800-2500
Uvular	6-7	/o,l/	↔ 400-600	⇓ 600-1000	⇑ 2900-3200
Pharyngeal	4-5	/a,ʌ/	⇑ 800-900	⇓ 900-1200	↑ 2700-3000
Laryngopharyngeal	1-2	/æ/	↑ 500-900	↑ 1400-1800	↑ 2500-2700

Table 2.1: Important constriction locations in speech, and their approximate effect on formant frequencies. Position of the constriction (x) is given for a female speaker with a vocal tract length of 15cm. Typical formant frequencies associated with each constriction location are given for a male speaker with a vocal tract length of 17.5cm, and with neutral vowel formants of 500, 1500, and 2500Hz. Direction of the arrow indicates the change of each formant relative to the formants of a neutral vowel.

Combining and differentiating, we obtain

$$\frac{\partial F_n}{\partial A(x)} \propto \frac{\epsilon}{4\rho c^2}\left[Z_0(x)^2|U(x,j2\pi F_n)|^2 - |P(x,j2\pi F_n)|^2\right], \quad Z_0(x) \equiv \frac{\rho c}{A(x)} \tag{2.11}$$

Consider, for example, a constriction at the lips. At the lips, all of the first three standing wave patterns have a flow maximum. A negative area perturbation at the lips therefore causes negative movement in all of the first three resonant frequencies, F_1, F_2, and F_3.

Table 2.1 lists some of the particular constrictions which are important in speech. For each constriction, a value of x is given which shows approximately where that constriction is located in the plot of figure 2.2. The table shows the direction in which F_1, F_2, and F_3 move (up a lot ⇑, up a little ↑, down a little ↓, down a lot ⇓, or neither ↔) when a constriction is made at each location. In addition, approximate values of the formant frequency targets observed in male speech are given; recall that the formant frequencies of /ə/ for a male speaker are approximately 500, 1500, and 2500Hz.

Table 2.1 lists vowels in terms of "constriction location," which is an unusual way to think about vowel features. Instead, vowels are usually classified in terms of tongue height , tongue tension, and tongue fronting. Fortunately, there is a convenient mapping between distinctive features based on tongue position and the frequencies of the formants F1 and F2.

- In general, the value of F1 is inversely proportional to tongue height. If the vowels are sorted from lowest F1 to highest F1, the order would be [+high,+tense] (/u,i/), [+high,-tense] (/ɪ,ʊ/), [-high,-low] (/o,e/), [+low,-tense] /ʌ,ɛ/, [+low,+tense] (/a,æ/).
- F2 is higher for front vowels than it is for back vowels.

Finally, note that there is one more place of articulation which is not mentioned in table 2.1: laryngeal. The laryngeal consonant /h/ does not change the shape of the vocal tract, and therefore it does not change the values of the formants. The formants in /h/ are whatever would be appropriate for the phonemes on the left and right of /h/.

Formant Targets and Loci

The previous discussion is based on a false assumption called the "segmental fallacy." The segmental fallacy is the assumption that the vocal tract transfer function $T_g(s)$ is approximately static during each phoneme, and then it jumps semi-discontinuously from one phone to the next. Many useful algorithms are based on this assumption (most notably the hidden Markov model of speech recognition), but it is false. *Formants are never stationary in continuous speech.*

In continuous speech, a phone is defined by a set of formant "targets," the values of which are listed in the previos section. In continuous speech, formants move from target to target in a zig-zag fashion, sometimes reaching all of the targets, and sometimes undershooting the targets by 200-300 Hertz. Reading

a spectrogram requires you to find the formant "targets," rather than simply looking at static formant frequencies.

In general, the only quasi-static segments in continuous speech are fricatives and nasal consonants. Fricatives show a relatively static spectrum because the spectrum is dominated by the frication transfer function, $T_f(s)$, which changes less rapidly than the glottal transfer function $T_g(s)$. Nasals show a static spectrum because the resonances in a nasal spectrum are resonances of the entire throat-nose system, not resonances of $T_g(s)$. The nasal resonances are relatively constant for any given speaker; in fact, some speaker identification systems use nasal spectra as one measurement of speaker identity.

Measuring the spectrum of any [-continuant] or [-sonorant] consonant does not tell you the values of the vocal tract formants associated with that consonant. Instead, the formants associated with a consonant are measured by observing formant motion at the end of the preceding vowel, or at the beginning of the following vowel. Formants toward the end of the preceding vowel usually move smoothly toward target values which characterize the place of articulation of the consonant (some of these target values are listed in table 2.1). Formants at the beginning of the following vowel usually move smoothly away from these same target values.

Formant targets which must be inferred from movement in a neighboring segment are called "formant loci." Thus, for example, the formant loci of a labial stop or nasal consonant are $F_1 = 200$Hz, $F_2 = 1000$Hz, and $F_3 = 2000$Hz.

Frication Transfer Function

$T_f(s)$ is also an all-pole transfer function:

$$T_f(s) = \frac{1}{\prod_{n=1}^{\infty}(1 - s/s_{f,n})(1 - s/s^*_{f,n})} \tag{2.12}$$

$$s_{f,n} = -\pi B_{f,n} + j2\pi F_{f,n} \tag{2.13}$$

$$F_{f,n} \approx \frac{(2n-1)c}{4l_f} \tag{2.14}$$

where l_f is the length of the cavity in front of the constriction. Different "places of articulation" of obstruent consonants have different front cavity lengths, and the values of l_f do not overlap very much:

	l_f	$F_{f,1}$
Labial (/p,b,f,v/)	0-0.5	17700-∞
Dental (/θ,ð/)	1-1.5	5900-8800
Alveolar (/t,d,s,z/)	1.5-2.5	3500-5900
Palatal, Retroflex (/tr,dr,ʧ,ʤ,ʃ,ʒ/)	2.5-3.5	2500-3500
Velar (/k,g/)	3.5-8	1100-2500

2.2.3 Radiation Characteristic

The radiation characteristic is the ratio between the sound pressure recorded by the microphone, $P_r(s)$, and the flow or pressure at the mouth.

Calculating the amplitude of $P_r(s)$ requires you to know the position of the microphone relative to the speaker's mouth. If the amplitude of $P_r(s)$ is unimportant, it is possible to simplify the equations by calculating $P_m(s)$, the pressure at the lips, instead.

There are two cases, depending on whether the source is a flow or a pressure.

1. If the source is a pressure, then
$$R(s) = \frac{P_m(s)}{P_m(s)} = 1 \tag{2.15}$$
2. If the source is a flow, we make use of Flanagan's resistor-inductor model of the radiation impedance, and we obtain
$$R(s) = \frac{P_m(s)}{U_m(s)} = Z_r(s) \approx \frac{sL_rR_r}{R_r + sL_r} \tag{2.16}$$
where L_r and R_r are as given in [4].

2.2.4 Complete Synthesis Models

Combining sources, transfer functions, and radiation characteristics yields

$$X(s) = Z_r(s)T_g(s)U_g(s) + T_g(s)P_h(s) + T_f(s)P_f(s) \tag{2.17}$$

Since $Z_r(s)$ and $T_g(s)$ are linear, they can be commuted. Many speech synthesis algorithms are based on the equation

$$X(s) = T_g(s)(U'_g(s) + P_h(s)) + T_f(s)P_f(s) \tag{2.18}$$

$$U'_g(s) \equiv Z_r(s)U_g(s) \tag{2.19}$$

2.3 One-Dimensional Acoustic Waves

2.3.1 One-Dimensional Constitutive Equations

$$-\frac{\partial p(x,t)}{\partial x} = \frac{\rho}{A}\frac{\partial u(x,t)}{\partial t} \tag{2.20}$$

$$-\frac{\partial u(x,t)}{\partial x} = \frac{A}{\rho c^2}\frac{\partial p(x,t)}{\partial t} \tag{2.21}$$

2.3.2 One-Dimensional Acoustic Wave Equations

$$\frac{\partial^2 p(x,t)}{\partial x^2} = \frac{1}{c^2}\frac{\partial^2 p(x,t)}{\partial t^2} \tag{2.22}$$

$$\frac{\partial^2 u(x,t)}{\partial x^2} = \frac{1}{c^2}\frac{\partial^2 u(x,t)}{\partial t^2} \tag{2.23}$$

2.3.3 General Form of the Solution: Time Domain

$$p(x,t) = p_+(t - x/c) + p_-(t + x/c) \tag{2.24}$$

$$u(x,t) = u_+(t - x/c) - u_-(t + x/c) = Y_0\left[p_+(t - x/c) - p_-(t + x/c)\right] \tag{2.25}$$

$$Y_0 \equiv \frac{1}{Z_0} \equiv \frac{A}{\rho c} \tag{2.26}$$

2.3.4 General Form of the Solution: Frequency Domain

$$P(x,s) = P_+(s)e^{-sx/c} + P_-(s)e^{sx/c} \tag{2.27}$$

$$U(x,s) = Y_0\left[P_+(s)e^{-sx/c} - P_-(s)e^{sx/c}\right] \tag{2.28}$$

2.4 Parallel Admittance and Series Impedance Methods

2.4.1 Series Impedance Method

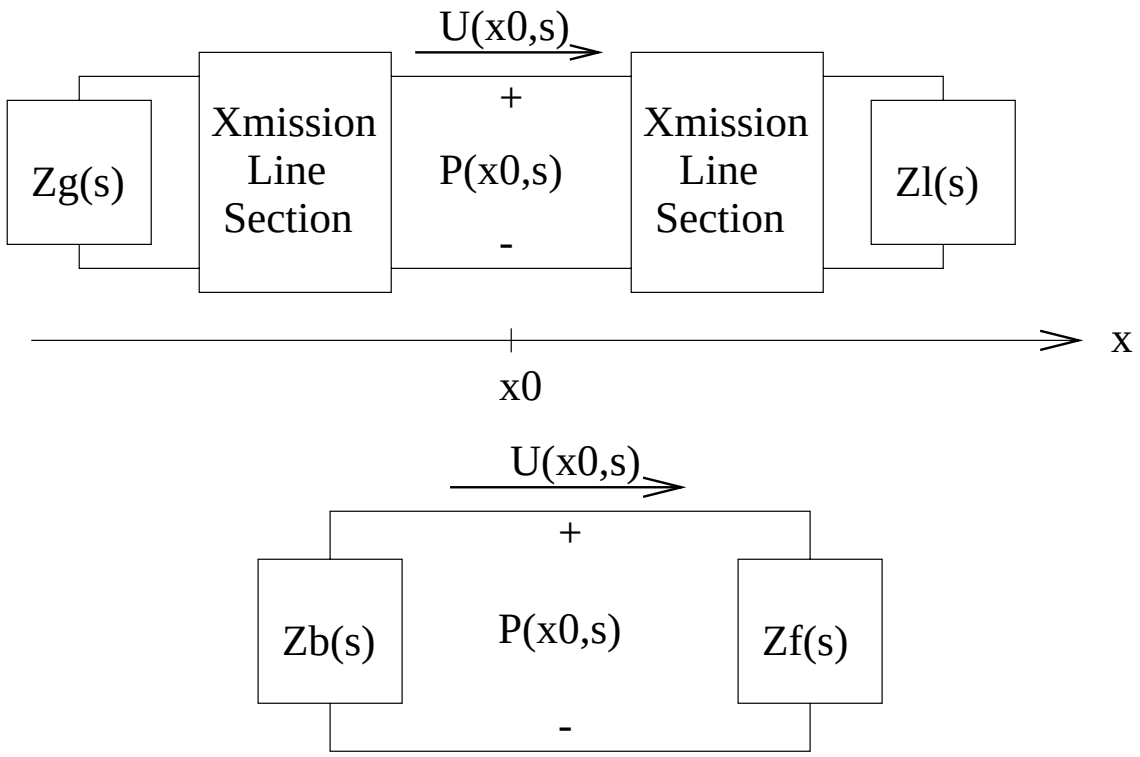
Suppose you want to know the pressure, $P(x_0,s)$, and flow, $U(x_0,s)$, at some point $x = x_0$ in a one-dimensional resonator, with no external inputs, as shown in figure 2.4.1

The part of the system at $x < x_0$ (the "back cavity") dictates a relationship between pressure and flow,

$$P(x_0,s) = -Z_b(x_0,s)U(x_0,s) \tag{2.29}$$

The part of the system at $x > x_0$ (the "front cavity") dictates another relationship,

$$P(x_0,s) = Z_f(x_0,s)U(x_0,s) \tag{2.30}$$



The pressure and flow at x_0 must satisfy both of these relationships simultaneously:

$$(Z_b(x_0,s) + Z_f(x_0,s))\, U(x_0,s) = 0 \tag{2.31}$$

$U(x_0,s)$ can only be non-zero at complex frequencies $s = j\Omega_n$ for which

$$\boxed{Z_b(x_0,j\Omega_n) + Z_f(x_0,j\Omega_n) = 0} \tag{2.32}$$

The frequencies $f_n = \Omega_n/2\pi$ which satisfy equation 2.32 are called the **resonant frequencies** of the system. The resonant frequencies are the same *regardless of your choice of* x_0.

