

Acoustic phonetics

John Kingston

Linguistics Department  
University of Massachusetts  
150 Hicks Way, 226 South College  
Amherst, MA 01003-9274  
[jkingston@linguist.umass.edu](mailto:jkingston@linguist.umass.edu)  
1-413-545-6833, fax -2792

Like the rest of linguistics, acoustic phonetics consists of both description and theory. Descriptions are images of some acoustic property of speech, or quantitative measures taken from these images, while the theory is concerned with how a sound's articulation determines its acoustics.

### Description

The three most commonly used images of speech are the *waveform*, *spectrum*, and *spectrogram* (Figure 1). The waveform displays differences in sound pressure level over time (Figure 1a,d,g,j), the spectrum differences in sound pressure level over frequency (Figure 1b,e,h,k), and the spectrogram differences in frequency over time (Figure 1c,f,i,l; darkness indicates differences in sound pressure level at particular frequencies and moments in time in the spectrogram). In the waveform, sound pressure level is measured in units of pressure, kiloPascals (kPa), and in the spectrum it is measured in deciBels (dB), a logarithmic transform of the ratio between the target sound's pressure level and a reference pressure level:

$$(1) \quad dB = 20 \log_{10} \frac{P_{target}}{P_{reference}}$$

Any difference in pressure by a factor of 2 is equivalent to a difference of 6 dB.

Figure 1a-c displays the waveform, spectrum, and spectrogram of a pure tone. The waveform shows that this sound is *periodic*, i.e. sound pressure level repeats itself regularly, here every 0.0022727 sec. The duration of a single *cycle* is the wave's *period* ( $T$ ). If the horizontal axis were not time elapsed but distance traveled, then the extent of a single cycle is its *wavelength* ( $\lambda$ ). Because both the period and wavelength measure a single cycle's extent, they are reciprocally related to the wave's frequency ( $F$ ), the number of cycles per second (Hz are equivalent to cycles per second):

$$(2) \quad F \text{ (cycles / sec)} = \frac{1}{T \text{ (sec / cycle)}}$$

$$(3) \quad F \text{ (cycles / sec)} = \frac{c \text{ (cm / sec)}}{\lambda \text{ (cm / cycle)}}$$

The numerator in (3) is not 1 but instead  $c$ , the speed of sound. A pure tone has energy at just one frequency; the single peak in the spectrum and the single line in the spectrogram shows that this pure tone has energy at 440 Hz. Pure tones are also called *sine waves* because the moment-to-moment variation in sound pressure level has the shape of the sine function [the pure tone displayed in Figure 1a was generated from  $0.5 \text{ kPa} \cdot \sin(2\pi \cdot 440 \cdot t)$ , where  $t$  represents time].

Please put Figure 1 about here.

The sound displayed in Figures 1d-f differs in all these respects: sound pressure level varies randomly not regularly (Figure 1d). This sound is thus *aperiodic*. The sound has more or less equal amounts of energy at all frequencies (Figures 1e,f). On analogy

with white light, which has equal amounts of energy at all visible, this sound is called *white noise*, because it has equal amounts of energy at all acoustic frequencies.

The sounds displayed in Figures 1g-i and 1j-l are speech sounds that resemble these non-speech sounds. In Figure 1g-i is a brief portion of the vowel [i] (as in *heed*), in which sound pressure level varies regularly even if not as strictly as that of the pure tone, while in Figure 1j-l is a brief portion of the fricative [s] (as in *see*), in which sound pressure level varies nearly as irregularly as in the white noise. The vowel is (quasi)periodic, and the fricative aperiodic. The spectrum and spectrogram of [i] shows that it has energy at more than a single frequency; the peaks in its spectrum (Figure 1h) and the horizontal bands in its spectrogram (Figure 1i) show that it has considerable energy below about 500 Hz, and at roughly 2000, 2600, and 3500 Hz, while the corresponding images of [s] (Figures 1k,l) show that it has considerable energy spanning a broad band from 4000-7000 Hz, but much less below that frequency.

Whether a sound is periodic and where in its spectrum energy is concentrated are nearly sufficient to distinguish all speech sounds from one another acoustically, and they also reflect the two components of the theoretical model for transforming articulations into acoustics. All voiced sounds are periodic, as are the trills, and besides the fricatives, stops and affricates are also aperiodic. Differences in quality between vowels and differences in place of articulation between consonants are both realized acoustically as differences in where energy is concentrated in their spectra.

The remaining property is duration, i.e. how long an acoustic property lasts. Besides distinguishing long and short vowels and consonants from one another, differences in duration contribute to conveying voicing contrasts in stops (via differences in when voicing begins relative to the stop release in prevocalic stops and differences in vowel duration before post-vocalic stops) and manner contrasts: noise is produced in a brief burst when a stop closure is released but sustained throughout a fricative constriction, noise is sustained longer in fricatives than affricates, and the distribution of energy across the spectrum changes frequency faster in stops than glides.

## Theory

The speech sounds radiating from the lips are the product of applying a filter to a sound source.

## Sound sources

Speech is produced by transforming the energy embodied in the flow of air into sound, thereby producing a sound *source*. All sound sources are produced by using valves to manipulate air flow through the vocal tract.

Periodic sound sources are produced when the flow of air causes a valve to open and close rapidly and regularly, or to *vibrate*. As the valve opens and closes, air pressure rises and falls regularly just downstream. The most common periodic sound source is vocal fold vibration, which is a rapid, regular opening and closing of the glottis, but trills made with the back edge of the soft palate, the tongue tip, or lips also produce such sound sources.

