Encyclopedia of Language & Linguistics, 2nd Ed.
Edited by Brown, K. Elsevier, Oxford 2006
V.9, PP. 442-460

Phonetics, Acoustic

C H Shadle, Haskins Laboratories, New Haven, CT, USA

Introduction

Phonetics is the study of characteristics of human sound-making, especially speech sounds, and includes methods for description, classification, and transcription of those sounds. Acoustic phonetics is focused on the physical properties of speech sounds, as transmitted between mouth and ear (Crystal, 1991); this definition relegates transmission of speech sounds from microphone to computer to the domain of instrumental phonetics, and yet, in studying acoustic phonetics, one needs to ensure that the speech itself, and not artifacts of recording or processing, is being studied. Thus, in this chapter we consider some of the issues involved in recording, and especially in the analysis of speech, as well as descriptions of speech sounds.

The speech signal itself has properties that make such analysis difficult. It is nonstationary; analysis generally proceeds by using short sections that are assumed to be quasistationary, yet in some cases this assumption is clearly violated, with transitions occurring within an analysis window of the desired length. Speech can be quasiperiodic, or stochastic (noisy), or a mixture of the two; it can contain transients. Each of these signal descriptions requires a different type of analysis. The dynamic range is large; for one speaker, speaking at a particular level (e.g., raised voice), the range may be −10 to 50 dB SPL (decibels Sound Pressure Level) over the entire frequency range (Beranek, 1954), but spontaneous speech may potentially range over 120 dB and still be comprehensible by a human listener. Finally, the frequency range is large, from 50 to 20 000 Hz. Though it is well known

that most of the information in speech occurs in the range of 300–3500 Hz (telephone bandwidth), if one is trying to describe and classify speech sounds a bigger range is needed.

Aspects of the recording method and the recording environment can also introduce artifacts. Breath noise can occur if the microphone is directly in front of the speaker's lips; moving the microphone further from the speaker can reduce breath noise, but then the speech signal will have a lower amplitude at the microphone, requiring a quiet recording room and possibly a more sensitive microphone. The microphone can also be moved to one side, but then the directional characteristics of speech must be considered. Higher frequencies are progressively more directional, meaning that they are highest amplitude on-axis (directly in front of the speaker's mouth) and decreasing in amplitude with angle off-axis. For instance, in the band 5–10 kHz, at 60 degrees off-axis the amplitude is 5 dB lower than at 0 degrees (on axis) (Beranek, 1954). This difference may be important for comparisons, across subjects and recording session, of parameters such as spectral tilt or formant amplitude. Differences in microphone placement can be corrected for as long as the location relative to the speaker's mouth (distance and angle) is known and the microphone is in the acoustic far-field; if in the near-field, more parameters such as the exact shape and size of the lip opening are needed.

The acoustic far field is the region where the sound pressure decreases linearly with distance from the source. The distance r from the source at which the far field begins depends on the source extent and the frequencies of interest. For instance, for frequencies greater than or equal to 350 Hz, far field begins at $r = 1$ m, and the source could be as much as 16 cm across (which is much larger than a typical lip opening, or about the size of a medium loudspeaker) (Beranek, 1954). A far-field pressure can be used to compute the equivalent source strength of the radiating surface (the air between the lips) and is, thus, important for studies in which source strength is derived from the radiated acoustic signal, or when absolute sound pressures measured at different locations need to be compared.

Background noise is often a limiting factor in microphone placement. If it is 3 dB or more below the signal, it can be corrected for (or, if 10 dB or more below, ignored), but this must be true at all frequencies of interest. There can be a big amplitude difference between the peak of the first formant of a vowel and the amplitude of a weak fricative at frequencies above 10 kHz. Solutions are to reduce the background noise by making recordings in sound-proofed, even anechoic chambers, or to use directional microphones that are more sensitive to sounds coming from their 'front' than their 'back.' Directional microphones work well at reducing background noise, but their frequency characteristics tend to be much less flat across all frequencies than those of omnidirectional microphones. Another solution is to measure the ambient noise, compare it to the signal plus noise, and filter out the frequency bands where the noise dominates the signal. This is commonly done for very low frequencies (e.g., less than 20 Hz, or often to eliminate mains hum at 50 or 60 Hz).