2.4.2 Parallel Admittance Method

Equations 2.29 and 2.30 can be solved to eliminate $U(x_0,s)$, giving:

$$(Y_b(x_0,s) + Y_f(x_0,s))\, P(x_0,s) = 0 \tag{2.33}$$

Therefore $P(x_0,s)$ can only be non-zero (if there is no external excitation) at complex frequencies $s = j\Omega_n$ for which

$$\boxed{Y_b(x_0,j\Omega_n) + Y_f(x_0,j\Omega_n) = 0} \tag{2.34}$$

2.4.3 Equivalence of the Parallel Admittance and Series Impedance Methods

As it turns out, equations 2.32 and 2.34 are satisfied at exactly the same frequencies Ω_n, unless zeros in either $Z_b(j\Omega_n)Z_f(j\Omega_n)$ or $Y_b(j\Omega_n)Y_f(j\Omega_n)$ get in the way:

$$Z_b(x_0,j\Omega_n) + Z_f(x_0,j\Omega_n) = \frac{Y_f(x_0,j\Omega_n) + Y_b(x_0,j\Omega_n)}{Y_b(x_0,j\Omega_n)Y_f(x_0,j\Omega_n)} \tag{2.35}$$

$$Y_b(x_0, j\Omega_n) + Y_f(x_0, j\Omega_n) = \frac{Z_f(x_0, j\Omega_n) + Z_b(x_0, j\Omega_n)}{Z_b(x_0, j\Omega_n) Z_f(x_0, j\Omega_n)} \tag{2.36}$$

The frequencies which satisfy equations 2.32 and 2.34 are called the "resonant frequencies" of the system.

2.4.4 Summary

To calculate resonant frequencies using the method of series impedances/parallel admittances:

1. Choose any x coordinate, x_0. The resonant frequencies *do not depend on your choice of* x_0, so you usually want to choose an x_0 which makes the math as easy as possible.
2. The "back cavity" places a constraint on the relationship between $P(x_0, s)$ and $U(x_0, s)$. Express this constraint as an impedance:
$$P(x_0, s) = -Z_b(x_0, s) U(x_0, s) \tag{2.37}$$
3. The "front cavity" also places a constraint on the pressure-flow relationship at x_0. Express this constraint as another impedance:
$$P(x_0, s) = Z_f(x_0, s) U(x_0, s) \tag{2.38}$$
4. Find the resonant frequencies $f_n = \Omega_n / 2\pi$ by solving either equation 2.32 or equation 2.34. The answer will be the same *regardless of which equation you solve* (except in pathological cases), so you should usually choose whichever equation which is easier to solve:
$$\boxed{Z_b(x_0, j\Omega_n) + Z_f(x_0, j\Omega_n) = 0} \tag{2.39}$$
$$\boxed{Y_b(x_0, j\Omega_n) + Y_f(x_0, j\Omega_n) = 0} \tag{2.40}$$

2.4.5 The Four Basic Lossless Impedances

Hard Wall

A "hard wall" is a wall into which air can not flow. In other words, perpendicular to the wall,

$$U(x_0, j\Omega) = 0 \tag{2.41}$$

$$\boxed{Y(x_0, j\Omega) = \frac{U(x_0, j\Omega)}{P(x_0, j\Omega)} = 0} \tag{2.42}$$

The impedance of a hard wall is undefined:

$$Z(x_0, j\Omega) = \frac{P(x_0, j\Omega)}{0} \tag{2.43}$$

Empty Space

We often pretend that a finite flow $U(x_0, j\Omega)$ out into "empty space" can never cause any significant change in pressure:

$$P(x_0, j\Omega) = 0 \tag{2.44}$$

Based on this assumption, we can write that the impedance of empty space is

$$\boxed{Z(x_0, j\Omega) = \frac{P(x_0, j\Omega)}{U(x_0, j\Omega)} = 0} \tag{2.45}$$

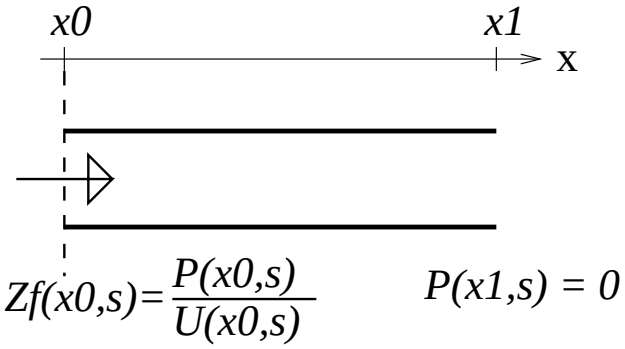


Figure 2.3: Tube open at opposite end.

Of course this is not strictly true — when flow leaves a finite-area section to enter an infinite-area section (e.g. when flow leaves the lips), there is a non-zero radiation impedance $Z(x_0, j\Omega) = Z_r(j\Omega)$. However, the approximation $Z(x_0, j\Omega) \approx 0$ is useful in many situations.

The admittance of empty space is undefined:

$$Y(x_0, j\Omega) = \frac{U(x_0, j\Omega)}{0} \tag{2.46}$$

Tube Open at Opposite End

This section will derive the front-part impedance $Z_f(x_0, s)$ and admittance $Y_f(x_0, s)$ of a tube whose opposite end opens into empty space, as shown above. Note that:

1. In order to calculate a "front part" impedance $Z_f(x_0, s)$, we only look at the shape of the tube for $x > x_0$ (that is, the part which is shown in the figure above).
2. In order to calculate a "back part" impedance $Z_b(x_0, s)$, we would only look at the shape of the tube for $x < x_0$ (which is not shown above!). The calculations for Z_b are otherwise identical to the calculations for Z_f.

Since this section is concerned with a front-part impedance, we will only concern ourselves with **one boundary condition**: the "empty space" boundary at $x_1 > x_0$.

1. **Boundary Condition:** $P(x_1, j\Omega) = 0$

   First, evaluate the general form of $P(x, j\Omega)$ at $x = x_1$:

   $$P(x_1, j\Omega) = P_+(j\Omega)e^{-j\Omega x_1/c} + P_-(j\Omega)e^{j\Omega x_1/c} \tag{2.47}$$

   The boundary condition is

   $$P(x_1, j\Omega) = 0 \tag{2.48}$$

   which yields

   $$P_-(j\Omega) = -P_+(j\Omega)e^{-2j\Omega x_1/c} \tag{2.49}$$

2. **Calculate the impedance at** $x = x_0$

   Next, let's evaluate at $x = x_0$ to get the impedance:

   $$P(x_0, j\Omega) = P_+(j\Omega)e^{-j\Omega x_0/c} + P_-(j\Omega)e^{j\Omega x_0/c} = j(2P_+(j\Omega)e^{-j\Omega x_1/c})\sin(\Omega(x_1 - x_0)/c) \tag{2.50}$$

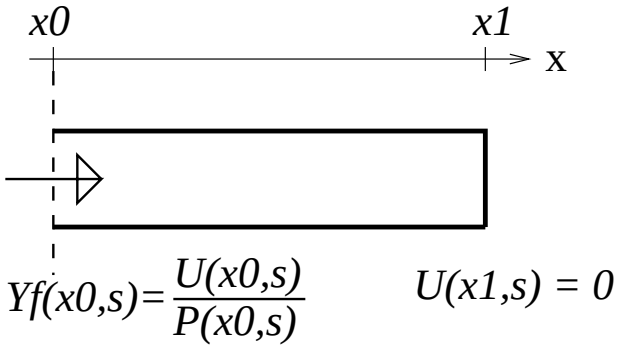


Figure 2.4: Tube closed at opposite end.

$$U(x_0, j\Omega) = Y_0 \left[P_+ e^{-j\Omega x_0/c} - P_- e^{j\Omega x_0/c} \right] = Y_0 (2P_+(j\Omega) e^{-j\Omega x_1/c}) \cos(\Omega(x_1 - x_0)/c) \tag{2.51}$$

$$\boxed{Z_f(x_0, j\Omega) = \frac{P(x_0,j\Omega)}{U(x_0,j\Omega)} = jZ_0 \tan(\Omega(x_1 - x_0)/c)} \tag{2.52}$$

$$\boxed{Y_f(x_0, j\Omega) = \frac{U(x_0,j\Omega)}{P(x_0,j\Omega)} = \frac{1}{jZ_0 \tan(\Omega(x_1 - x_0)/c)}} \tag{2.53}$$

Tube Closed at Opposite End

This section derives the "front-part" impedance $Z_f(x_0, s)$ and admittance $Y_f(x_0, s)$ of a tube whose opposite end is a hard wall, as shown above.

1. **Boundary Condition:** $U(x_1, j\Omega) = 0$

$$U(x_1, j\Omega) = Y_0 \left[P_+(j\Omega) e^{-j\Omega x_1/c} - P_-(j\Omega) e^{j\Omega x_1/c} \right] \tag{2.54}$$

The boundary condition is

$$U(x_1, j\Omega) = 0 \tag{2.55}$$

therefore

$$P_-(j\Omega) = P_+(j\Omega) e^{-2j\Omega x_1/c} \tag{2.56}$$

2. **Calculate the impedance at** $x = x_0$

$$P(x_0, j\Omega) = P_+(j\Omega) e^{-j\Omega x_0/c} + P_-(j\Omega) e^{j\Omega x_0/c} = (2P_+(j\Omega) e^{-j\Omega x_1/c}) \cos(\Omega(x_1 - x_0)/c) \tag{2.57}$$

$$U(x_0, j\Omega) = Y_0 \left[P_+ e^{-j\Omega x_0/c} - P_- e^{j\Omega x_0/c} \right] = jY_0 (2P_+(j\Omega) e^{-j\Omega x_1/c}) \sin(\Omega(x_1 - x_0)/c) \tag{2.58}$$

$$\boxed{Y_f(x_0, j\Omega) = \frac{U(x_0,j\Omega)}{P(x_0,j\Omega)} = jY_0 \tan(\Omega(x_1 - x_0)/c)} \tag{2.59}$$

$$\boxed{Z_f(x_0, j\Omega) = \frac{P(x_0,j\Omega)}{U(x_0,j\Omega)} = \frac{1}{jY_0 \tan(\Omega(x_1 - x_0)/c)}} \tag{2.60}$$

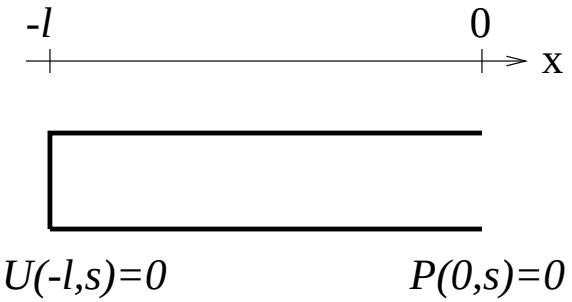


Figure 2.5: Quarter-wave resonator

2.5 Concatenated Tubes Method for Finding Resonances

2.5.1 The Four Basic One-Dimensional Resonators

Quarter-Wave Resonator

This section will derive the resonant frequencies of the tube above (a "quarter-wave resonator") using four different methods: a graphical standing-wave method, a direct Laplace-transform method, the series impedance method evaluated at $x_0 = -l/2$, and the parallel admittance method evaluated at $x_0 = -l$.