Aperiodic sound sources are produced by closing a valve enough to obstruct air flow through it. The valve can be closed completely, as in a stop, or partially, as in a fricative. If air continues to flow into the cavity through the glottis from the lungs, intraoral air pressure ( $P_o$ ) rises behind the obstruction. This rise in  $P_o$  causes oral air flow ( $U_o$ ) to speed up enough to become turbulent and thus noisy. Noisy, turbulent air flow is produced continuously through the narrow channel of a fricative constriction, but it is limited to a brief burst following the *release* of the complete obstruction in stops.

The essential and most useful characteristic of both the periodic and aperiodic sound sources is that they introduce acoustic energy into the oral cavity across a very broad range of frequencies, although energy falls off with increasing frequency in voicing, and there is relatively little energy below about 1000 Hz in aperiodic sound sources.

In voicing, there is energy at the rate of vocal fold vibration, also known as the *fundamental frequency* ( $F_0$ ), and at all its integer multiples, also known as *harmonics*. Energy is present at the harmonics of the fundamental because the shape of the sound pressure level variation above the glottis is saw-toothed rather than sinusoidal (Figure 2). It has this shape because air flow through the glottis ( $U_g$ ) drops off more abruptly than it rises. As a consequence, the pressure level of the sound produced rises gradually but falls abruptly. Voice qualities differ in how abruptly  $U_g$  drops off and thus how steeply pressure falls: compared to modal voice (Figure 2a),  $U_g$  drops off much more abruptly than it rises in tenser and creakier voice qualities (Figure 2b), while it may drop off no more abruptly than it rises in laxer and breathier voice qualities (Figure 2c). The faster  $U_g$  drops off, the more energy is present in the higher harmonics: energy falls off at 6 dB per octave in modal voice, just 3 dB per octave in tense voice, and at 12 dB per octave in lax voice; with each doubling in frequency, energy is reduced to 50%, 71%, and 25%, respectively, of its previous value (Figure 3).

Please put Figures 2 and 3 about here.

## Resonance

The broadband distribution of energy in both periodic and aperiodic sound sources is useful because it provides energy at all the frequencies at which the air inside the oral cavity resonates. Those frequencies depend on the shape of the resonating cavity. If the articulators are in their rest positions and vocal folds are in the position for voicing, the shape of this cavity approximates that of a tube with the same cross-sectional area along its entire length, that is closed at one end, the glottis, and open at the other, the lips (Figure 4). The acoustic energy in the source propagates away from the source and reflects back and forth off the two ends of the tube, thereby establishing a resonance in the form of a *standing wave*. In a standing wave, the locations of zero and maximum pressure variation are fixed.

Please put Figure 4 about here.

Up to now, waveforms have represented sound waves as pressure levels varying across time or space. To understand how air resonates in a tube, it is more useful to represent them as *changes* in pressure level distributed across time or space. The pressure level represents the extent to which the air molecules are compressed or rarefied at particular moment or location, while the change in pressure level instead

represents the extent to which the air molecules are displayed longitudinally, or equivalently the *velocity* of the pressure change.

The air inside the oral cavity is most free to undergo longitudinal displacement at the open end, the lips, and least free to do so at the closed end, the glottis. In other words, a velocity maximum is expected at the lips, and a velocity minimum at the glottis. (Minima are also called *nodes*, and maxima *anti-nodes*.) The open and closed ends of the tube are thus *boundary conditions* that determine the frequencies at which the air inside the oral cavity will resonate; they are the frequencies of waves that have a velocity maximum at the lips and a velocity minimum at the glottis. The air resonates less robustly at the frequencies of waves whose velocity maximum and minimum are at greater distances from the lips and glottis. Because the air resonates at some frequencies and not at others, it *filters* the sound source, passing energy in the source at frequencies close to the air's preferred resonance frequencies and stopping it at more distant frequencies.

Figures 4b-d show the first three sound waves that fit these boundary conditions, in that each has a velocity maximum at the lips and a velocity minimum at the glottis. They are referred to as the first, second, and third resonances, because they have the three lowest frequencies of all waves that fit boundary conditions. It does not matter that the second and third waves have velocity maxima and minima elsewhere along the tube, as only the two ends impose boundary conditions. Although an infinite number of additional waves fit these boundary conditions, the first three are sufficient for distinguishing nearly all speech sounds from one another.

What then are these waves' frequencies? This question can be answered using just three facts: (1) frequency ( $F$ ) is the reciprocal of wavelength ( $\lambda$ ), (2) the proportion of the sound's wavelength that spans the distance from the glottis to the lips, and (3) the distance from the glottis to the lips itself, the oral cavity's length ( $L_{oc}$ ). Figures 4a-c show that one quarter of the first resonance's wavelength spans the distance from the glottis to the lips, three quarters of the second's, and five quarters of the third's. More generally:

$$(4) \quad L_{oc} = \frac{2n - 1}{4} \lambda_n$$

where  $n$  is the resonance number. If we solve for wavelength and substitute the result into (3), we obtain:

$$(5) \quad F_n = \frac{c}{\frac{2n - 1}{4} L_{oc}}$$

Substituting 35000 cm/sec for  $c$  and 17.5 cm for  $L_{oc}$  (the average length of the adult male oral cavity) into (5) yields 500, 1500, and 2500 Hz as the frequencies of the first three resonances. These values are very close to those for the neutral vowel, schwa.

The only variable in (5) is  $L_{oc}$ . Because it is in the denominator, resonance frequencies are all higher when  $L_{oc}$  is shorter and lower when it is longer. This explains why resonance frequencies are generally higher in children than adults and in women than in men. It also explains why resonance frequencies are higher in vowels

pronounced with the lips spread, such as the [i] in *heed*, than in those pronounced with the lips rounded, such as the [u] in *who'd*. Spreading the lips shortens the oral cavity, while the lip protrusion that is part of lip rounding lengthens it. From these observations, we formulate the *length rule*: the frequencies of all resonances vary inversely with the length of the resonating cavity.

