Talking Heads: Speech Production

Measuring and Modeling Speech Production

The Acoustic Theory of Speech Production: the source-filter model

Acoustic speech output in humans and many nonhuman species is commonly considered to result from a combination of a source of sound energy (e.g. the larynx) modulated by a transfer (filter) function determined by the shape of the supralaryngeal vocal tract. This combination results in a shaped spectrum with broadband energy peaks. This model is often referred to as the "source-filter theory of speech production" and stems from the experiments of Johannes Müller (1848) in which a functional theory of phonation was tested by blowing air through larynges excised from human cadavers. "Müller ... noticed that the sound that came directly from the larynx differed from the sounds of human speech. Speechlike quality could be achieved only when he placed over the vibrating cords a tube whose length was roughly equal to the length of the airways that normally intervene between the larynx and a person's lips. The sound then resembled the vowel [uh], the first vowel in the word *about* ..." (from Lieberman, 1984). In this model the source of acoustic energy is at the larynx – the supralaryngeal vocal tract serves as a variable acoustic filter whose shape determines the phonetic quality of the sound (Fant, 1960).

When the larynx serves as a source of sound energy, voiced sounds are produced by a repeating sequence of events. First, the vocal cords are brought together (adduction), temporarily blocking the flow of air from the lungs and leading to increased subglottal pressure. When the subglottal pressure becomes greater than the resistance offered by the vocal folds, they open again. The folds then close rapidly due to a combination of factors, including their elasticity, laryngeal muscle tension, and the Bernoulli effect. If the process is maintained by a steady supply of pressurized air, the vocal cords will continue to open and close in a quasiperiodic fashion. As they open and close, puffs of air flow through the glottal opening. The frequency of these pulses determines the fundamental frequency (FØ) of the laryngeal source and contributes to the perceived pitch of the produced sound. An example of the spectrum of the result of such glottal air flow is plotted at the top left of Figure 2. Note that there is energy at the fundamental frequency (FØ = 100 Hz) and at the harmonics of the fundamental, and that the amplitude of the harmonics falls off gradually. The bottom left panel shows the comparable case for a fundamental frequency of 200 Hz. The rate at which the vocal folds open and close during phonation can be varied in a number of ways and is determined by the tension of the laryngeal muscles and the air pressure generated by the lungs. The shape of the spectrum is determined by details of the opening and closing movement, and is partly independent of fundamental frequency. In normal speech fundamental frequency changes constantly, providing linguistic information, as in the different intonation patterns associated with questions and statements, and information about emotional

content, such as differences in speaker mood. In addition, the fundamental frequency pattern determines naturalness of utterance production. This can be illustrated by creating a synthetic version of a natural utterance in which the spectral properties are left largely unchanged while the normally varying fundamental is replaced with a fundamental of constant frequency.

The supralaryngeal vocal tract, consisting of both the oral and nasal airways (Figure 1), can serve as a time-varying acoustic filter that suppresses the passage of sound energy at certain frequencies while allowing its passage at other frequencies. Formants are those frequencies at which local energy maxima are sustained by the supralaryngeal vocal tract and are determined, in part, by the overall shape, length and volume of the vocal tract. The detailed shape of the filter (transfer) function is determined by the entire vocal tract serving as an acoustically resonant system combined with losses including those due to radiation at the lips. An idealized filter function for the neutral vowel /ʻ/ is shown in the center panels of Figure 2 for a supralaryngeal vocal tract approximately 17cm long, approximated by a uniform tube. The formant frequencies, corresponding tothe peaks in the function, represent the center points of the main bands of energy that are passed by a particular shape of the vocal tract. In this idealized case they are 500, 1500 and 2500 Hz with bandwidths of 60 to 100 Hz, and are the same regardless of the fundamental frequency (i.e., they are the same in both the top and bottom center panels).