- **Standing-Wave Method**

  The boundary conditions on this tube are

$$P(0, j\Omega) = 0 \tag{2.61}$$

$$U(-l, j\Omega) = 0 \tag{2.62}$$

  The resonant frequencies of the tube are the frequencies at which standing waves in the tube match these two boundary conditions. The first three standing wave patterns which match the boundary conditions are pictured in figure 2.5.1

  It is obvious that in order to meet both boundary conditions, the length of the tube must be an odd multiple of $\lambda/4$, where λ is the wavelength of the standing wave:

$$l = \frac{\lambda_n}{4}, \frac{3\lambda_n}{4}, \ldots, (2n-1)\frac{\lambda_n}{4}, \quad n = 1, 2, \ldots \tag{2.63}$$

  Solving for the resonant frequencies f_n, we get

$$\boxed{f_n = \tfrac{c}{\lambda_n} = (2n-1)c/4l} \tag{2.64}$$

- **Laplace-Transform Method**

  The boundary conditions on the tube are, again,

$$P(0, j\Omega) = 0 \tag{2.65}$$

$$U(-l, j\Omega) = 0 \tag{2.66}$$

  1. **Boundary Condition 1:** $P(0, j\Omega) = 0$

     The general solution to the wave equation at $x = 0$ is

$$P(0, j\Omega) = P_+(j\Omega)e^0 + P_-(j\Omega)e^0 = 0 \tag{2.67}$$

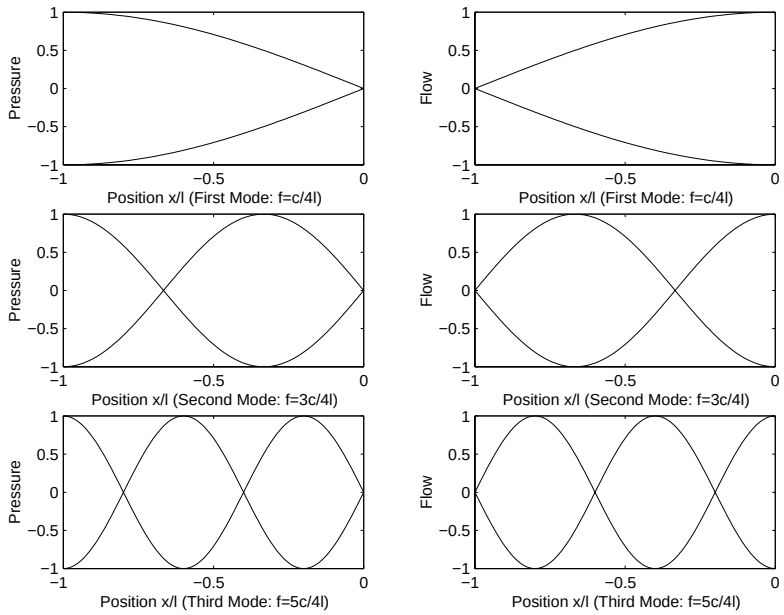


Figure 2.6: Modes of a quarter-wave resonator.

which yields

$$P_+ = -P_- \tag{2.68}$$

2. **Boundary Condition 2:** $U(-l, j\Omega) = 0$
   The general solution at $x = -l$ is

$$U(-l, j\Omega) = Y_0 \left[P_+(j\Omega) e^{j\Omega l/c} - P_-(j\Omega) e^{-j\Omega l/c} \right] = 0 \tag{2.69}$$

3. **Solve:**
   Combining equations 2.68 and 2.69 gives

$$U(-l, j\Omega) = 2Y_0 P_+(j\Omega) \cos(\Omega l/c) = 0 \tag{2.70}$$

   Therefore, $P_+(j\Omega)$ can only be nonzero at frequencies Ω_n such that

$$\cos(\Omega_n l/c) = 0 \tag{2.71}$$

$$\frac{\Omega_n l}{c} = \frac{\pi}{2}, \frac{3\pi}{2}, \frac{5\pi}{2}, \ldots \tag{2.72}$$

$$\boxed{f_n = (2n-1)c/4l, \quad n = 1, 2, \ldots} \tag{2.73}$$

- **Series Impedance Method at** $x_0 = -l/2$
  For any x_0, the resonant frequencies Ω_n satisfy

$$Z_b(x_0, j\Omega_n) + Z_f(x_0, j\Omega_n) = 0 \tag{2.74}$$

  Suppose we cut the tube at $x_0 = -l/2$. The front-part and back-part impedances are

$$Z_f(-l/2, j\Omega) = jZ_0 \tan(\Omega l/2c) \tag{2.75}$$

$$Z_b(-l/2, j\Omega) = \frac{1}{jY_0 \tan(\Omega l/2c)} \tag{2.76}$$

Substituting equations 2.75 and 2.76 into 2.74, we get

$$jZ_0 \tan(\Omega_n l/2c) + \frac{1}{jY_0 \tan(\Omega_n l/2c)} = 0 \tag{2.77}$$

Re-arranging, and noting that $Z_0 = 1/Y_0$, we get

$$\tan^2(\Omega_n l/2c) = 1 \tag{2.78}$$

Which has solutions at

$$\frac{\Omega_n l}{2c} = \frac{\pi}{4}, \frac{3\pi}{4}, \frac{5\pi}{5}, \ldots \tag{2.79}$$

$$\boxed{f_n = (2n-1)c/4l, \quad n = 1, 2, \ldots} \tag{2.80}$$

- **Parallel Admittance Method at $x_0 = -l$**

  For any x_0, the resonant frequencies Ω_n satisfy

$$Y_b(x_0, j\Omega_n) + Y_f(x_0, j\Omega_n) = 0 \tag{2.81}$$

The above equation is extremely easy to solve at $x_0 = -l$, because the "back part" is just a hard wall:

$$Y_b(-l, j\Omega) = 0 \quad \text{for all } \Omega \tag{2.82}$$

Substituting equation 2.82 into equation 2.81, we find that the resonant frequencies Ω_n satisfy

$$Y_f(j\Omega_n) = 0 \tag{2.83}$$

$$\frac{1}{jZ_0 \tan(\Omega_n l/c)} = 0 \tag{2.84}$$

$$\frac{\Omega_n L}{c} = \frac{\pi}{2}, \frac{3\pi}{2}, \ldots, (2n-1)\frac{\pi}{2}, \quad n = 1, \ldots \tag{2.85}$$

$$\boxed{f_n = (2n-1)c/4l, \quad n = 1, 2, \ldots} \tag{2.86}$$

Open-Ended Half-Wave Resonator

- **Question:**

  Which is easiest to use here, the impedance method, or the admittance method?

- **Answer:**

  The "back part" impedance at $x = -l$ is

$$Z_b(-l, j\Omega) = 0 \quad \text{for all } \Omega \tag{2.87}$$

The impedance method therefore yields the following trivial formula for the resonant frequencies Ω_n:

$$Z_f(-l, j\Omega_n) = 0 \tag{2.88}$$

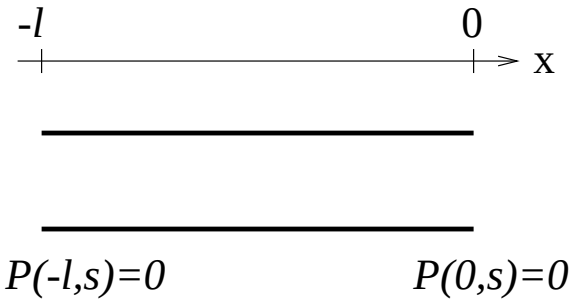


Figure 2.7: Open-ended half-wave resonator.

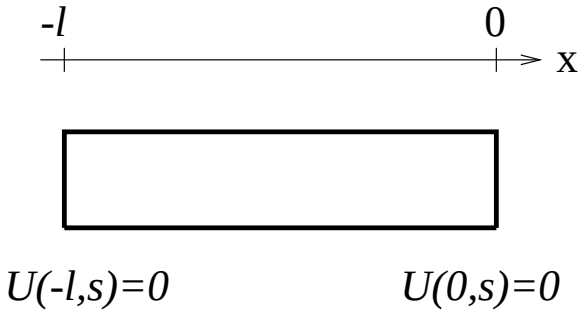


Figure 2.8: Closed-ended half-wave resonator.

which can be solved to yield:

$$jZ_0 \tan(\Omega_n l/c) = 0 \tag{2.89}$$

$$\frac{\Omega_n l}{c} = 0, \pi, 2\pi, \ldots, n\pi, \quad n = 0, \ldots \tag{2.90}$$

$$\boxed{f_n = nc/2l} \tag{2.91}$$

At the resonant frequencies, the length of the tube l is a multiple of half the wavelength λ_n:

$$l = \frac{nc}{2f_n} = n\left(\frac{\lambda_n}{2}\right) \tag{2.92}$$

Closed-Ended Half-Wave Resonator

- **Question:**
  Which is easiest to use here, the impedance method, or the admittance method?
- **Answer:**
  The "back part" admittance at $x = -l$ is

$$Y_b(-l, j\Omega) = 0 \quad \text{for all } \Omega \tag{2.93}$$

The admittance method therefore yields the following trivial formula for the resonant frequencies Ω_n:

$$Y_f(-l, j\Omega_n) = 0 \tag{2.94}$$

which can be solved to yield:

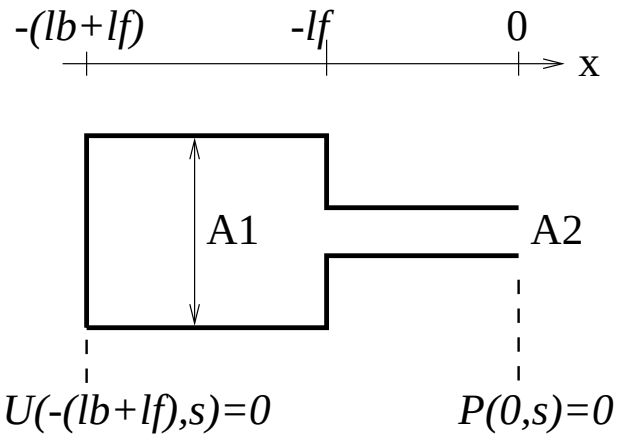


Figure 2.9: Helmholtz resonator.

$$jY_0 \tan(\Omega_n L/c) = 0 \tag{2.95}$$

$$\frac{\Omega_n l}{c} = 0, \pi, 2\pi, \ldots, n\pi, \quad n = 0, \ldots \tag{2.96}$$

$$\boxed{f_n = nc/2l} \tag{2.97}$$

At the resonant frequencies, the length of the tube l is a multiple of half the wavelength:

$$l = \frac{nc}{2f_n} = n\left(\frac{\lambda_n}{2}\right) \tag{2.98}$$

Helmholtz Resonator

In a two-tube problem like this one, the series impedance method and the parallel admittance method are both easiest to solve at the boundary between the two tubes, $x_0 = -l_f$. For example, the series impedance method gives us this formula for the resonances Ω_n:

$$Z_b(-l_f, j\Omega_n) + Z_f(-l_f, j\Omega_n) = 0 \tag{2.99}$$

where the impedances are

$$Z_b(-l_f, j\Omega) = \frac{1}{jY_{0,b}\tan(\Omega l_b/c)} = \frac{1}{j(A_b/\rho c)\tan(\Omega l_b/c)} \tag{2.100}$$

$$Z_f(-l_f, j\Omega) = jZ_{0,f}\tan(\Omega_n l_f/c) = j(\rho c/A_f)\tan(\Omega l_f/c) \tag{2.101}$$

At high frequencies, equation 2.99 can be solved using the method of concatenated tubes discussed in the next section. At low frequencies, however, we make the approximation that

$$\tan(x) \approx x \tag{2.102}$$

And we get

$$Z_b(-l_f, j\Omega) \approx \frac{1}{j\Omega(A_b l_b/\rho c^2)} \equiv \frac{1}{j\Omega C_b} \tag{2.103}$$

$$Z_f(-l_f, j\Omega) \approx j\Omega\left(\frac{\rho l_f}{A_f}\right) \equiv j\Omega M_f \tag{2.104}$$

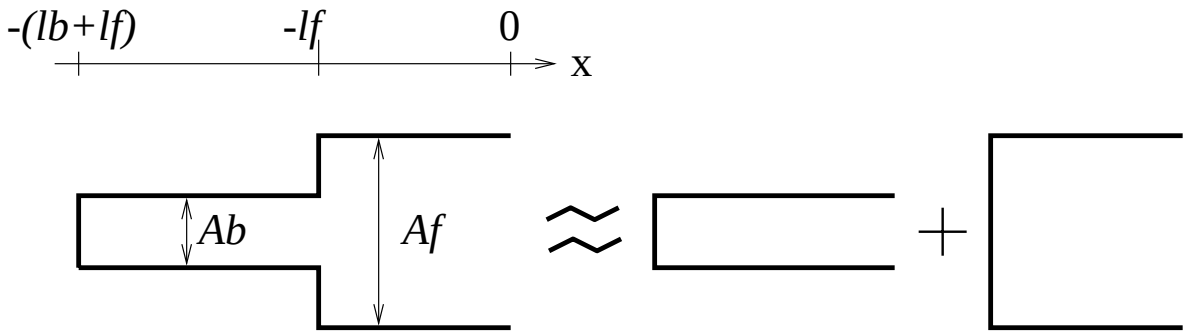


Figure 2.10: Two-tube model of the vowel /aa/.

where C_b and M_f are called the "acoustic compliance" and "acoustic mass," respectively, and correspond to the capacitance and inductance in an electrical analogy. The first non-zero resonant frequency is

$$\Omega_1 = \frac{1}{\sqrt{M_f C_b}} = \sqrt{\frac{c^2 A_f}{A_b l_b l_f}} \tag{2.105}$$

$$\boxed{F_1 = \frac{1}{2\pi}\sqrt{\frac{c^2 A_f}{A_b l_b l_f}}} \tag{2.106}$$

2.5.2 A Two-Tube Example: /aa/

The exact resonances of this system are found using the equation:

$$Z_b(-l_f, j\Omega_n) + Z_f(-l_f, j\Omega_n) = \frac{1}{jY_{0,b}\tan(\Omega_n l_b/c)} + jZ_{0,f}\tan(\Omega_n l_f/c) = 0 \tag{2.107}$$

If $A_b \ll A_f$, then $\frac{1}{Y_{0,b}} \gg Z_{0,f}$, and the solutions to equation 2.107 are approximately the solutions to the following two equations:

$$Z_b(-l_f, j\Omega_n) + 0 \approx 0 \tag{2.108}$$

$$-\infty + Z_f(-l_f, j\Omega_n) \approx 0, \quad \text{or more correctly,} \quad 0 + Y_f(-l_f, j\Omega_n) \approx 0 \tag{2.109}$$