The spectrum of the vowel [i] in Figure 1h has peaks at frequencies decidedly different from those predicted for schwa. The lowest peak is well below 500 Hz, the next well above, although the third is close to 2500 Hz.

Before continuing, a new term is needed to refer to these peaks. Because the resonant response of the air in the oral cavity is a broadly rather than narrowly tuned filter, the air resonates across a range of frequencies and amplifies a number of contiguous harmonics rather than just a single one (Figure 1h; also Figure 5). These ranges are called *formants*, and henceforth these peaks are referred to as the first, second, and third formants (F1, F2, F3).

The formant frequencies of [i] undoubtedly differ from those predicted for schwa because [i]'s articulation differs from schwa's. [i] is produced by raising the tongue body toward the front of the palate, while spreading the lips at the same time. These articulations change the cross-sectional area of the oral cavity, as well as its length. We have already formulated a rule for predicting how shortening changes formant frequencies, but we still need one for predicting what changes in cross-sectional area do.

Two heuristics are widely used for this purpose. These heuristics are tested here by seeing how well they predict the differences in formant frequencies between the three point vowels [i, u, a], and between the three places of articulation in the syllables [ba, da, ga]. The spectra in Figure 5 show that that F1 is lower in the high vowels [i, u] (Figures 5a, b) than in schwa, but higher in the low vowel [a] (Figure 5c); they also show that F2 is higher in front unrounded [i] than schwa but lower in low back unrounded [a] and especially high back rounded [u]. F1 varies inversely with tongue height, and F2 varies directly with tongue backness and lip rounding. The spectrograms in Figure 6 show that F1 starts low and rises following [b, d, g], both F2 and F3 start low following [b], both formants start higher following [d], and they diverge from very similar frequencies following [g].

Please put Figure 5 and 6 about here.

The first heuristic treats the constriction as dividing the oral cavity into separate resonating cavities (Figure 7) and applies the length rule independently to each of them. The first three formants radiating from the lips are the three lowest of all the resonances produced by the two cavities. It is called the *cavity association* heuristic because each formant can be associated with the cavity from which it came. There are two complications. First, the cavity behind the constriction is effectively closed at both ends, so its resonances must have velocity minima at both ends. The formula for predicting their frequencies is:

$$(6) \quad F_n \approx \frac{c}{\frac{2}{n}L_{rc}}$$

where  $L_{rc}$  is the length of the resonating cavity. The second complication is that the constriction interacts acoustically with the cavity behind it, producing a *Helmholtz* resonance. Its frequency ( $F_h$ ) is:

$$(7) \quad F_h \approx \frac{c}{2L_c} \sqrt{\frac{A_c}{A_b L_b}}$$

where  $A_c$  is the constriction's cross-sectional area,  $L_c$  is its length, and  $A_b$  and  $L_b$  are the cross-sectional area and length of the cavity behind the constriction. If a constriction with a cross-sectional area of 1 cm<sup>2</sup> and a length of 3 cm is moved incrementally from 3 cm above the glottis to 0.5 cm back of the lips along a 17.5 cm oral cavity, the back and front cavities produce the resonance frequencies displayed in Figure 8, along with the Helmholtz resonance. The cross-sectional areas of the unconstricted portions of the cavity are 8 cm<sup>2</sup>. The arrows projected down from the intersections between back and front cavity resonances show that F3 changes its association from the front to the back cavity when constriction's center is 7 and then again 13 cm from the glottis, and F2 changes its association from the front to the back cavity when the constriction's center is 11 cm from the glottis.

Please put Figures 7 and 8 about here.

As indicated below the horizontal axis in the figure, the centers of the lingual constrictions in the point vowels [a, u, i] are roughly one quarter (4.5 cm), two thirds (11.5 cm), and three quarters (13 cm) of the distance from the glottis to the lips. The lingual constriction for [g] is roughly the same as that for [u], while that for [d] is about seven eighths the distance from the glottis to the lips (15.5 cm), and [b]'s constriction is of course at the lips (17.5 cm).

The Helmholtz resonance is lowest for all constrictions, and lowers progressively as the constriction is moved forward. This resonance thus constitutes F1 for all constrictions. The cavity association heuristic successfully predicts that the low vowel [a] has a higher F1 than the high vowels [u, i], and also that F1 is low following all of [g, d, b]. The heuristic also successfully predicts that [a]'s F2 (the first resonance of the front cavity) is low, but it incorrectly predicts that [i]'s F2 (the first resonance of the back cavity) is also low, indeed lower than [u]'s. However, it does successfully predict that F2 and F3 (the first resonances of the front and back cavities) start at very similar frequencies following [g], because a velar constriction is close to where these formant frequencies cross at 11 cm from the glottis. The heuristic is also successful in predicting the low starting values for F2 and F3 following [b] (the first and second resonances of the back cavity), but fails to predict the higher values for these formants following [d], which should also be the first and second resonances of the back cavity.

There are "solutions" to the failed predictions. For both [i] and [d], F2 would be high if it is the front rather than the back cavity's first resonance. For [u], the calculations are incomplete. They left out the acoustic effects of lip rounding, which lengthens the front cavity and introduces another constriction at the mouth opening. This labial constriction turns the front cavity resonance into one that is closed at both ends like the back cavity resonance and introduces another Helmholtz resonance. None of the front cavity resonances produced by this configuration are lower than the back cavity

resonances, but the Helmholtz resonance produced by the labial constriction and the front cavity is low enough to constitute the F2 observed in [u] (657 Hz if the labial constriction has a cross-sectional area of 1 cm<sup>2</sup> and a length of 2 cm and the front cavity is 4.5 cm long).