If it is important to know the absolute sound level of a speech signal, and keep that information intact for every kind of analysis, a calibration signal needs to be recorded as part of the original recording session and put through the same stages (amplification, filtering, sampling, analysis) as the speech. Whatever factor is needed to return the calibration signal to its known level can then be applied to the speech signal. If this is desirable, the microphone and amplifier should be of instrumentation quality, and there must not be any automatic gain control applied. This is important if one needs to compare sound levels across speakers and recording sessions.

Signal Preprocessing

While preprocessing is a relative term, it tends to be used for processes that are applied to every signal in a given system before the 'elective' processes. Thus, amplification (which may have more than one stage), filtering to remove low-frequency noise, antialiasing filtering, sampling, and preemphasis tend to be common preprocessing stages. They are best understood as changes to the spectrum of the signal. Some of the changes are reversible, such as amplification and preemphasis; some are not, because a part of the original signal is permanently lost, as in high-pass (e.g., to remove low-frequency noise) or low-pass (e.g., antialiasing) filtering. Sampling is reversible, provided a suitable antialiasing filter has been used first. Theoretically, the filter should remove all frequencies greater than half the sampling rate, that is, the cut-off frequency of the filter $f_{co} = f_s/2$. In practice, no real filter can cut off abruptly, so the cut-off frequency should be set somewhat lower than $f_s/2$; how much lower will depend on the characteristics of the filter.

If the signal being sampled includes frequencies that are greater than $f_s/2$, whether because antialiasing was not done or the cutoff was too high, they will be aliased to lower frequencies. Thus, a 6 kHz component in a signal sampled at 10 kHz will appear as energy at 4 kHz, adding to whatever energy originally occurred at 4 kHz. In general, an aliased signal cannot be unscrambled. The

anti-aliasing needs to be done for every sampling stage, whether the original sampling to convert a continuous-time signal to a discrete-time (sampled) signal, or a later downsampling to lower the sampling rate of a discrete-time signal (McClellan *et al.*, 1998). In general, one should use the highest sampling rate likely ever to be needed for that signal and apply antialiasing for that f_s; this will, of course, generate the largest number of samples and, therefore, largest file sizes, so for particular parts of the analysis where such high time resolution is not needed, the signal can be refiltered and downsampled. Systems that sample, such as DAT recorders and sound cards, now often have antialiasing filters built in; analysis software, such as MATLAB, will not necessarily perform this step automatically.

Preemphasis was originally devised to make optimal use of the small dynamic range of analog tape. A speech spectrum tends to fall off with frequency; that is, amplitudes are lower at higher frequencies. The pre-emphasis filter tilts upward smoothly and thus flattens out the speech spectrum while leaving its important peaks (such as formants and harmonics) intact relative to each other. This is still useful before computing a spectrogram, since the upper frequencies will show up better if they have been boosted in amplitude. Since it is a reversible operation and simple to describe, there is no reason not to do it, but it is important to remember when it has and has not been applied to aid comparisons.

Signal Analysis

The techniques used to analyze speech should be appropriate to the local signal properties as well as consistent with the aims of the analysis. The information that is desired is typically related to the type of speech sound – whether it is voiced or not, continuant or not, the place of constriction, and so on. We will consider speech production models later; let us first consider analysis methods in relation to the properties of the signal.

Analysis of Periodic Signals

A perfectly periodic signal repeats exactly at some time interval T0 and so has a fundamental frequency $F0 = 1/T0$. It may have harmonics, which occur at integral multiples of F0, i.e., 2 F0, 3 F0, and so repeat exactly at T0/2, T0/3, ..., respectively. There is no noise; the signal is entirely deterministic.