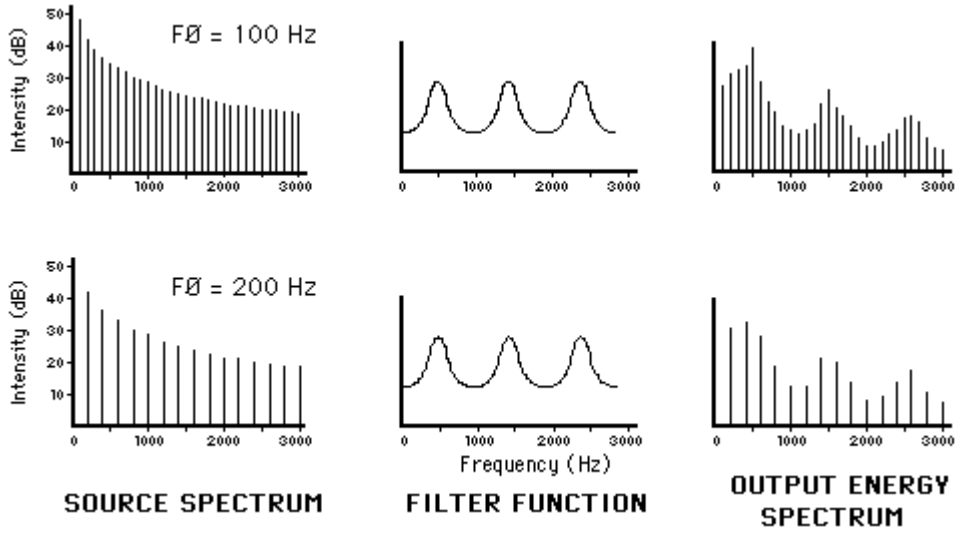

Figure 2: The source-filter model of speech production.
The source spectrum represents the spectrum of typical glottal air flow with a fundamental frequency of 100 Hz. The filter, or transfer, function is for an idealized neutral vowel /ʻ/, with formant frequencies at approximately 500 Hz, 1500 Hz and 2500 Hz. The output energy spectrum shows the spectrum that would result if the filter function shown here was excited by the source spectrum shown at the left.

The spectrum of the glottal air flow, which has energy at the fundamental frequency (100 Hz) and at the harmonics (200 Hz, 300 Hz, etc.), is plotted at the top left of Figure 2. The amplitude of the harmonics, which for the purposes of this figure combines the effects of both the source spectrum and radiation, decreases by approximately 6dB per octave. At the top right of the figure is shown the spectrum that results from filtering the laryngeal source spectrum at the top left with the idealized filter function shown in the center

of the figure. Note that the laryngeal source has been "shaped" by the filter function. Energy is present at all harmonics of the fundamental frequency of the glottal source, but the amplitudes of individual harmonics are determined by both the source amplitudes and the filter function. The bottom half of Figure 2 shows the effect of using a different source function, while retaining the same filter function. In this case, the fundamental frequency of the glottal source is 200 Hz, with harmonics at integer multiples of the fundamental (400 Hz, 600 Hz, etc.). The spectrum that results from combining this glottal source with the filter function for an idealized /ʻ/ has the same overall pattern as that shown above it. However, there are differences in the details. Note, for example, that the lowest formant for /ʻ/ has a center frequency of 500 Hz. A glottal source with a fundamental of 100 Hz will have a harmonic at this frequency. A source with a fundamental of 200 Hz will have harmonics that straddle the lowest formant (i.e., at 400 and 600 Hz), as shown at the bottom right of Figure 2. Since the overall shapes are the same, these details do not change the perceived vowel quality, which would be that of an /ʻ/. However, the top example would be perceived to have lower pitch because of its lower fundamental frequency.

The flexibility of the human vocal tract, in which the articulators can easily adjust to form a variety of shapes, results in the potential to produce a wide range of sounds. For example, the particular vowel quality of a sound is determined mainly by the shape of the supralaryngeal vocal tract, and is reflected in the filter function. Figure 3 illustrates this. Detailed accounts of the acoustic properties of the vocal tract can be found in a number of sources, including Fant (1960), Flanagan (1965), Fry (1979) and Lieberman & Blumstein (1988).

Introduction | Acoustic Theory | Measuring Production
Tract Model | Gestural Modeling | State of the Art

PREVIOUS CONTENT NEXT