But equations 2.108 and 2.109 are exactly the same resonant equations we would get by "decoupling" the tubes, as shown in the figure above! The resonances of the back and front tube sections separately are:

1. Resonances of section 1: $f_{n,1} = (2n-1)c/4l_b$
2. Resonances of section 2: $f_{n,2} = (2n-1)c/4l_f$

So, according to equations 2.108 and 2.109, the resonances of the system include all of the resonances $f_{n,1}$, plus all of the resonances $f_{n,2}$. For example, if $l_b = 8$cm, $l_f = 9$cm, and $c = 35000$cm/s, then the resonances of the complete system are

$$F_n \approx \frac{c}{4l_f}, \frac{c}{4l_b}, \frac{3c}{4l_f}, \ldots \tag{2.110}$$

$$\approx 970\text{Hz}, 1090\text{Hz}, 2910\text{Hz}, 3270\text{Hz}, \ldots \tag{2.111}$$

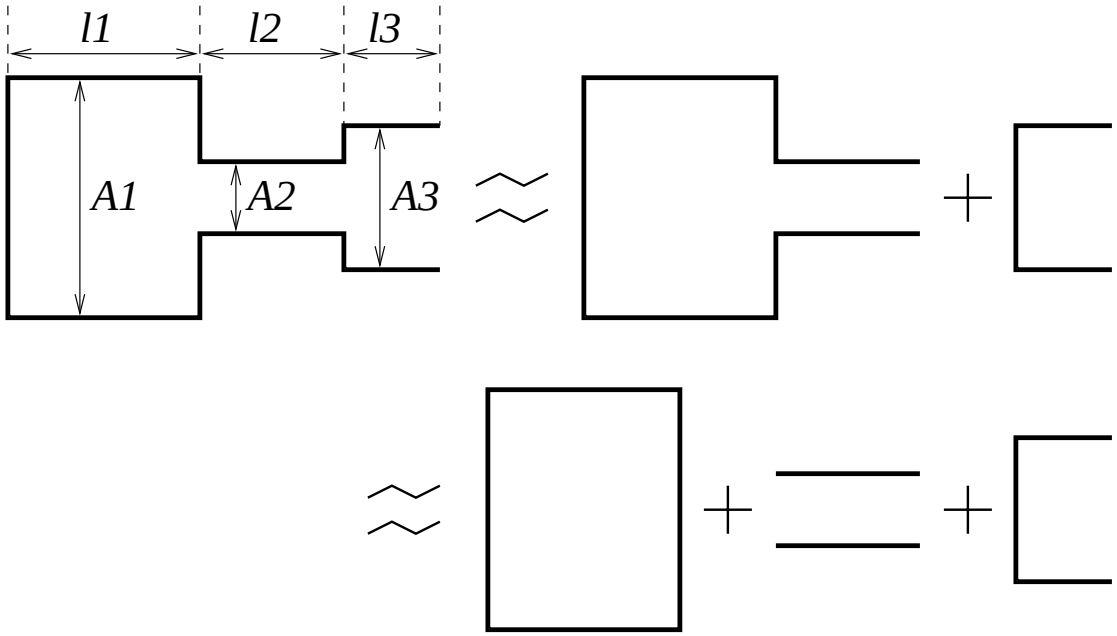


Figure 2.11: Three-tube model of the vowel /iy/.

2.5.3 A Three-Tube Example: /iy/

If $A_1 \gg A_2$ and $A_2 \ll A_3$, the resonances of this system can be calculated approximately by "decoupling" the tubes as shown. At very low frequencies, the back two sections act like a Helmholtz resonator:

$$F_1 = \frac{1}{2\pi\sqrt{C_1 M_2}} = \frac{1}{2\pi}\sqrt{\frac{c^2 A_2}{A_1 l_1 l_2}} \tag{2.112}$$

At higher frequencies, all three tubes can be considered individually, and we have

1. Resonances of section 1: $f_{n,1} = nc/2l_1$
2. Resonances of section 2: $f_{n,2} = nc/2l_2$
3. Resonances of section 3: $f_{n,3} = (2n-1)c/4l_3$

The resonances of the complete system include the Helmholtz resonance, F_1, and the resonances of the individual sections, $f_{n,1}$, $f_{n,2}$, and $f_{n,3}$.

2.6 Summary

1. The frequency-domain solution to the acoustic wave equations is

$$P(x,s) = P_+(s)e^{-sx/c} + P_-(s)e^{sx/c} \tag{2.113}$$

$$U(x,s) = Y_0\left[P_+(s)e^{-sx/c} - P_-(s)e^{sx/c}\right] \tag{2.114}$$

$$Y_0 \equiv \frac{1}{Z_0} \equiv \frac{A}{\rho c} \tag{2.115}$$

2. The series impedance/parallel admittance method for finding resonances consists of the following steps:
   (a) Choose a convenient x_0.

(b) Derive $Z_f(x_0, j\Omega)$.

(c) Derive $Z_b(x_0, j\Omega)$.

(d) Solve either of the following equations to find the resonances Ω_n:

$$Z_b(x_0, j\Omega_n) + Z_f(x_0, j\Omega_n) = 0 \tag{2.116}$$

$$Y_b(x_0, j\Omega_n) + Y_f(x_0, j\Omega_n) = 0 \tag{2.117}$$

3. The four basic lossless acoustic impedances are:

   (a) Hard wall: $Y(x_0, j\Omega) = 0$.

   (b) Empty space: $Z(x_0, j\Omega) = 0$.

   (c) Tube of length l closed at the opposite end: $Y(x_0, j\Omega) = jY_0 \tan(\Omega l/c)$.

   (d) Tube of length l open at the opposite end: $Z(x_0, j\Omega) = jZ_0 \tan(\Omega l/c)$.

4. The four basic one-dimensional resonators are:

   (a) Tube closed at one end, open at the other: $f_n = (2n-1)c/4l,\ n = 1, 2, \ldots$.

   (b) Tube closed at both ends: $f_n = nc/2l,\ n = 0, 1, \ldots$.

   (c) Tube open at both ends: $f_n = nc/2l,\ n = 0, 1, \ldots$.

   (d) Helmholtz resonator: $F_1 = (2\pi\sqrt{C_b M_f})^{-1},\ \ C_b \equiv (A_b l_b/\rho c^2),\ \ M_f \equiv (\rho l_f/A_f)$.

5. The concatenated-tubes method for finding the approximate resonant frequencies of a system consists of the following steps:

   (a) Break apart the system at every rapid change in area.

   (b) Identify the resonances of each decoupled tube segment.

   (c) Look for pairs of decoupled tube segments which might recouple, at low frequencies, to produce a Helmholtz resonance. Calculate the Helmholtz resonance of each such pair.

   (d) Combine and sort all of the resonant frequencies from parts (b) and (c) to get the resonances of the whole system.

2.7 Reflection and Transmission

2.7.1 Continuity of Flow and Pressure at a Boundary

Suppose tubes of admittances $Y_{0,k} = (A_k/\rho c)$ and $Y_{0,k+1} = (A_{k+1}/\rho c)$ are joined at $x = x_0$, so that the cross-sectional area of the vocal tract is discontinuous at that point. Pressure and flow are constrained to be continuous across the boundary:

$$u_k(x_0, t) = u_{k+1}(x_0, t) \tag{2.118}$$

$$p_k(x_0, t) = p_{k+1}(x_0, t) \tag{2.119}$$

The pressures and flows can each be written in terms of forward and backward traveling waves:

$$u_k^+(t - x_0/c) - u_k^-(t + x_0/c) = u_{k+1}^+(t - x_0/c) - u_{k+1}^-(t + x_0/c) \tag{2.120}$$

$$\frac{1}{Y_{0,k}}(u_k^+(t - x_0/c) + u_k^-(t + x_0/c)) = \frac{1}{Y_{0,k+1}}(u_{k+1}^+(t - x_0/c) + u_{k+1}^-(t + x_0/c)) \tag{2.121}$$

Solving for the outgoing waves, $u_{k+1}^+(t - x_0/c)$ and $u_k^-(t + x_0/c)$, produces the equations

$$\boxed{u_{k+1}^+(t - x_0/c) = (1 + r_k)u_k^+(t - x_0/c) + r_k u_{k+1}^-(t + x_0/c)} \tag{2.122}$$

$$\boxed{u_k^-(t + x_0/c) = (1 - r_k)u_{k+1}^-(t + x_0/c) - r_k u_k^+(t - x_0/c)} \tag{2.123}$$

where the reflection coefficient is defined to be

$$\boxed{r_k \equiv \frac{Y_{0,k+1} - Y_{0,k}}{Y_{0,k+1} + Y_{0,k}} = \frac{A_{k+1} - A_k}{A_{k+1} + A_k}} \tag{2.124}$$

2.7.2 Termination

Termination at the Glottis

The reflection coefficient at the glottis depends on the glottal admittance, Y_g, and the area of the first vocal tract section, A_1:

$$r_g \equiv \frac{\frac{A_1}{\rho c} - Y_g}{\frac{A_1}{\rho c} + Y_g} \tag{2.125}$$

If $Y_g = 0$ (i.e. no flow into the glottis), then $r_g = 1$.

Termination at the Lips

The reflection coefficient at the lips depends on the radiation impedance, Z_r, and the area of the last vocal tract section, A_N:

$$r_L \equiv \frac{Y_r - Y_{0,N}}{Y_r + Y_{0,N}} = \frac{\frac{1}{Z_r} - \frac{A_N}{\rho c}}{\frac{1}{Z_r} + \frac{A_N}{\rho c}} \tag{2.126}$$

If $Z_L \approx 0$ (zero pressure at the lips), the reflection coefficient at the lips is $r_L = 1$.

2.8 Transfer Functions

2.8.1 Uniform Tube

Imagine a uniform tube of length l, terminated by a reflection coefficient of r_L at the lips, and a reflection coefficient of r_g at the glottis.

Suppose the flow from the glottis is an impulse at time zero,

$$u_g^+(t) = \delta(t) \tag{2.127}$$

At time $\tau = l/c$, the forward-going impulse will arrive at the lips. It will be transmitted with a transmission coefficient of $1 + r_L$, and also reflected with a reflection coefficient of $-r_L$. At time $\tau = 2l/c$, the backward-going impulse arrives at the glottis. The wave is transmitted through the glottis with a transmission coefficient of $1 - r_g$, and reflected forward again with a reflection coefficient of r_g.

This process continues forever. The flow from the lips is thus

$$u_L(t) = (1 + r_L)\left\{\delta(t - l/c) - r_L r_g \delta(t - 3l/c) + \ldots + (-r_L r_g)^k \delta(t - 2kl/c - l/c) + \ldots\right\}, k = 0, 1, \ldots \tag{2.128}$$

The transfer function is

$$H(s) = \frac{U_L(s)}{U_g(s)} = (1 + r_L)e^{-sl/c} \sum_{k=0}^{\infty} (-r_L r_g e^{-2sl/c})^k \tag{2.129}$$

$$H(j\Omega) = \frac{(1 + r_L)e^{-j\Omega l/c}}{1 + r_L r_g e^{-2j\Omega l/c}} = \frac{(1 + r_L)/2}{\frac{1 - r_L r_g}{2} e^{j\Omega l/c} + r_L r_g \cos(\Omega l/c)} \tag{2.130}$$

- If $r_L = 1$ and $r_g = 1$, the impulse response $h(n)$ has the same magnitude for all time, and the peaks of the transfer function $H(j\Omega)$ have infinite magnitude. In this case, the system is said to be "lossless."
- If $|r_L| < 1$ or $|r_g| < 1$, part of the energy of the system is transmitted through the glottis or the lips during every period of $2l/c$. As a result, the impulse response decays at a rate of $(r_L r_g)^{tc/2l}$, and the peaks of the transfer function do not have infinite magnitude. In this case, the system is said to be "lossy."