In the second heuristic, the proximity of a constriction to one of a formant's velocity minima or maxima determines how it *perturbs* that formant's frequency away from its neutral value: a constriction near a velocity minimum raises the formant's frequency, while one near a maximum instead lowers it (expansions have the opposite effects). These effects will be referred to as the *minmax* rule. How does one determine whether a constriction is near a velocity minimum or maximum? Figures 4a-c show that minima occur at even quarters of the wavelength and maxima at odd quarters, and that their locations are at fixed proportions of the length of the oral cavity. Because the locations of constrictions are also a certain proportion of the distance from the glottis to the lips, it is easy to calculate whether they coincide with a minimum or maximum for a particular formant, by multiplying both sides of (4) by the proportion of the oral cavity's length that corresponds to a constriction at a particular location and rounding the result on the right hand side to the nearest quarter, as shown in Table I.

Please put Table 1 about here.

The minmax rule successfully predicts the effects of the labial, palatal, velar, and pharyngeal constrictions on all three formants of [b, i, g, a], and it successfully predicts the effects of the alveolar and velar constrictions on F1 and F3 in [d, u], but it fails to predict F2 raising after [d], and F2 lowering in [u]. The latter can again be predicted once the effects of lip rounding are added, as the simultaneous labial constriction, together with the protrusion of the lips, lowers F2 along with all other formants. To predict F2 raising after [d], its constriction would have to be somewhat further back and thus closer to a palatal constriction than it is.

Neither heuristic perfectly predicted all the observed effects of articulations on a speech sound's acoustics, but both succeeded for a majority of the sounds.

This entry has shown how speech sounds are described acoustically, and how their acoustic properties can be predicted from their articulations. Their articulations produce sound sources by transforming aerodynamic energy into acoustic form, and those sound sources in turn causes air inside to the oral cavity to resonate, at frequencies determined by the lengths of resonating cavities and where they are constricted.

Recommended reading.

- Fant, C. Gunnar M. (1960). *Acoustic Theory of Speech Production*, The Hague: Mouton.  
 Jakobson, Roman, Fant, C. Gunnar M., & Halle, Morris. (1952). *Preliminaries to Speech Analysis*, Cambridge, MA: MIT Press.  
 Ladefoged, Peter, & Maddieson, Ian. (1996). *Sounds of the World's Languages*, Oxford, UK: Blackwell Publishers.  
 Stevens, Kenneth N. (1998). *Acoustic Phonetics*, Cambridge, MA: MIT Press.



Place of constriction	Segment	Proportion of Oral Cavity Length	Proximity to Minimum or Maximum		
			Calculation	Odd/Even	Lower/Higher
Labial	b	1	$1 * \frac{1}{4} = \frac{1}{4}$	Odd	F1 lower
			$1 * \frac{3}{4} = \frac{3}{4}$	Odd	F2 lower
			$1 * \frac{5}{4} = \frac{5}{4}$	Odd	F3 lower
Alveolar	d	$\frac{7}{8}$	$\frac{7}{8} * \frac{1}{4} = \frac{7}{32} \sim \frac{1}{4}$	Odd	F1 lower
			$\frac{7}{8} * \frac{3}{4} = \frac{21}{32} \sim \frac{3}{4}$	Odd	F2 lower
			$\frac{7}{8} * \frac{5}{4} = \frac{35}{32} \sim \frac{4}{4}$	Even	F3 higher
Palatal	i	$\frac{3}{4}$	$\frac{3}{4} * \frac{1}{4} = \frac{3}{16} \sim \frac{1}{4}$	Odd	F1 lower
			$\frac{3}{4} * \frac{3}{4} = \frac{9}{16} \sim \frac{2}{4}$	Even	F2 higher
			$\frac{3}{4} * \frac{5}{4} = \frac{15}{16} \sim \frac{4}{4}$	Even	F3 higher
Velar	g, u	$\frac{2}{3}$	$\frac{2}{3} * \frac{1}{4} = \frac{2}{12} \sim \frac{1}{4}$	Odd	F1 lower
			$\frac{2}{3} * \frac{3}{4} = \frac{6}{12} \sim \frac{2}{4}$	Even	F2 higher
			$\frac{2}{3} * \frac{5}{4} = \frac{10}{12} \sim \frac{3}{4}$	Odd	F3 lower
Pharyngeal	a	$\frac{1}{4}$	$\frac{1}{4} * \frac{1}{4} = \frac{1}{16} \sim \frac{0}{4}$	Even	F1 higher
			$\frac{1}{4} * \frac{3}{4} = \frac{3}{16} \sim \frac{1}{4}$	Odd	F2 lower
			$\frac{1}{4} * \frac{5}{4} = \frac{5}{16} \sim \frac{1}{4}$	Odd	F3 lower

Table I: Calculating the proximity of a constriction to a velocity minimum or maximum of a resonance from the constriction's proportional distance from the glottis to the lips.

Figure legends:

Figure 1. Waveforms, spectra, and spectrograms of a pure tone (a-c), white noise (d-f), the vowel [i] as in *heed* (g-i), and the fricative [s] as in *see* (j-l).

Figure 2. Waveforms of the sound source produced by vocal fold vibration (heavy lines) and the first 16 harmonics of those voice qualities (light lines) for (a) modal, (b) tense, and (c) lax voice qualities. The vertical scale is twice as large in (b) as (a) or (c).

Figure 3. Spectra of modal, tense, and lax voice qualities; the dots represent the peak amplitudes of the harmonics.