In the real world, there is no such thing as a perfectly periodic signal. The closest equivalent in speech is quasiperiodic, meaning that F0 changes over time and has a small amount of noise. A typical example is a vowel, with the fundamental and many harmonics. We can look at and measure the time waveform, but if we want to know the distribution of energy at the frequency of each harmonic, we need to compute some type of spectrum. The classic first step for such a signal is the Discrete Fourier Transform (DFT). The signal is multiplied by a window, and the DFT is computed of the windowed signal. If we had a perfectly periodic signal, there would be no difference in the result if we included exactly one period, or exactly two, so we could think of the window as selecting exactly one period to minimize the amount of computation. With a quasiperiodic signal, the window can exclude parts of the signal in which F0 is very different. The signal within the window is approximately stationary, and so taking a single DFT is appropriate.

The window length and shape are important. The longer the window is, the finer the frequency resolution will be; the shorter it is, the coarser. In other words, the resolution in time is inversely proportional to the resolution in frequency. There is one wrinkle in this simple statement, however; the frequency resolution depends not only on the window length, but also on the number of points used to compute the DFT. If we want to be able to see every harmonic defined, we need fine resolution – perhaps 50 Hz between points on the DFT. But then the time window may be long enough for the signal to change properties somewhat; if so, the harmonics that are computed will be an average of the different sets of true values that occurred during the windowed signal.

How does the number of samples used to compute the DFT, which we call N_{DFT}, interact with the window length, and why would we ever want a N_{DFT} to be longer than the window, since all values outside the window are zero by definition? The short answer is that the number of points used to compute the DFT actually controls the frequency resolution. The Fourier transform of a discrete-time signal is a continuous function of frequency; the DFT samples that transform in the frequency domain, spreading N_{DFT} points evenly between $-f_s/2$ and $+f_s/2$. This means that the bigger N_{DFT} is – the more samples used to compute the DFT – the more tightly packed the samples are in the frequency domain and, thus, the finer the frequency resolution. The technique is called zero-padding, because the windowed signal is 'padded' with zeros to match the length of the DFT.

If the signal thus treated is perfectly periodic, it has energy only at the harmonics of its fundamental frequency. Increasing the frequency resolution beyond the harmonic spacing will not reveal anything else since there is not any other energy to see. There is a hazard, however; increasing resolution slightly beyond the harmonic spacing can mean that some of

the harmonics are missed. Increasing the resolution well beyond the harmonic spacing is less problematic; zeros between the harmonics will be revealed. If the windowed signal is not perfectly periodic, energy will exist between harmonics, and a longer N_{DFT} will define the shape of the transform of the samples occurring within the window more accurately. Zero-padding to use a longer N_{DFT} does not provide any more information about the properties of the signal, but does allow what is there to be seen better.

Finally, using a longer N_{DFT} does incur a computation cost, since increasing N_{DFT} increases the number of operations required to compute the DFT. The Fast Fourier Transform is an algorithm developed to compute the DFT efficiently; if N_{DFT} is a power of two (e.g., 64, 128) the computation will be faster. However, a 1024-point DFT will still take longer to compute than 128 or 512 points, and so N_{DFT} should always be justified in terms of the signal properties and the information sought by the analysis. We will return to this subject in the next section.

There are many window shapes, starting with the rectangular window, which weights every sample equally and cuts off to zero abruptly, and progressing to the gradually tapered windows typically used in speech analysis, the Hanning and Hamming. Since they are tapered at each end, there is no abrupt change in amplitude, which could create an artifact of seeming noise in the signal. They also have better properties in the frequency domain than does the rectangular window, minimizing the amount of leakage of one spectral component into neighboring components. **Figure 1** contrasts two Hanning window lengths used to analyze the same signal to produce DFTs and (as discussed below) LPC spectral envelopes. **Figure 1A** uses a window of 60 ms; every harmonic is clearly shown. **Figure 1B** uses a window of 10 ms. The major peaks are still visible, at approximately 250, 2600, and 3800 Hz, but the rest of the spectrum has been flattened. The peak at 250 Hz is wider, because it includes the energy for two or three harmonics, as we know from examining **Figure 1A**.