2.8.2 General Case

Now suppose that the vocal tract is composed of N concatenated tube sections, each of length l. Let the forward glottal flow be

$$u_g^+(t) = \delta(t) \tag{2.131}$$

The first impulse will arrive at the lips at time $N\tau = Nl/c$. The second impulse will arrive at time $N\tau + 2\tau$, with a magnitude which depends on the various reflection coefficients. In general, another impulse will arrive at the lips with some magnitude α_k at time $t = N\tau + 2k\tau$ for all $k = 0, 1, \ldots$:

$$h(t) = \sum_{k=0}^{\infty} \alpha_k \delta(t - N\tau - 2k\tau) \tag{2.132}$$

$$H(s) = e^{-N\tau s} \sum_{k=0}^{\infty} \alpha_k e^{-2k\tau s} = e^{-N\tau s} \hat{H}(s) \tag{2.133}$$

$$\hat{H}(j\Omega) = \sum_{k=0}^{\infty} \alpha_k e^{-j\Omega 2k\tau} = \hat{H}(j(\Omega + \frac{2\pi}{2\tau})) \tag{2.134}$$

The spectrum $\hat{H}(j\Omega)$ is periodic with period $\frac{2\pi}{2\tau}$; if N is even, $H(j\Omega)$ is periodic with the same period. This suggests that $H(j\Omega)$ is a discrete-time transfer function with a sampling period of $T = 2\tau$, and indeed that is the case. If we define

$$z \equiv e^{-Ts} = e^{-2\tau s} \tag{2.135}$$

Then $H(s)$ can be written

$$H(z) = z^{-\frac{N}{2}} \sum_{k=0}^{\infty} \alpha_k z^{-k} \tag{2.136}$$

Although $H(s) = H(z)$ is already a discrete-time transfer function, $u_g(t)$ and $u_L(t)$ have not yet been sampled. The input signal can be sampled as shown:

$$u_g[n] \equiv u_g(2n\tau) \tag{2.137}$$

Yielding the following model of the vocal tract:

$$u_L[n] = u_g[n] * \hat{h}[n] \tag{2.138}$$

$$u_L(2\tau(n + \frac{N}{2})) = u_L[n] \tag{2.139}$$

In words, the digital output signal $u_L[n]$ is obtained by filtering $u_g[n]$ through the transfer function $\hat{h}[n]$. The continuous output $u_L(t)$ is constructed by delaying the samples of $u_L[n]$ by $N\tau$ seconds ($N/2$ samples), and then passing them through a D/A converter.

2.9 How to Read Spectrograms

Reading a spectrogram is kind of like solving a crossword puzzle. You start by writing down the information which is easiest to figure out: where are the strident fricatives, where are the vowels and semivowels. That information will suggest things to look for next: are the fricatives voiced or unvoiced, how many vowels and semivowels are in each vocalic section. Gradually, in this way, you accumulate enough information to guess the phonemes.

This handout is organized around the following spectrogram reading steps. It's usually a good idea to start out by applying the steps in order, although you may find yourself backtracking sometimes.

1. Identify transitions between different broad phonetic classes. Broad phonetic classes which are usually easy to identify include vowel/semivowel, nasal consonant, stop consonant, strident fricative, and non-strident fricative.
2. Each broad class region may contain one or many segments. Look for clues to help you decide how many segments are in each region: how long is the region? Are there formant transitions?
3. If the phoneme is a stop or fricative, check to see whether it's voiced or unvoiced.
4. Measure the formant frequencies, and make 2-3 guesses about the identity of each segment.
5. Think about the effect of context on each phoneme, and add additional guesses if a segment is really ambiguous.
6. Try to find words which match the segments you wrote down.

Each of the sections in this handout is followed by spectrograms illustrating the main points of the section. For spectrograms of specific phonemes, Rabiner & Juang section 2.4 is a good reference.

2.9.1 Broad Class Transitions

Figure 2.12 is a spectrogram of the utterance "When the sunlight strikes raindrops in the air," as spoken by a male speaker.

Sonorant versus Obstruent

All speech sounds are either sonorant or obstruent. This is probably the easiest distinction to spot on the spectrogram, so it's a good one to look for first.

An obstruent sound is a sound produced with an obstructed air flow; obstruent sounds include stops, fricatives, and affricates. Since the air flow is obstructed, pressure builds up inside the vocal tract, and it becomes hard to maintain voicing. Thus, even voiced obstruents (e.g. /b/, /z/) usually have weak voicing.

A sonorant sound is one with unobstructed air flow; sonorant sounds include vowels, semivowels, and nasals.

Generally, if you can see F1 on the spectrogram, the sound is sonorant, and if F1 is missing or very weak, the sound is obstruent. For example, look at the n-d sequence in "raindrop" at around 1400ms in Fig. 2.12. F1 in the /n/ (1370-1440) is as strong as F1 in the neighboring vowel, but F1 in the /d/ release (1440-1480) is very weak.

Vowel/Semivowel

Vowels and semivowels are characterized by:

- Sonorant voicing.
- Clear formant structure.
- Continuous formant transitions.

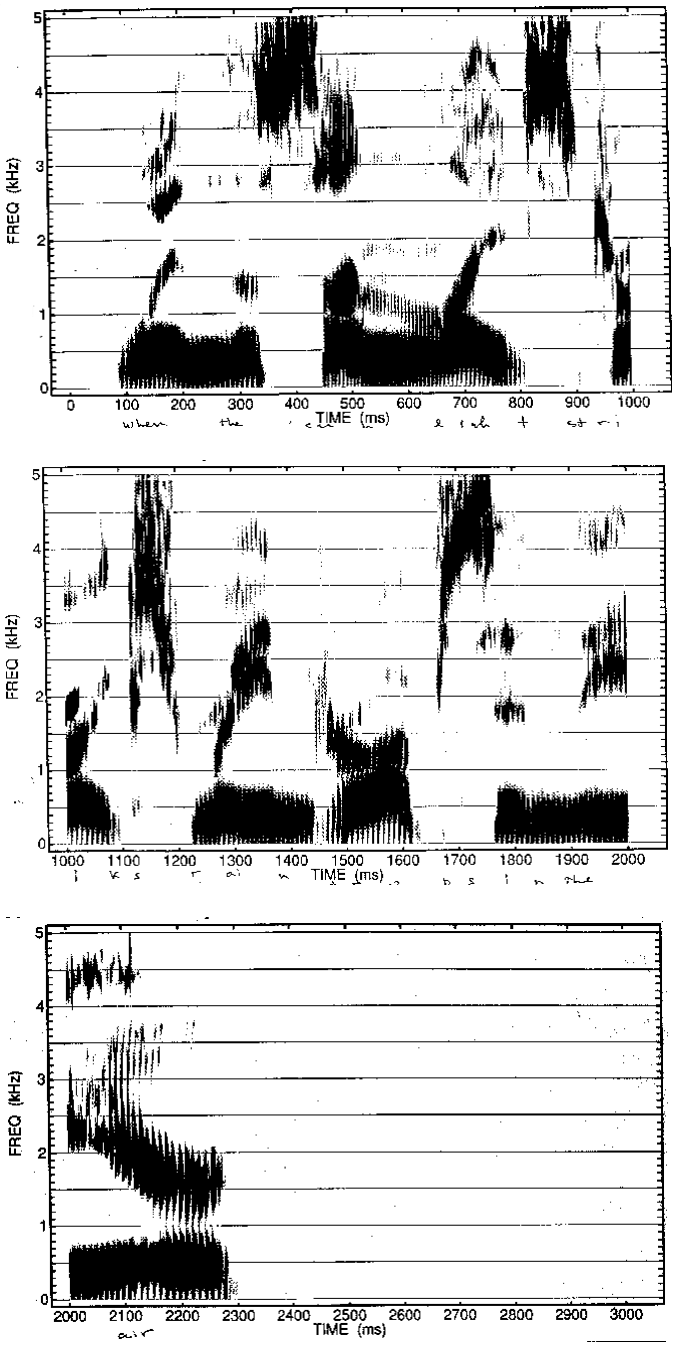


Figure 2.12: Spectrogram of the phrase, "When the sunlight strikes raindrops in the air..." A spectrogram is a plot of signal energy as a function of both time and frequency.

The first three formants are usually visible on the spectrogram. The formants move continuously from one phoneme to the next. For example, the /w/-/eh/ sequence from 100ms to 200ms in Fig. 2.12 is characterized by smooth upward formant transitions from the /w/ to the /eh/. In the /l/-/ai/ sequence between 650ms and 800ms, F1 and F3 fall while F2 and F4 rise, both with no discontinuities.

Nasal Consonant

Nasal consonants are characterized by:

- Sonorant voicing.
- Strong F1, but weak or invisible higher formants.
- Formants change discontinuously at CV and VC transitions.

The "formants" in a nasal consonant are resonances of the nose-throat system, not the mouth-throat system. The formant frequencies change discontinuously at the boundary between a vowel and a nasal consonant. For example, in Fig. 2.12, F1-F3 change discontinuously at the beginning of the /n/ in "sunlight" (520ms), and the amplitudes of the F2 and F3 peaks are much lower than during the preceding vowel. During the /n/ in "when the" (200-300ms), F2 and F3 are invisible, but F1 is still strong.

Stop Consonant

A stop consonant (also sometimes known as a plosive) is usually characterized by:

- Complete silence during closure, except possibly for a "voice bar" below 300Hz.
- A burst of broadband or high-frequency noise at release, which can last from 5ms to about 100ms.

Many stops can be located by simply looking for short silences (30-60ms). For example, the /t/ at 900ms and the /p/ at 1650ms in Fig. 2.12 are completely silent. The /t/ at 800ms and the /k/ at 1100ms are not completely silent: in both cases, the vocal folds continue to vibrate for a little while after glottal closure, resulting in a short "voice bar."

Not all silences are stops: notice the brief silence at 1200ms, between /s/ and /r/. This brief silence is clearly not a stop, because it is not followed by a burst of noise.

Strident Fricative

/s/, /z/, /ʃ/, and /ʒ/ are characterized by

- No sonorant voicing = no F1.
- High-amplitude frication noise, at high frequencies.

The spectrum of a strident fricative looks like a high-pass filtered noise spectrum, with a cutoff frequency between F3 and F5 of the neighboring vowel. The total energy of a strident fricative is similar to the energy of the neighboring vowel. Examples are visible in Fig. 2.12 at 350-450ms, 800-900ms, 1120-1200ms, and 1670-1760ms.

Non-Strident Fricative

/f/, /v/, /θ/, and /ð/ are characterized by

- Very weak, very broadband frication noise, often not visible on the spectrogram.
- Short duration.
- Easily assimilates characteristics of the neighboring sounds.

The word "the," which is supposed to begin with the non-strident fricative /ð/, appears twice in the sentence of Fig. 2.12. The first /ð/, at 290ms, seems to have assimilated the sonorant voicing of the preceding /n/, but there is still a small burst of apparent frication near 4400 Hertz, between 290ms and 300ms. The second /ð/, which should be near 1920ms, seems to have completely assimilated the sonorancy of the /n/.

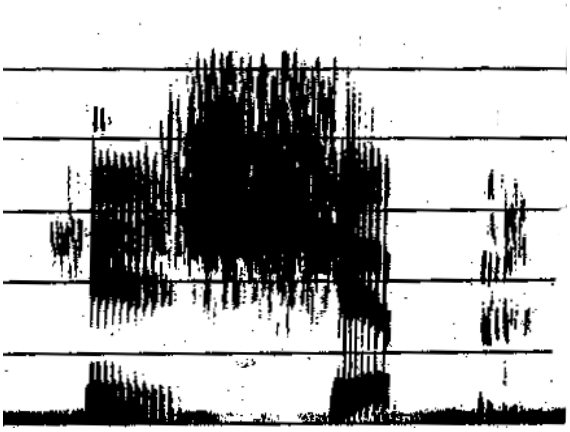


Figure 2.13: "His ship." The /z/ has become an /sh/ through processes of coarticuation, so that there are now two identical /sh/ phonemes in a row. The fricative noise section in the middle of the phrase looks just like a single /sh/; the only acoustic cue suggesting the presence of two fricatives is the duration, which is a bit more than 1.5 times the expected duration of a strident fricative in this context.

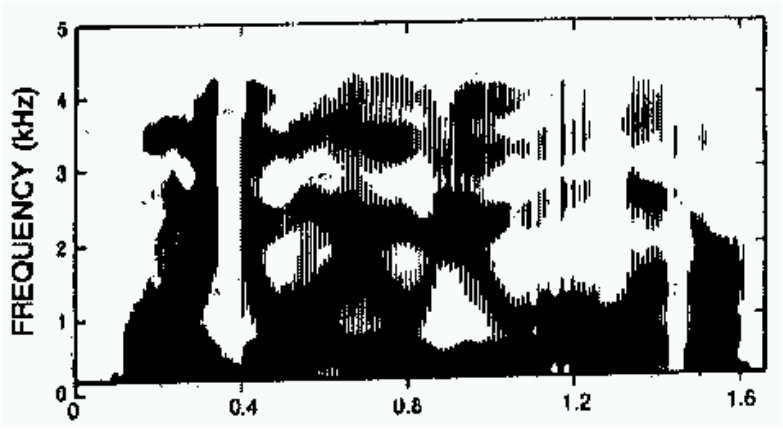


Figure 2.14: "Why do I owe you a letter?" Without knowing the phrase, we can guess the number of phonemes by counting the number of formant transitions, i.e. the number of significant up or down movements of either F1 or F2 or both.

2.9.2 How Many Segments?

Often, a "broad class" region will contain two or more consecutive phonemes. For example, consider the strident fricative region in "his ship," Figure 2.13. The /z/ in "his" has become an /∫/ through processes to be considered in section 2.9.5. As a result, there are two /∫/ phonemes in a row, with no visible boundary between them. The only reason that we know that there are two fricatives here instead of just one is that the fricative region is so long — roughly twice as long as either of the neighboring vowels.

A vowel/semivowel region can contain a long string of phonemes. Fig. 2.14, for example (R& J fig. 2.9), shows a spectrogram of the sentence "Why do I owe you a letter?" There are only two obstruents in this sentence (the /d/ at 0.35 seconds, and the /t/ at 1.45 seconds); all of the other phonemes are vowels or semivowels.