Figure 4. The oral cavity as a tube closed at one end and open at the other: (a-c) standing waves corresponding to the first three resonances, with a velocity minimum at the closed end and a velocity maximum at the open end.

Figure 5. Spectra of the vowels (a) [i] as in *heed*, (b) [u] as in *who'd*, and (c) [a] as in *hod*. The individual peaks are the harmonics of the voice sound source and the formants are the ranges of amplified harmonics. Peaks corresponding to F1-F3 are labeled at the top of each panel.

Figure 6. Spectrograms of the first 150 ms of the words (a) *bad*, (b) *dad*, and (c) *gad*. The onsets of F1-F3 are labeled.

Figure 7. The configuration of the oral cavity with a constriction partway along its length, with a 1 cm<sup>2</sup> cross-sectional area partway along its length.

Figure 8. The first three resonance frequencies of the back cavity (filled symbols) and front cavity (empty symbols) produced by the articulatory configuration in Figure 7. The crosses indicate the frequencies of the Helmholtz resonance produced by the acoustic interaction between the constriction and the cavity behind it.

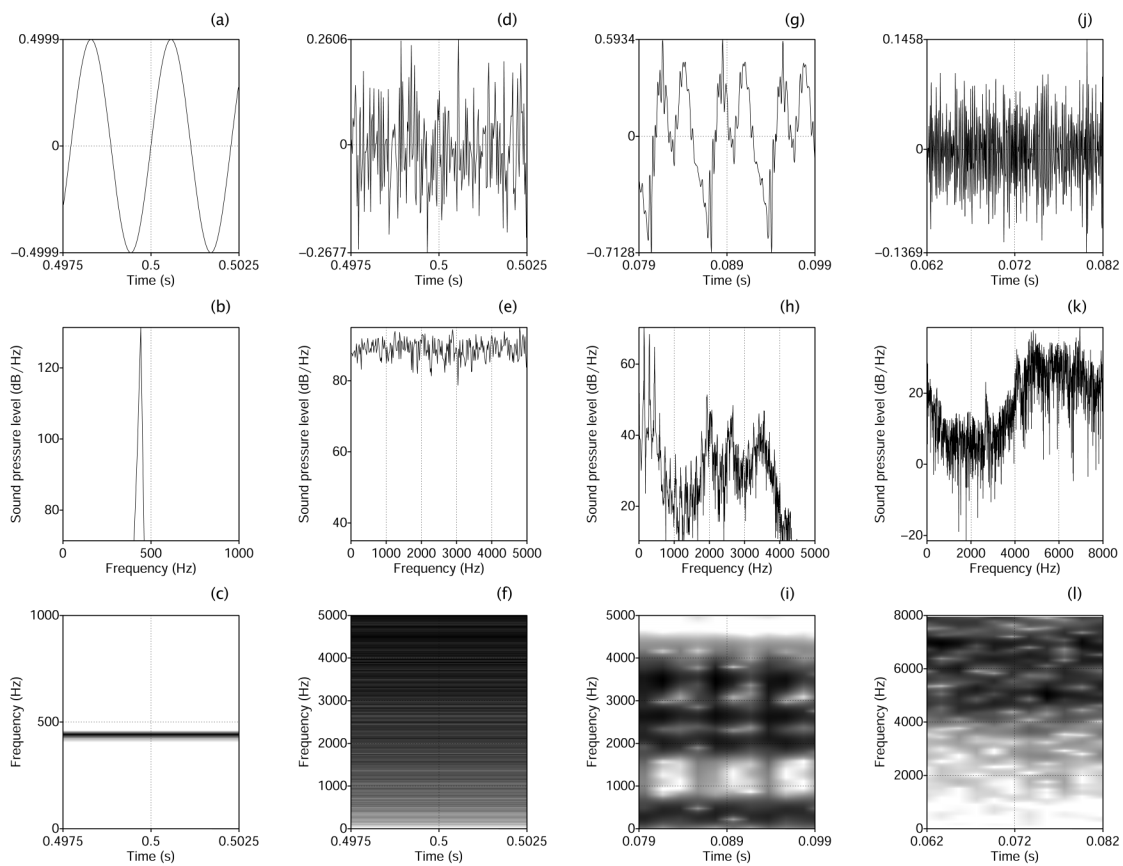


Figure 1.