The DFT is plotted as amplitude, or log amplitude, vs. frequency. The speech spectrogram is made up of a sequence of DFTs, each computed for the same length of windowed signal and plotted as frequency vs. time, representing the spectral amplitude in greyscale. A fine-grain effect is achieved by having a skip factor that is much shorter than the window length; Olive *et al.* (1993) specified that they used a 30-ms window for their wideband spectrograms, skipping that window along 1 ms at a time. The narrowband spectrogram uses a bandwidth of 25–50 Hz and resolves every harmonic; the wideband spectrogram uses a bandwidth of 200–300 Hz and blurs the harmonics together, which shows the more widely separated formants better.

Linear Prediction Coding (LPC) is often used for a different type of spectral analysis of quasiperiodic sounds. LPC analysis consists of finding the best set of coefficients to predict the entire signal in a frame from a few of its samples. The user chooses how many samples will be used and, thus, how many coefficients will be computed; this specifies the order of a polynomial. In the frequency domain, the order specifies how many poles there will be in what is known as the LPC spectral envelope. Two poles are needed for each peak in the envelope, plus another two for overall spectral tilt. Thus, if $f_s = 10$ kHz, and the order is 12, then the spectral envelope will be a smooth envelope that captures the main five peaks of the DFT spectrum. The peaks will often, but not always, line up with the formants; two formants near each other in frequency may be represented by one peak. Increasing the order to, say, 40 will allow 20 peaks to be found in the same spectrum; depending on the actual F0, these peaks may coincide with harmonics.

Referring to **Figure 1** again, we note that the two LPC spectral envelopes are similar though not identical. A minor peak around 3 kHz is more noticeable when the wider time window is used; in the range 5–8 kHz, three peaks are visible for the short window, only two for the long window, and there are other differences at higher frequencies. There is little evidence that the signal has changed properties substantially within the 60-ms window, or the harmonic peaks would be wider; therefore its spectral representations are likely to be more accurate in this case.

The cepstrum offers another way to compute a spectral envelope. If you took a DFT and then immediately an inverse DFT (the IDFT), you would recover the original time waveform. With the cepstrum, the DFT is computed; then the log is taken, and then the IDFT. The result is called the cepstrum, in a domain that is not quite time, not quite frequency, and is known as quefrency. It is best understood by thinking of the log spectrum as if it were a time waveform. The closely packed patterns, the harmonics, would represent evidence of high-frequency components if the DFT were a time waveform; the wider-spaced patterns, the formants, would represent lower frequency components. These elements end up separated in the cepstrum into what is referred to as high-time and low-time components, respectively. If only the low-time components are selected and then the DFT is taken, the result is essentially the spectral envelope without the harmonic spikes. Unlike the LPC spectral envelope, the cepstrum will fit the troughs as well as the peaks of the spectrum. The only caveat is that the