When a string of vowels/semivowels is this long, the only cue we can use to break it into phonemes is the movement of the formant frequencies. In general, each semivowel or regular vowel can be modeled as a single set of formant "targets," or frequencies which F1 and F2 should approach. In a long vowel/semivowel sequence, the formants move relatively smoothly and continuously from one set of formant targets to the next, so you can get a fair estimate of the number of phonemes by just counting the number of transitions.

The most important exceptions are diphthongs – segments like /ai/ and /oi/ – which, at first glance, look just like two vowels in a row (/a/+/i/ or /o/+/i/).

As an example, consider the formant transitions in Fig. 2.14. Making use of formant target from section 2.9.4, we can label the transitions from $t = 0.4$ to $t = 1.4$ as follows:

1. F1 and F2 approach 400 and 1100Hz, respectively, for the /u/ in "do."
2. F1 rises to 800Hz for the first target (the /a/ part) of /ai/ in "I."
3. F1 drops to 300Hz, and F2 rises to 1800Hz, executing the second target (the /i/ part) of /ai/ in "I."
4. F2 drops to 1100Hz for the /o/ in "owe."
5. F1 drops to 200Hz, and F2 rises to 1900Hz, for the /y/ in "you."
6. F2 drops to 1000Hz for the /u/ in "you."
7. F2 rises to 1200Hz, and F1 rises to 500Hz, for the /ah/ in "a."
8. F2 falls to 900Hz for the /l/ in "letter."
9. F2 rises to 1400Hz for the /eh/ in "letter."

2.9.3 Voiced or Unvoiced?

There are two useful cues which can be used to differentiate voiced and unvoiced obstruents:

- Voiced obstruents have longer voice bars. Note that *neither voiced nor unvoiced* obstruents are usually voiced all the way through, and *both voiced and unvoiced* obstruents are often voiced part way through, so you can't just rely on the presence/absence of a voice bar. What you can rely on is the length of the voice bar – the voice bar in an unvoiced obstruent usually stops within 30ms, while the voice bar in a voiced obstruent may last longer.
- *Total duration* of voiced obstruents is *shorter* (in English).

Fricatives

In voiced fricatives, the voice bar is actually a useful cue. figures 2.15 and 2.16 show spectrograms of a male speaker saying "the fuss," "the vanilla," "the thug," "the the," "the supper," "the zuh," "the shutter," and "the zhuh." All four of the voiced fricatives (/v/, /ð/, /z/, and /ʒ/) have voice bars all the way through the fricative. In more casual speech, these voice bars might be shortened to about half the length of the fricative.

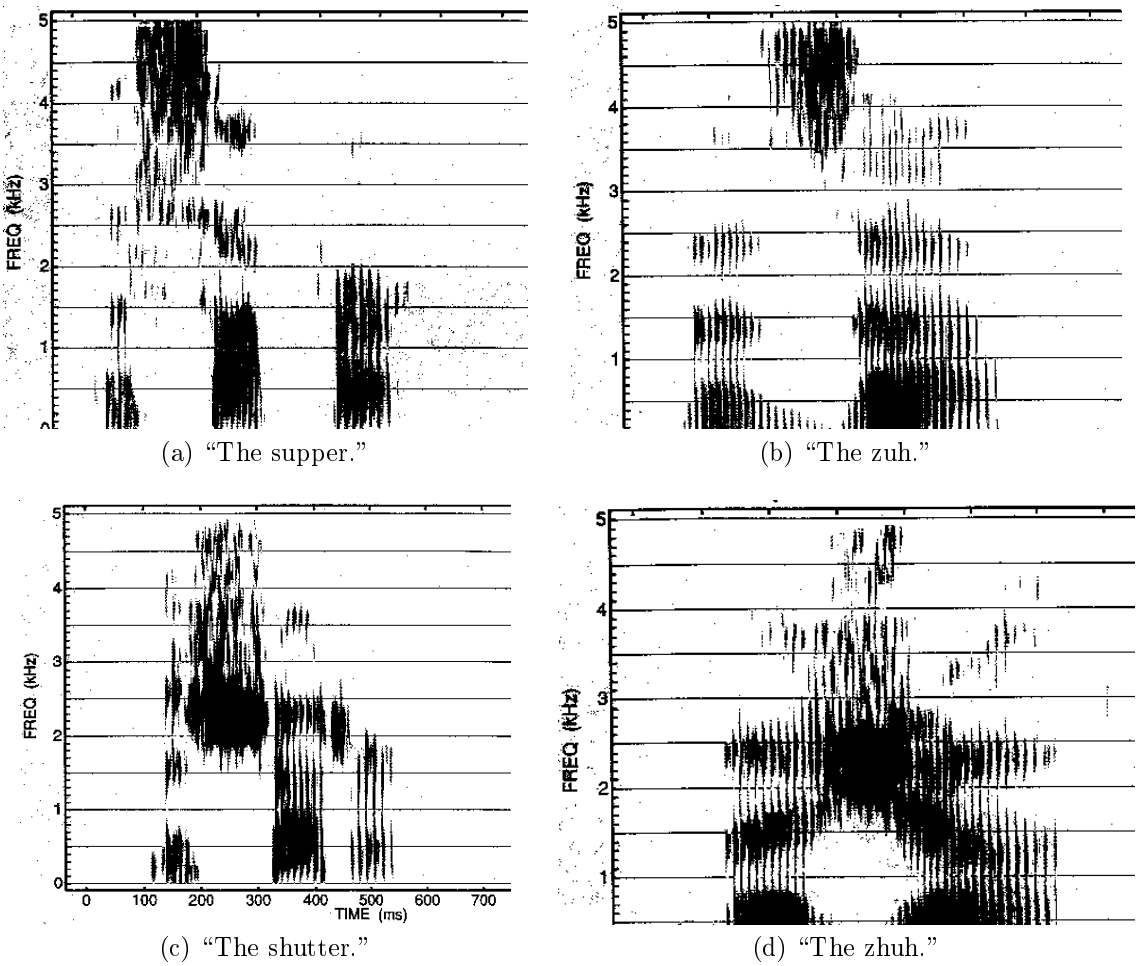
The total durations of the voiced fricatives in figures 2.15 and 2.16 are much shorter than the durations unvoiced fricatives, as shown in the following table:

	labial (/v/,/f/)	dental (/θ/,/ð/)	apical (/s/,/z/)	palatal (/ʃ/,/ʒ/)
Duration of Unvoiced Phoneme	130ms	120ms	130ms	140ms
Duration of Voiced Phoneme	50ms	70ms	120ms	80ms

Syllable-Initial Stops

The most reliable cue to voicing of a syllable-initial stop is the length of the noise burst following release. An unvoiced syllable-initial stop in English is "aspirated," which means that after noise stops being produced at the constriction, noise continues to be produced at the glottis for an additional 30-100ms. In the spectrogram, it looks as if the speaker were producing a stop followed by an /h/ sound.

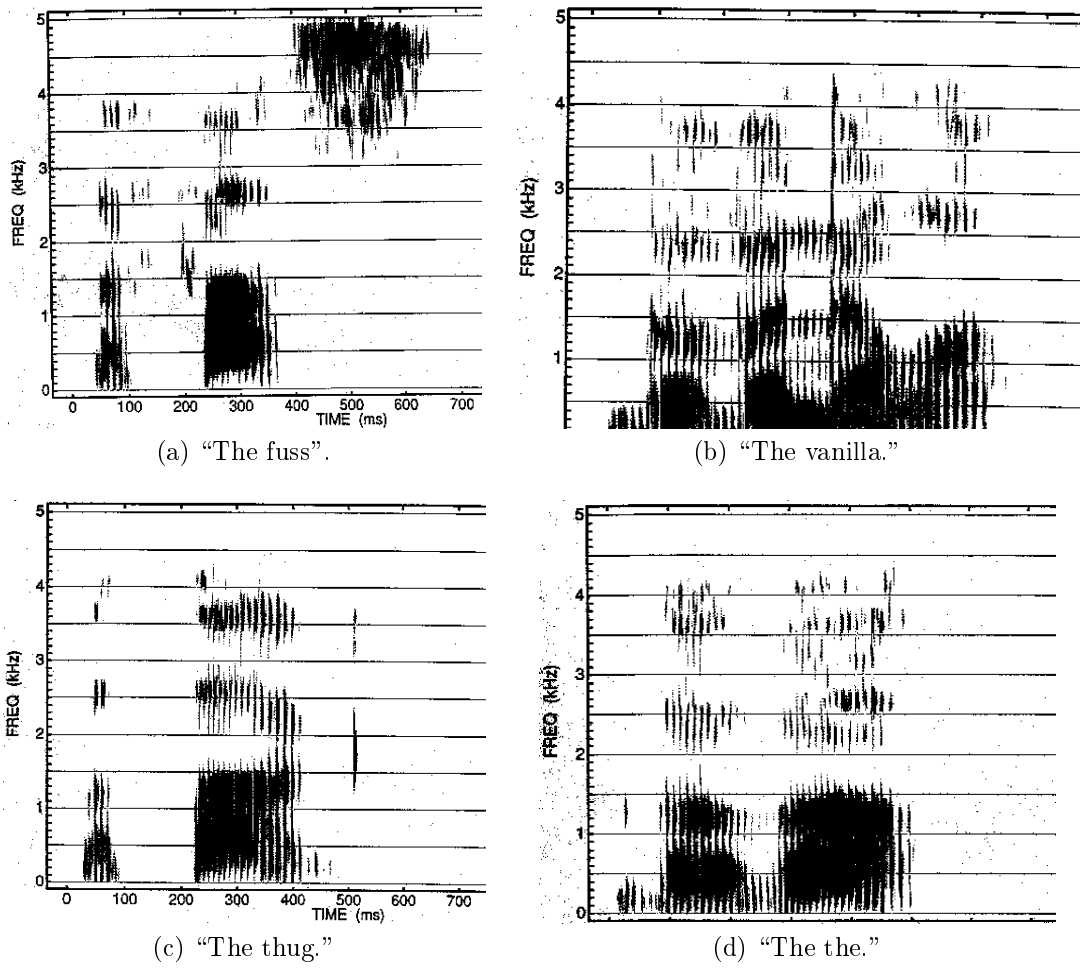
The six pages of Fig. 2.17 show spectrograms of a male speaker saying "the putt," "the bug," "the tug," "the duck," "the cut," and "the gut." The noise bursts following /d/ and /g/ are only about 25ms long,



(a) "The supper." (b) "The zuh."

(c) "The shutter." (d) "The zhuh."

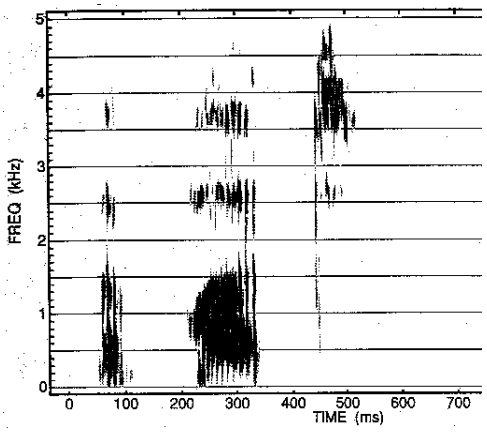
Figure 2.15: Strident fricative consonants, showing differences in place of articulation (e.g. "supper" versus "zuh") and voicing (e.g. "supper" versus "shutter.") "Strident fricatives" are formed by directing a turbulent jet of air against the teeth or some other obstacle, creating loud frication noise.



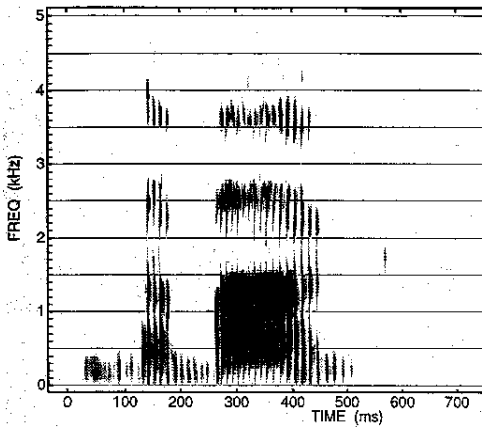
(a) "The fuss". (b) "The vanilla."

(c) "The thug." (d) "The the."

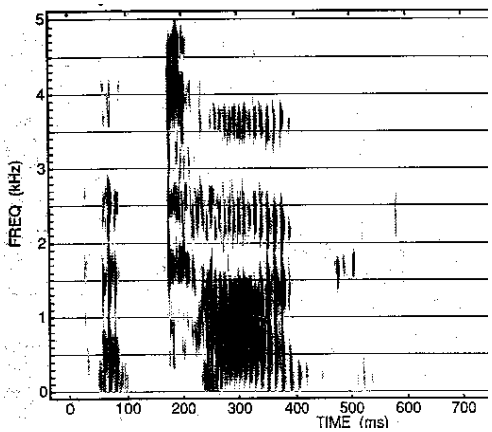
Figure 2.16: Non-strident fricative consonants, showing differences in place of articulation (e.g. "fuss" versus "thug") and voicing (e.g. "fuss" versus "vanilla.") Non-strident fricatives are similar to strident fricatives, but the turbulent jet of air usually does not hit an obstacle such as the teeth, so frication noise is not as loud.