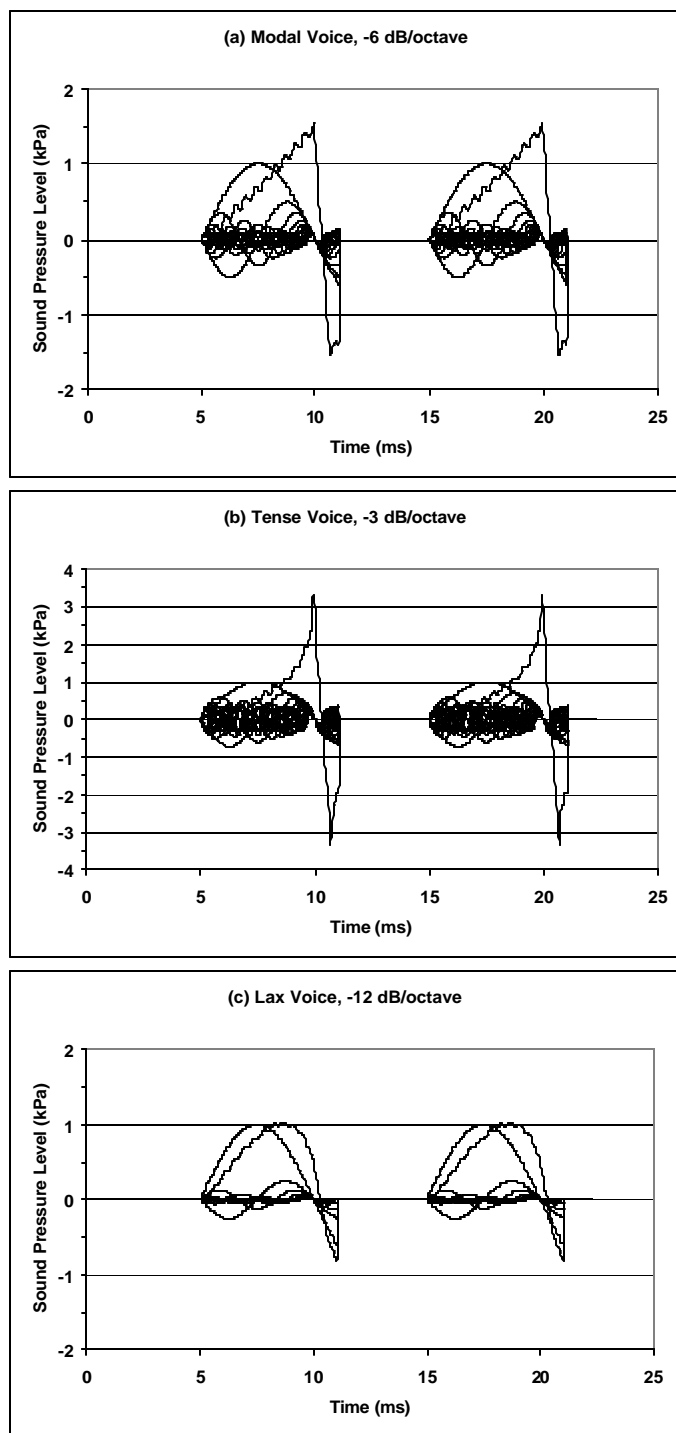


Figure 2.

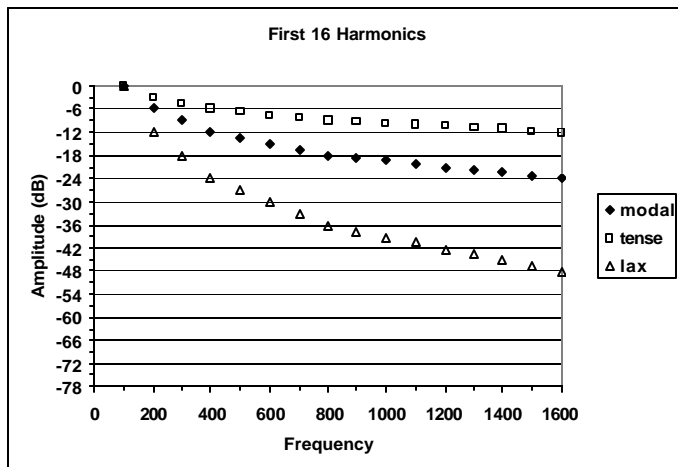


Figure 3.

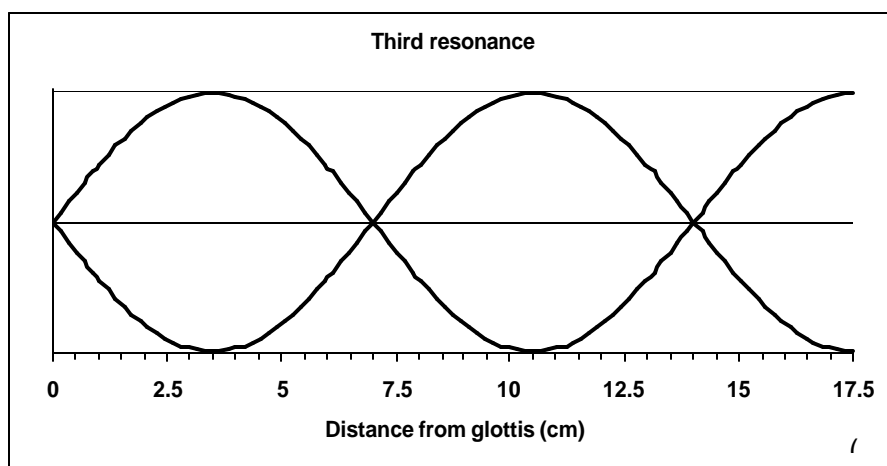
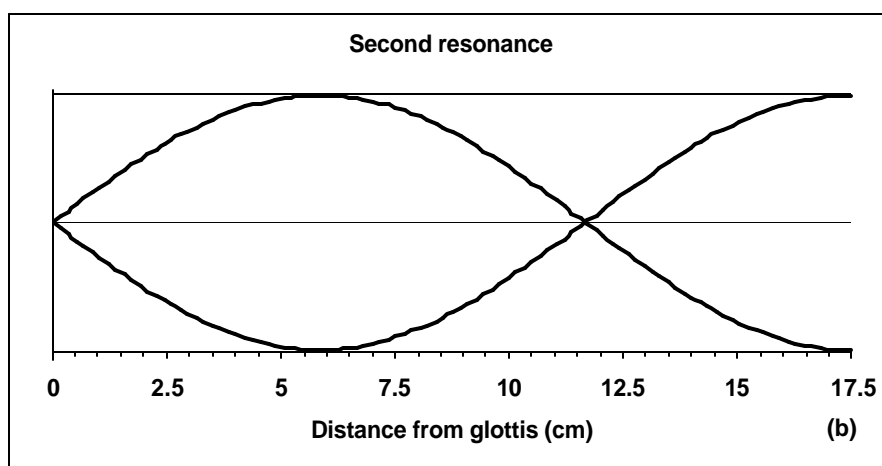
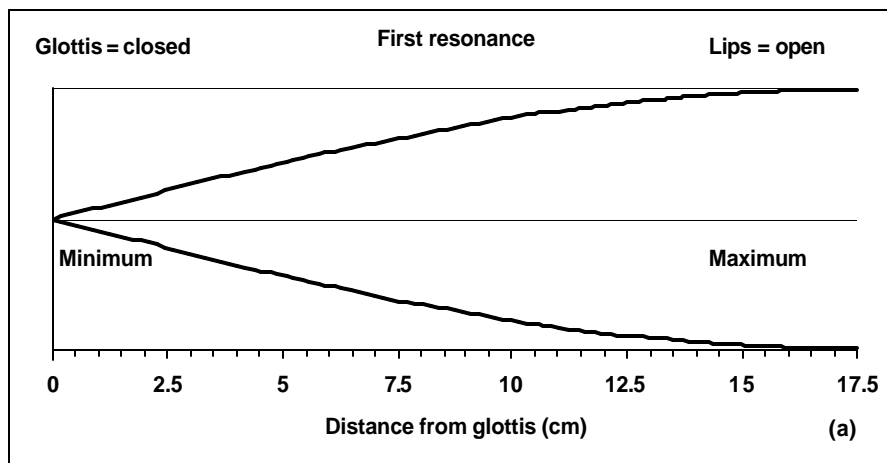