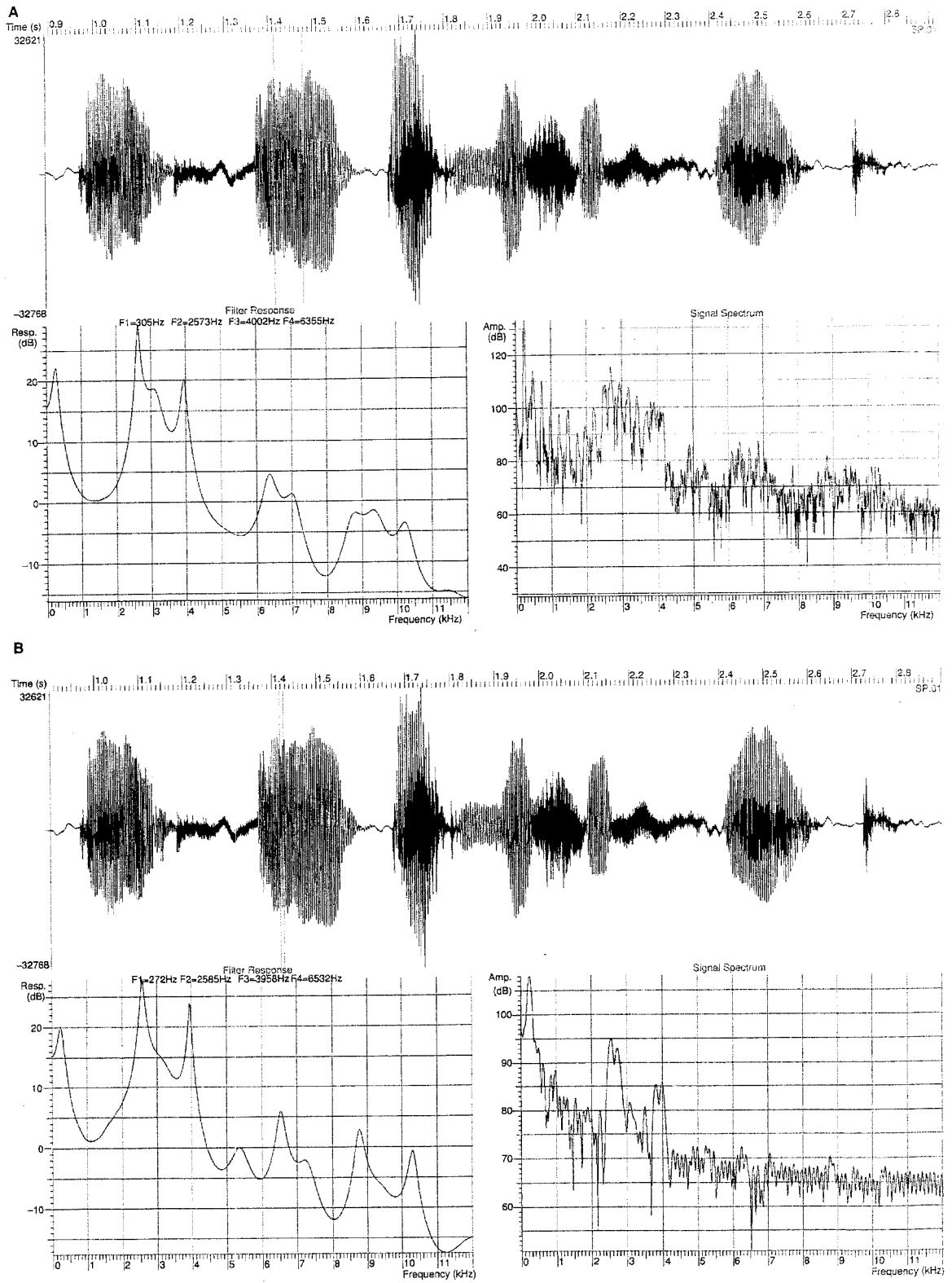

Figure 1 Same speech signal is analyzed with two lengths of Hanning window and LPC. (A) Waveform of "Don't feed that vicious hawk" is shown on top, with cursors marking the 60-ms window in [i] of 'feed.' DFT is lower right; LPC spectral envelope is lower left. (B) Same waveform is shown, with cursors marking a 10-ms window with the same starting point as in (A). DFT is lower right; LPC is lower left.

high- and low-time components must not overlap, which means that the process works well for low-F0 voices, and less well for higher F0 (Gold and Morgan, 2000).

With periodic signals, it is often desirable to find out what the period (or, equivalently, fundamental frequency) is; a secondary question is to determine the entire source spectrum. Many F0 trackers exist and can be roughly grouped into time-domain and frequency-domain algorithms. If a person measured F0 from a time waveform, they would look for a repeating pattern using any number of cues such as the highest-amplitude peaks or longest up- or down-slope; they would check earlier and later to make sure that, even though the pattern is slowly changing, the interval of quasirepetition seems consistent; and finally, they would measure the time interval between repetitions and invert that value to obtain a local estimate of F0. Such manual tracking, when done by people with some training, is extremely consistent across trackers and has been used as the gold standard by which to evaluate computer algorithms. It should not be surprising, then, that some of the most successful algorithms use similar simple parameters defined on the time waveform (Rabiner and Schafer, 1978; Gold and Morgan, 2000).