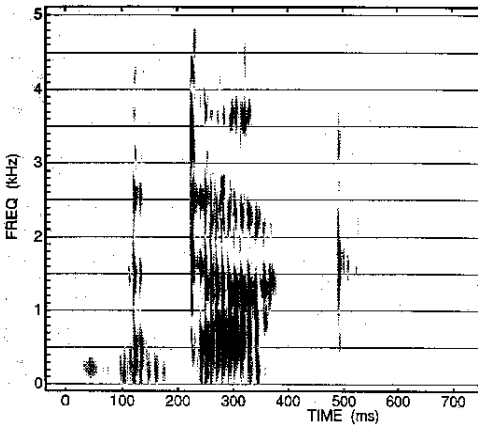
(a) "The putt."



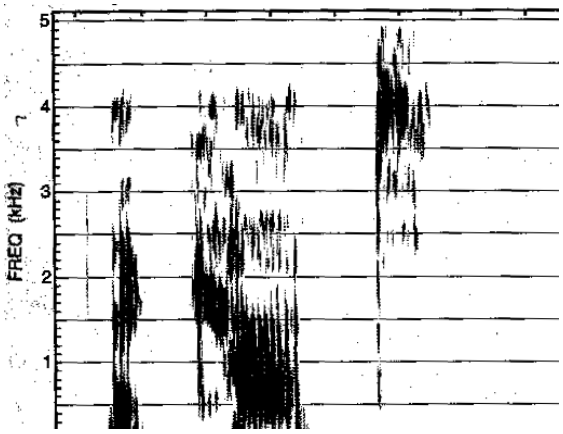
(b) "The bug."



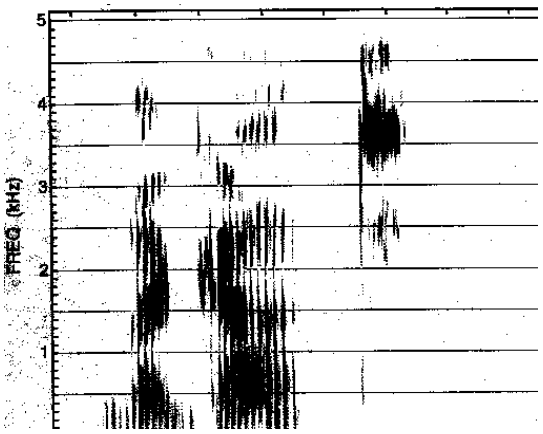
(c) "The tug."



(d) "The duck."

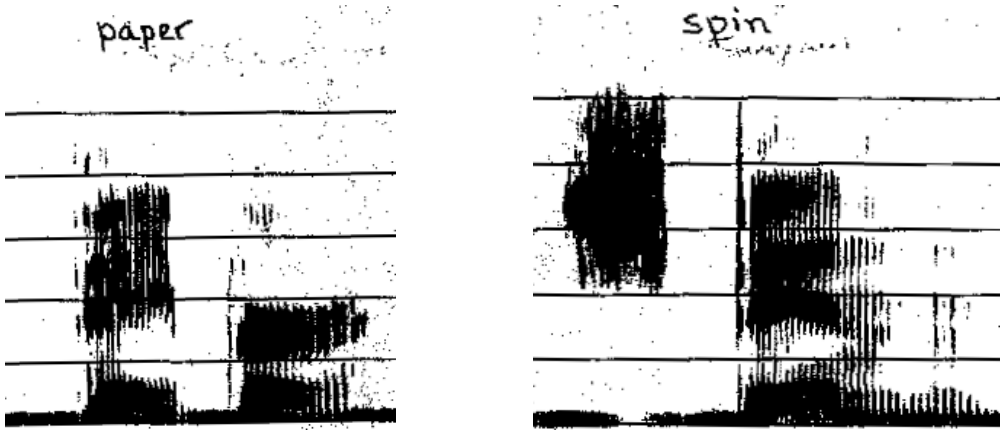


(e) "The cut."



(f) "The gut."

Figure 2.17: Stop consonants, showing differences in place of articulation (e.g. "putt" versus "tug") and voicing (e.g. "putt" versus "bug.")



(a) "paper" (b) "spin"

Figure 2.18: Syllable-initial unvoiced stops in English are always aspirated. These two examples show two different /p/ releases which are not strictly syllable-initial, and are therefore not aspirated. The middle /p/ in "paper" is both syllable-final and syllable-initial, so it has about half the normal amount of aspiration. The /p/ in spin is not syllable-initial at all, so it is unaspirated.

while the aspiration following /t/ and /k/ is about 60ms. Both /p/ and /b/ are shorter: /p/ only has 25-30ms of aspiration, while /b/ has no visible noise burst.

Notice that, unlike the fricatives, none of the stops has a voice bar which continues all the way through closure. The voiced stops tend to have longer voice bars than the unvoiced stops, but even this distinction may disappear in casual speech!

Non-Syllable-Initial Stops

An unvoiced stop is only aspirated in English if it is the first phoneme in the syllable. That means that unvoiced stops are not aspirated in the following positions:

- **Syllable-Final Stops**

  The best way to tell whether a syllable-final stop is voiced or unvoiced is by looking at duration — of both the stop itself, and of the vowel preceding the stop. Unvoiced stops tend to have a *longer* duration than voiced stops in English, and the vowel preceding an unvoiced stop is usually *shorter* than the vowel preceding a voiced stop. For example, look at the utterances "the tug" and "the duck" in Fig. 2.17. The /k/ in "duck" is not strongly aspirated — the noise burst after release only lasts about 30ms, which is short for a /k/ — but the closure duration is 120ms, which is 50% longer than the duration of the /g/ in "tug."

  Why is duration of the stops a cue for duration? Recent research suggests that the actual tongue or lip gesture is nearly the same regardless of whether the stop is voiced or unvoiced. The difference occurs at the glottis: in order to emphasize the lack of voicing in an unvoiced stop, English speakers start to cut off voicing (in this case, by spreading the glottis) before the tongue or lip constriction is fully formed. The result is that the vowel ends early before an unvoiced stop, and there is a period of silence which is longer than the actual duration of the stop constriction.

- **Ambisyllabic Stops**

  In some words, stops can be associated with both the preceding and the following syllables. Usually, these "ambisyllabic" stops act like syllable-final stops. For example, look at the word "paper" in Fig. 2.18a. The first /p/ is unquestionably syllable-initial, and shows 30ms of aspiration. The second /p/ is ambisyllabic, meaning it can be associated with either the first or second syllable, so it acts like a syllable-final stop: the noise burst is only 10ms long, but the closure is almost 100ms in length.

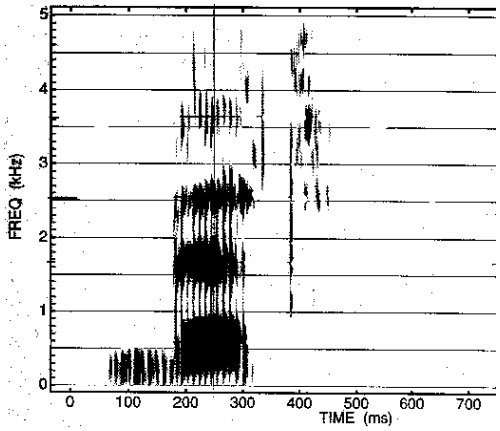


Figure 2.19: "Bet." In this example, the formant frequencies are relatively constant throughout the vowel.

(image missing)

Figure 2.20: Formant frequencies of the vowels of English, as measured from a large database of male speakers [17].

- **Consonant Clusters**

  If an unvoiced stop is preceded by /s/, the stop is not syllable initial, and it is therefore not aspirated! Observe, for example, the /p/ in "spin," in Fig. 2.18b: the noise burst is only about 15ms in length.

  Fortunately, stops in /s/-stop clusters in English are always unvoiced: "ska," "stay," and "spa" are possible English words, but "sga," "sday," and "sba" are not.

2.9.4 Measure the Formant Frequencies

Once you have found a phoneme in the spectrogram, identified its broad class, and figured out whether it's voiced or unvoiced, the only thing left to do is identify it. In phonetics, this last little bit of information about a phoneme is sometimes called the "place" of the phoneme, and it corresponds roughly to the position of the tongue and lips while the phoneme is being produced.

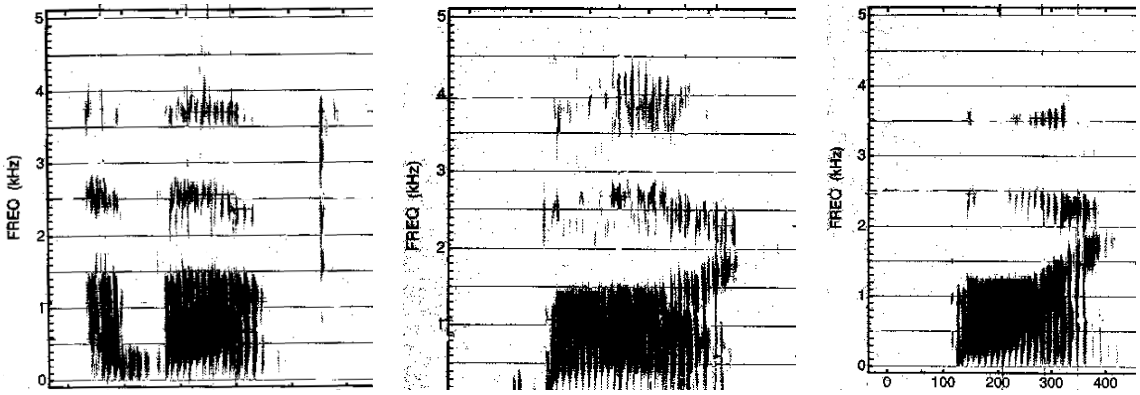
In all cases, we determine the "place" of a phoneme by looking at the frequencies of the formant resonances. However, the way we *use* the formant frequencies to identify a phoneme varies somewhat from one class to another.

Vowels

The identity of a regular vowel or a semivowel is determined entirely by the values of the formant targets, so the first step is to identify the targets. A single vowel surrounded by obstruents may show constant formants, like the word "bet" in Fig. 2.19 (male speaker). A sequence of vowels and semivowels, however, will often look like a series of formant transitions, as shown in Fig. 2.14, with no steady-state formants. The formant targets for each phoneme will usually be close to a local minimum or local maximum in either F1 or F2.

Once you have identified the formant targets, the best way to try to identify a vowel or semivowel is usually to just use F1 and F2 as indices into a chart like that shown in Fig. 2.20, which shows the vowels of American English arranged in an F1-F2 vowel space for male speakers [17]. Similar charts exist for female speakers; if you don't have a chart for female speakers available, you can get pretty good results by multiplying a female speaker's formants by a scaling factor of 6/7 (the average ratio of female to male vocal tract lengths), and then looking up the vowel in Fig. 2.20.

Unfortunately, in continuous speech, the formant targets of a vowel are often "missed" by a few hundred Hertz. In speech recognition systems, this variability is modeled as "noise" in the speech production system.



(a) "about" (b) "buy" (c) "boy"

Figure 2.21: The words "about," "buy," and "boy," showing the three dipthongs of English: /AW/, /AI/, and /OI/, respectively (ARPABET notation). Each diphthong starts with a relatively static set of formant targets, then moves toward a briefly marked second set of formant targets.

Diphthongs

Diphthongs, like /aɪ/ in "buy," /ɔɪ/ in "oy," and /aʊ/ in "about," are sounds with two distinct formant targets. Fig. 2.21 shows spectrograms of a male speaker saying "about," "buy," and "boy." Notice that:

- Diphthongs are long. A diphthong is usually longer than 100ms.
- There are two distinct formant targets. The formants often sit close to the first target for about 50ms, and then slide off toward the second target.

Semivowels

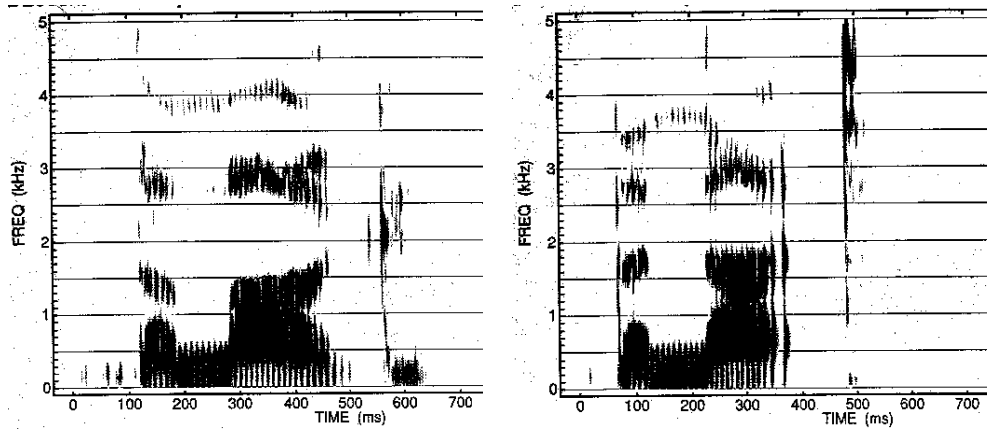
The semivowels are similar to vowels, but with more extreme formant values. In addition, F3 is often useful in identifying semivowels.