Figure 4.

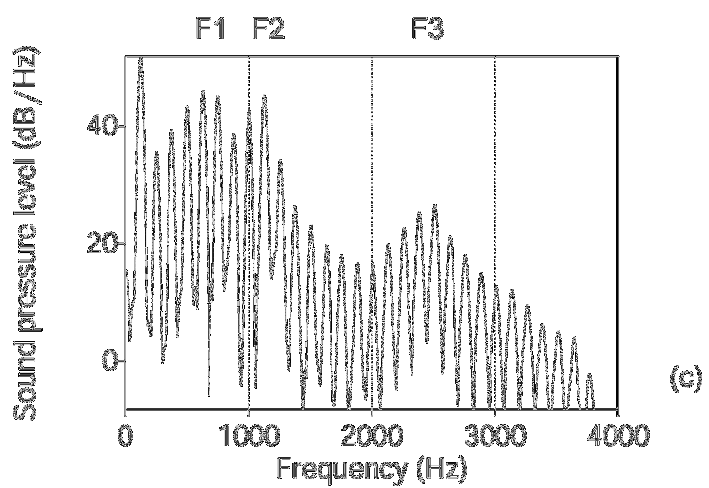
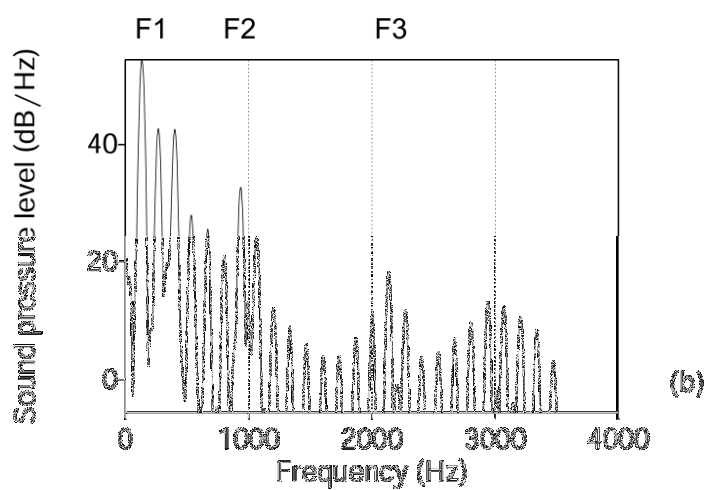
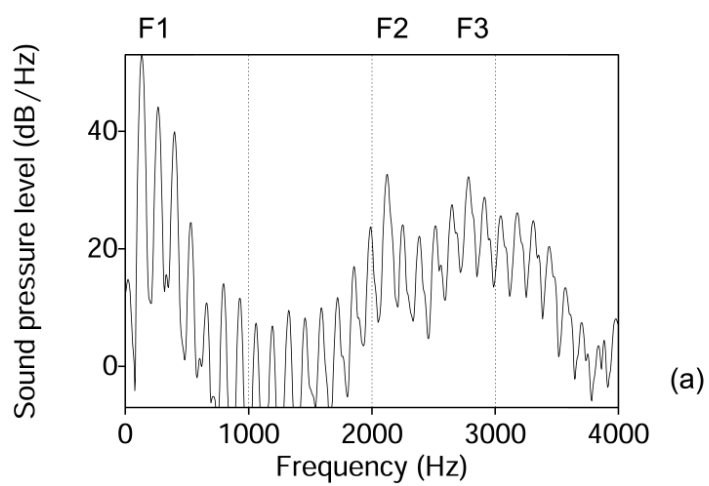


Figure 5.