Another time-domain algorithm takes a different approach, beginning by computing the autocorrelation of a windowed part of a signal with itself. The signal and its copy are aligned, the product is computed of each sample with its aligned counterpart, and the products are summed. The value resulting is that for the lag $\tau = 0$. Then the copy is shifted by one sample, and the process is repeated, with products being formed of each sample with its one-sample-earlier counterpart. The new sum is computed for $\tau = 1$. As the signal and its copy get more and more out of alignment, the sum of products decreases – until they are misaligned by one pitch period, and then the sum will have a high value again. When the total lag equals two and three pitch periods the sum will peak again, but because the two signals overlap less and less, successive peaks will be smaller. The algorithm computes the autocorrelation and then finds the peaks in the signal. The lag τ of the first peak is taken as T0, of the second peak is 2 T0, and so on until the peaks are too low in amplitude to be reliable indicators. Autocorrelation-based F0 trackers work better on high F0 voices, because the pitch periods are shorter so more of them fit within the same size window (Rabiner and Schafer, 1978).

Frequency-domain F0 trackers use some form of a spectrum in which the harmonics are visible. The peaks are found, and their lowest common divisor is determined. This method can work even if the fundamental and some of the harmonics are missing (as in telephone speech). Preprocessing, especially using low-pass filtering, though sometimes more elaborate, is used. In one algorithm LPC analysis is used to find the formants; an inverse filter is then devised and multiplied by the original signal to remove the formants, leaving the harmonics of now nearly uniform amplitude. Then LPC analysis with a higher order is used, and the peak frequencies, and their lowest common divisor, are found (Gold and Morgan, 2000).

F0 trackers have been compared extensively. Some work better with speech recorded in noisy environments; some work better with high, or low, voices. Generally, voices become very difficult to track when they verge into vocal fry, diplophonia, or other kinds of vocal instability. A manual tracker may be able to discern periodicity where an automatic tracker has declared a signal unvoiced. Most trackers include heuristic thresholds that, for instance, do not allow octave skips in the output F0 values. This is unfortunate when the speaker has actually produced a sudden octave change by going into falsetto or yodeling.

Analysis of Stationary Noise

In completely random noise, adjacent samples are uncorrelated, and the noise must be described statistically. The time waveform can be described by the probability distribution of amplitudes, and that distribution can be described by its mean, variance, and higher moments. The noise can also be described by its power spectrum, and can be classified in general terms as wideband or narrowband noise. White noise is flat across all frequencies and therefore is wideband. One can think of the bandwidth of noise in terms of the rapidity of the variation possible in the time domain.

For all such descriptions of noise, stationarity means that the properties of the noise do not change with time. If this is true, we can collect a very long example of the signal to analyze; equally, we could collect sections of it today, tomorrow, and next year and assume that the mean, variance, and higher moments are the same in all of our samples.

In the real world, signals carrying information are not perfectly stationary. As with periodicity, though, we can declare something to be quasistationary if its properties do not change very fast compared to the intervals we are interested in; alternatively, we can assume that a signal is stationary and, as part of the analysis, try to determine if that assumption is valid.

In speech, the central portions of unvoiced fricatives are often treated as if stationary; sometimes the entire fricative is treated this way, even though the

transitions are clearly regions of rapid change. If nonstationary noise is treated as stationary, the result is likely to be a sort of muddling together of the changing values describing the noise. However, nonstationary noise is sometimes analyzed as if it were a deterministic signal, and this is likely to lead to erroneous conclusions.

From comments in the previous section on how the frequency resolution depends on the length of the DFT (N_{DFT}), it might seem that the best way to analyze a noisy signal would be to use a relatively short window so that the noise within it is close to stationary, and then use a big N_{DFT} so that the resulting transform is sampled with a fine frequency resolution. But it is possible to prove that taking a single DFT of noise results in a spectrum with an error of the same magnitude as the true value. Some form of averaging must be done in order to describe noisy signals. Using a longer window (and DFT) before taking a single DFT, which intuitively seems to be a good idea because more samples are included, does not help; the frequency resolution becomes finer, but the values still have a large error. If, on the other hand, the samples in the long window are subdivided into many short windows, the DFT is computed for each short window separately, and the results averaged at each frequency, the resulting averaged power spectrum converges to the true value. If each window contains independent, identically distributed samples of the same underlying process, the variance of the estimate decreases as the number of such windows increases (Bendat and Piersol, 2000). This is shown graphically for white noise being time-averaged with an increasing number of averages in **Figure 2**.