- /j/ is like /i/, but with an even higher F2 (usually above 2000 Hertz), and an even lower F1 (usually below 300Hz). F3 is also high.
- /w/ is like /u/, but with even lower formants – F1 is below 300Hz, F2 is below 800Hz, and F3 is also low.
- /r/ is one of the easiest phonemes to find in any spectrogram — whenever F3 dips below 2000 Hertz, there is an /r/ there! Male speakers always produce /r/ with F3 below 2000 Hertz; some female speakers may produce /r/ with F3 as high as 2200 Hertz.
- /l/ is also easy to find. In an /l/, F1 is neutral (near 500Hz), but F2 dips down very close to F1 (600-900Hz), while F3 goes very high (2900-3200Hz). The result is a large energy gap in the mid-frequency range where F2 and F3 normally are.

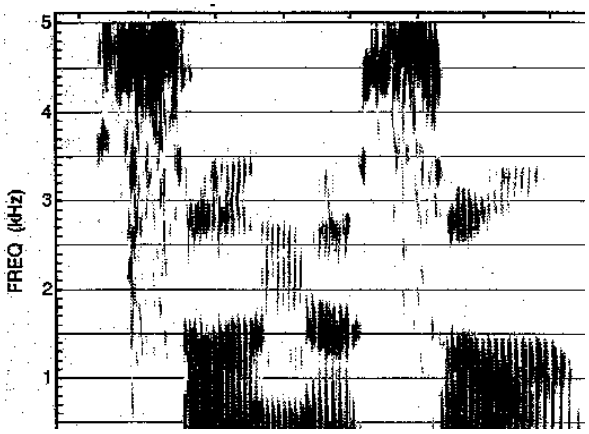
/r/ and /l/ can be the nucleus of a syllable, as in the words "bert" and "able." When this happens, the phonemes are sometimes transcribed as /ɜ˞/ and /l̩/.

Nasals and Stops

Nasal and stop consonants can be identified based on the movement of formants in the neighboring vowels. The formants of the preceding vowel tend to move toward a *locus* which characterizes the place of articulation of the stop or nasal, and the formants of the following vowel tend to start near the same *locus* and move away from it.



(a) "The mug." (b) "The nut."



(c) "Sung a song."

Figure 2.22: Nasal consonants. The nasal consonants in English have the same three places of articulation (lips, tongue blade, and tongue body) as the stop consonants.

For this section, consider the stop consonants shown in Fig. 2.17, and the nasal consonants shown in Fig. 2.22. Fig. 2.22 shows spectrograms of the utterances "the mug," "the nut," and "sung a song" as produced by a male speaker.

- **Labial (/p/,/b/,/m/)**

  As the lips close, the resonances of the vocal tract move toward the resonances of a tube closed at both ends:

$$F_n = \frac{nc}{2l}, \quad n = 0, 1, 2, \ldots \tag{2.140}$$

  where $c \approx 35000$cm/s is the speed of sound in air at body temperature. By assuming that $l \approx 17.5$cm, we can calculate that the formant loci of a labial stop are approximately $F_1 = 0$, $F_2 = 1000$, and $F_3 = 2000$ Hertz. The actual first locus of a labial stop is about $F_1 = 180$Hz, because of the yielding walls of the vocal tract, but the F_2 and F_3 loci are correct.

- **Velar (/k/,/g/,/ŋ/)**

  In a velar stop, the tongue body cuts the vocal tract neatly into thirds – 2/3 of the vocal tract sits behind the constriction, and 1/3 sits in front of the constriction. The formant loci are a combination of the resonances of the back cavity and the front cavity:

$$f_{b,n} = 0, \frac{c}{2l_b}, \ldots \tag{2.141}$$

$$f_{f,n} = \frac{c}{4l_f}, \frac{3c}{4l_f}, \ldots \tag{2.142}$$

  As with labial stops, the yielding walls of the vocal tract raise the first formant locus from zero up to about $F_1 = 200$Hz.

  The loci of F_2 and F_3 are the same! Since $l_b = 2l_f$,

$$\frac{c}{2l_b} = \frac{c}{4l_f} \tag{2.143}$$

  In fact, the position of a velar constriction moves around quite a bit, so the actual locus of F_2 and F_3 may be anywhere between about 1500Hz and 2200Hz. Strangely enough, no matter what the F_2 and F_3 locus may be, they are always close together! This convergence of F_2 and F_3 is often easy to spot on a spectrogram, and is such a useful cue that it has been given a name — we call it the "velar pinch."

  In Fig. 2.17, the velar pinch is at about 2000Hz in "the cut" and "the gut," and at about 1700Hz in the syllable-final /k/ and /g/ in "the tug" and "the duck." In Fig. 2.22, the velar pinch in "sing" is at about 2200Hz. The word "song" in Fig. 2.22 demonstrates the only common exception to the "velar pinch" rule — a velar stop or nasal following an /a/ or /l/ may show a dropping F_2, as in this spectrogram.

- **Alveolar (/t/,/d/,/n/)**

  Unlike labial and velar stops, the formant loci of an alveolar stop are not easy to calculate using a production model. In part, this is because speakers seem to adjust the shapes of their mouths so that, regardless of context, the F_2 locus of an alveolar stop is always the same! With remarkable consistency, the F_2 locus of an alveolar stop is almost always

$$F_2 = \begin{cases} 1700\text{Hz} & \text{for male speakers} \\ 2000\text{Hz} & \text{for female speakers} \end{cases} \tag{2.144}$$

  A velar stop may have a similar F_2 locus, but unlike the velar, an alveolar will never show convergence of F_2 and F_3. The F_3 locus of an alveolar stop tends to be moderately high — typically between 2500 and 3000 Hertz.

  The F_1 locus of an alveolar stop is usually the same as the locus of a labial stop, that is, $F_1 = 180$Hz.

Strident Fricatives

The spectrum of a strident fricative typically looks like high-pass filtered white noise. The place of the fricative is determined by the cutoff frequency of the noise spectrum:

- **Palatal (/ʒ/,/ʃ/)**

  The cutoff frequency is usually between F2 and F3 of the neighboring vowel. See, for example, "the shutter" and "the zhuh" in Fig. 2.15.

  As shown in Fig. 2.15, /ʃ/ and /ʒ/ are the only non-velar sounds which sometimes show a "velar pinch" of F2 and F3 in the neighboring vowel. Usually we don't need this cue, because the cutoff frequency is such an easy cue to use.

- **Alveolar (/z/,/s/)**

  The cutoff frequency is usually between F4 and F5 of the neighboring vowel (typically around 4000Hz). See, for example, "the supper" and "the zuh" in Fig. 2.15.

  The cutoff frequency of an alveolar fricative will sometimes sweep downward toward a neighboring labial sound, such as /p/, /w/, or even /r/. This sweep is called a "labial tail," and it is usually transient; most of the fricative will show the typical 4500Hz cutoff frequency of an /s/. For example, the /s/ at 1200ms in Fig. 2.12 shows a labial tail toward the following /r/ ("strikes raindrops"), and the /s/ at 1670ms (at the end of "raindrops") begins with a labial tail caused by the preceding /p/. Notice that, in both cases, the cutoff frequency in the center of the fricative is much too high for an /ʃ/, so we can easily identify these fricatives as /s/ despite the labial tails.

Non-Strident Fricatives

Non-strident fricatives are difficult to identify as fricatives, because they tend to assimilate voicing (and sometimes even sonorancy) from the neighboring sounds. In order to compensate for their tendency to lose general class features, however, non-strident fricatives cling to their place features very stubbornly. For example, most English speakers pronounce the /ð/ in "in the" (Fig. 2.12) as if it were a nasal consonant, but the tongue is still pushed up against the teeth as if it were a /ð/.

The most reliable cues for place of a non-strident fricative, therefore, are usually the formant loci.

- **Dental (/ð/,/θ/)**

  The F2 locus of a dental consonant is lower than that of an alveolar consonant (perhaps in part because the back cavity is longer). The F2 locus of /θ/ and /ð/ is typically

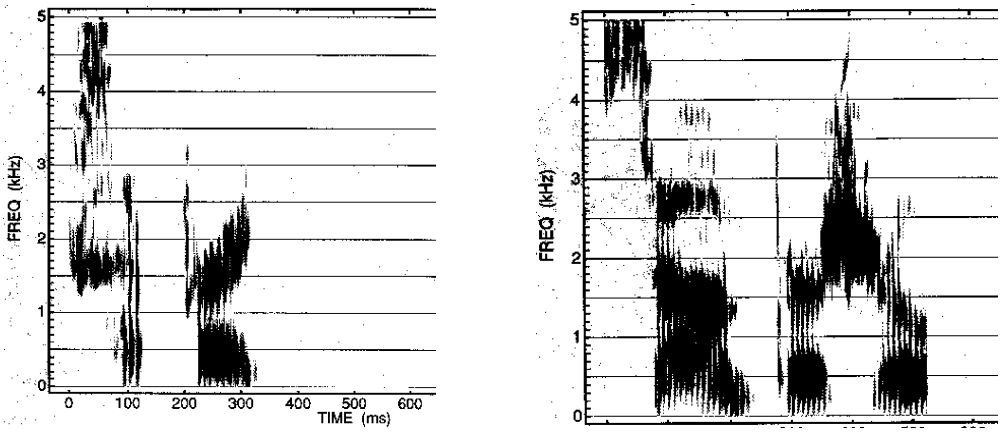
  $$F_2 = \begin{cases} 1400\text{Hz} & \text{for male speakers} \\ 1600\text{Hz} & \text{for female speakers} \end{cases} \tag{2.145}$$

  If a dental fricative shows any frication at all, the frication has a very high cutoff frequency, typically at 4500Hz or higher.

- **Labiodental (/v/,/f/)**

  The sounds /v/ and /f/ are made by putting the lips against the teeth, so they are called "labiodental" sounds ("labio"=lips, "dental"=teeth). The formant loci are essentially the same as for a labial sound: 180Hz, 1000Hz, and 2000Hz.

  Of all of the non-strident fricatives, /f/ is the most likely to actually include frication noise which is strong enough to see on a spectrogram — see, for example, "the fuss" in Fig. 2.16. Labiodental frication tends to be scattered over several frequencies, and sometimes shows a weak harmonic structure, as in the example of Fig. 2.16 (frication noise at 1800, 2700, and 3700Hz). There is no definite cutoff frequency, but you usually do not see frication noise below 1000Hz.



(a) "ksep fer (except for)" (b) "soum pressure (sound pressure)

Figure 2.23: Examples of phoneme deletion and feature assimilation caused by context effects.

2.9.5 Context Effects

The sound of a phone may change drastically in casual speech. These changes can usually be predicted as a function of the surrounding phone-level and word-level context. Several examples of contextual change have already been described; this section briefly presents two more examples.

Devoicing of Unstressed Syllables

Unstressed vowels can be devoiced — the vowel is still "there," in some sense, but there is no sonorant voicing to mark its place. In Fig. 2.23(a), for example, "except for" has been implemented as "ksep fer." Native speakers of English still feel that this sequence contains three syllables, in part because the sequence "ksep" is not an allowable syllable in English. For practical speech recognition purposes, however, the sequence should be considered a two-syllable reduction of the original phrase. For example, it is possible to show that the duration of this "ksep fer" is slightly longer than the duration of phonetically similar two-syllable phrases (e.g. "skip for" or "clip for"), but only by carefully comparing phrases spoken by the same speaker in identical sentence contexts. [10].

This example also illustrates two other common processes: reduction of unstressed vowels ("for" becomes /fər/), and deletion of stops in the middle of a multi-obstruent sequence (the /t/ in "except" is dropped).

Place Assimilation

The first consonant in a stop-stop or nasal-stop sequence often adopts the place of articulation of the second consonant. See for example Fig. 2.23(b), "sound pressure" becomes "soum pressure."

2.10 Exercises

1. Use matlab to analyze one of the sentences in ~ee214a. Listen to the sentence – what is the speaker saying? In each of the following parts, be sure to tell me the time of the segment you are analyzing, so I can check your work!

   (a) Find a vowel. What vowel is it? What are the first three formants? What is F_0?

   (b) Find a semivowel. What semivowel is it? What are the first three formants? What is F_0?

   (c) Find a nasal consonant. What consonant is it? How many resonance frequencies can you identify, and what are they? What is F_0?

   (d) Find a fricative consonant. What consonant is it? What is the frequency of the most prominent resonance?

   (e) Find a plosive release. What consonant is it? What are the locus frequencies of the first three formants in the following vowel?