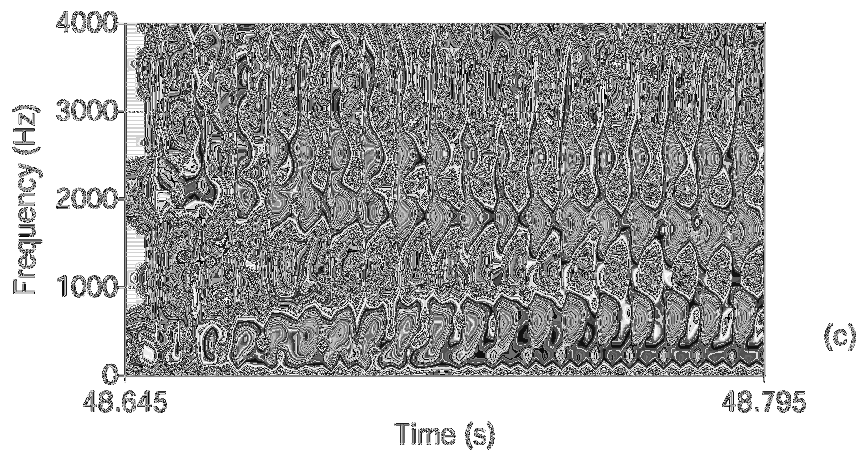
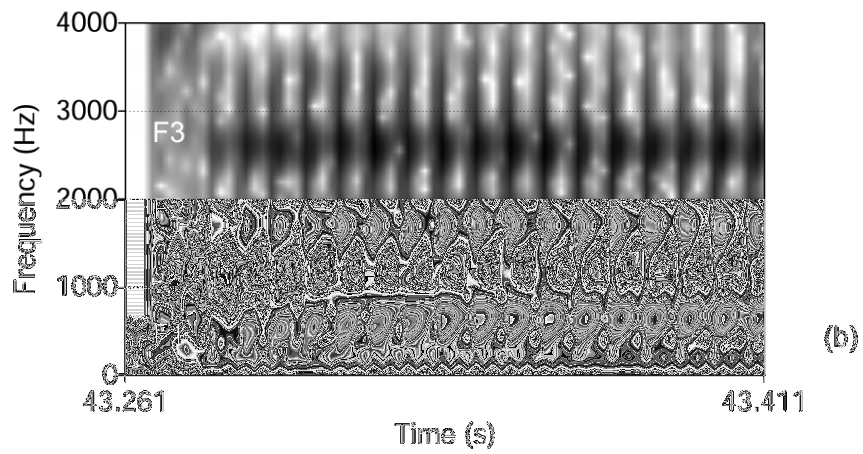
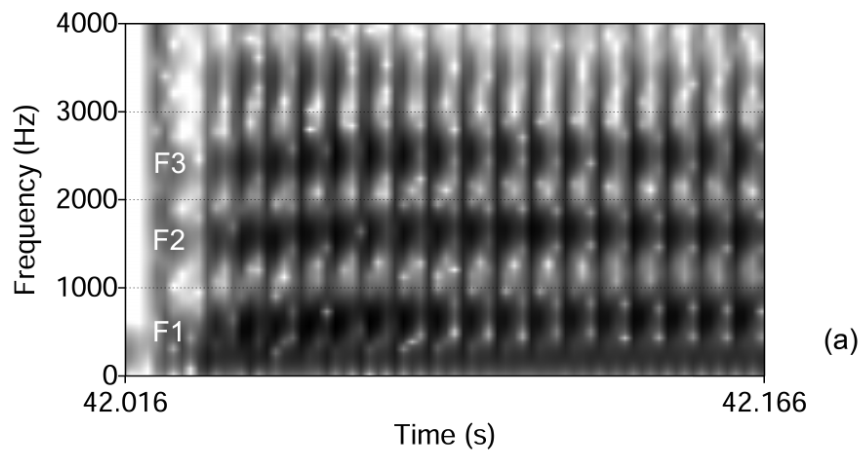


Figure 6.



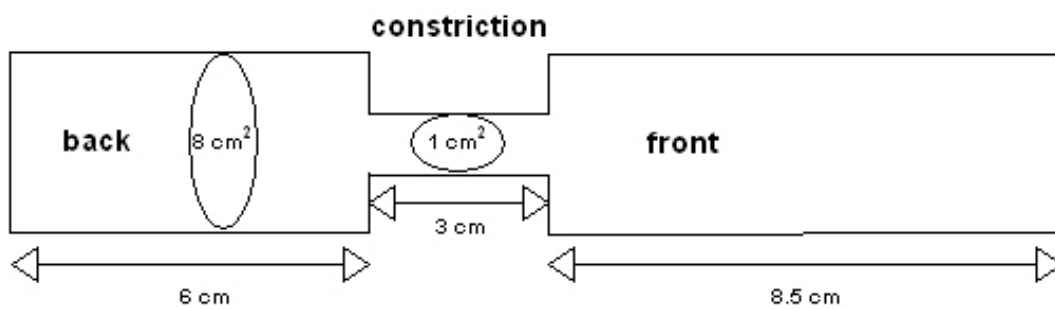


Figure 7.

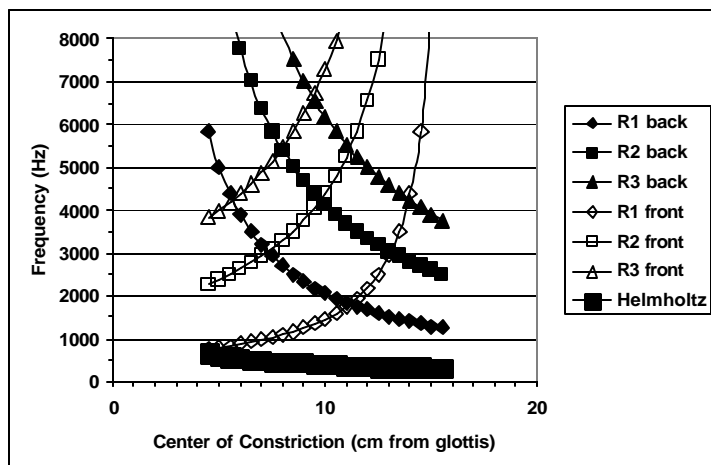


Figure 8.