There are three ways in which averaging can be done, each of which will reduce the error of the spectral estimate, but each also with its own pros and cons. The method just described, of chopping a long interval into short windows, is called time averaging (see **Figure 3A**). If Hanning or Hamming windows are being used, the samples at the tapered edges can be 'reused' by overlapping windows to some degree: rules of thumb range from 30 to 50% overlap. In this way, 100 ms of signal could be chopped into nine overlapping 20-ms windows, which could significantly improve the variance of the estimate – provided that the signal is more or less stationary during the 100 ms. The practice used in some speech studies of overlapping the windows much more than this (e.g., using a 20-ms window and a skip factor of 1 ms, so that 40 ms of signal is used to generate 21 windows and, thus, 21 DFTs) has two disadvantages: the variance of the estimate is not reduced proportionate to the number of averages, and the result is weighted toward the characteristics of the middle 20 ms of the 40 ms, since that is the most heavily overlapped portion.

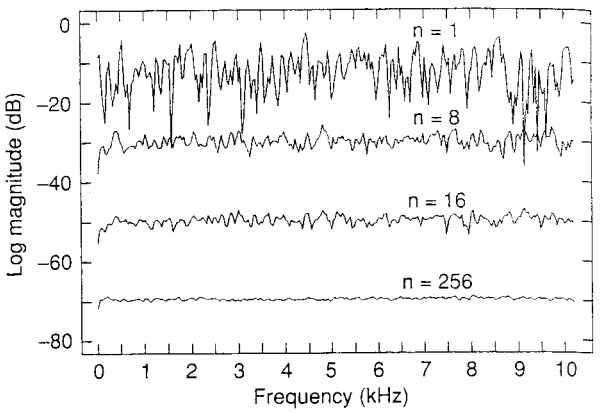

Figure 2 White noise, analyzed with time-averaging. The number of DFTs computed and averaged at each frequency is shown as an n value with each curve. (From Shadle (1985) The acoustics of fricative consonants. PhD thesis, MIT, Cambridge, MA. RLE Tech. Report 504, with permission.)

A second way is to compute the ensemble average (see **Figure 3B**). An ensemble of signals means essentially that different signals have been produced under identical conditions, leading to the same properties for the noise in each signal. The noise properties can vary in time, but the time variations must be the same for each member of the ensemble. For instance, if our signal is the sound of raindrops falling on the roof, and they fall louder and faster as the wind blows harder, then an ensemble could consist of raindrops falling in ten different storms, in all of which the wind increased at the same rate. We place our windows at the same time in each signal (relative to the wind speed, or other controlling parameter), compute the DFTs, and average as for the time average. The obvious problem here is in knowing that every member of the ensemble had the same controlling parameters at the same times. However, each individual signal does not need to be stationary for more than the length of the short window.

A third way is to compute the frequency average, by computing a single DFT and then averaging in the frequency domain (see **Figure 3C**). Ten adjacent frequency components can be averaged to produce a single component. This reduces the frequency resolution but improves the error. However, this works well only if the spectrum is fairly flat. If the spectrum has significant peaks or troughs, the frequency averaging will flatten them and so introduce bias to the estimate, meaning that it will converge to the wrong value.

For speech, all of these methods have been used, but none is ideal. Another method exists and is beginning to be used in speech research: multitaper

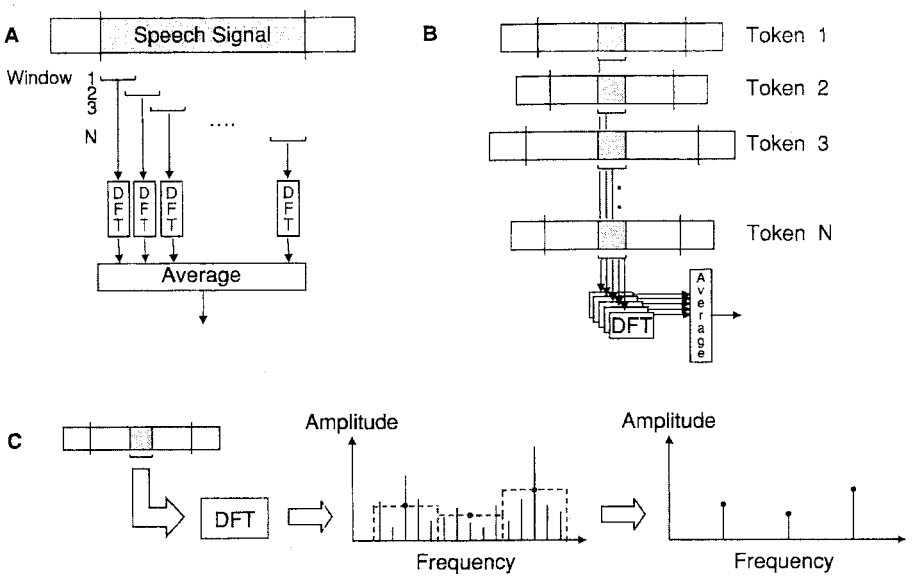

Figure 3 Diagrams indicating which parts of the signal(s) are used to generate averaged power spectra. Each rectangular box represents a part of a speech time waveform; shaded regions indicate the part being analyzed. Brackets indicate length of the window for which the DFT is computed. The 'Average' boxes compute an average of the DFT amplitude values at each frequency. (A) Time averaging. (B) Ensemble averaging. (C) Frequency averaging.

analysis. With this method, a single short signal segment is used, but it is multiplied by many different windows – called tapers – before computing and averaging their DFTs. The particular shape of the tapers satisfies the requirement for statistical independence of the signals being averaged. **Figure 4** compares a multitaper estimate and a DFT spectral estimate of the same central portion of an [s]. The jaggedness of the DFT curve can provide a rough visual indication of its greater error compared to the multitaper curve. Spectrograms can be constructed of a sequence of multitaper estimates and plotted similarly. There are important choices to be made about the number of tapers to use and other parameters, but the method offers advantages in speech analysis over the three averaging techniques described above (Blacklock, 2004).

Note that spectrograms, although they do not include spectral averaging explicitly, are not as misleading as using single DFTs for noisy sounds. Essentially, the eye averages the noise, aided by the use of a small skip factor in the computation. The same is not true of spectral slices derived from a spectrogram; since these are constructed from a single DFT, there is nothing shown for the eye to average. This problem was recognized in an early article about the use of the spectrogram (Fant, 1962: Figure 6, p. 20).

Analysis of Mixed Noise and Periodic Signals

Mixed-source signals would seem to call for two different analysis techniques. Examples in speech include voiced fricatives and affricates, and also breathy or hoarse productions of vowels, liquids, and nasals. In all of these cases the signal analysis is complicated by the fact that the noise and voicing source are not independent; in voiced fricatives the noise can be modulated by the voicing source, and breathy or hoarse sounds are likely to change as the vocal folds vibrate, even if the noise is not specifically modulated by the acoustic signal.

Mixed-source signals should be analyzed with time averaging, ensemble averaging, or multitaper. If the

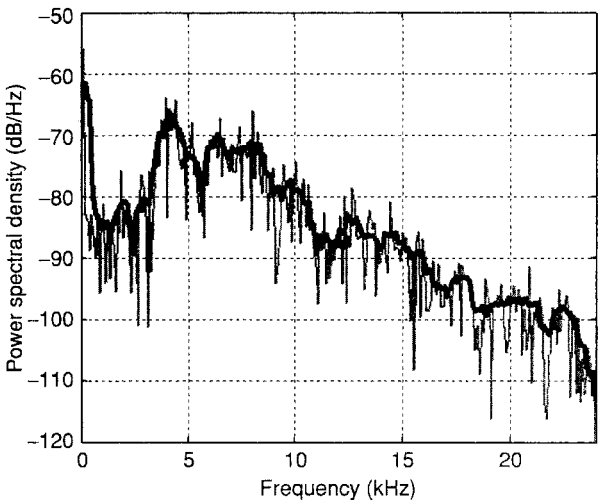

Figure 4 Multitaper spectrum of [s] in 'bassoon' in blue (smooth curve) overlaid on DFT of same signal in red (jagged curve). British male speaker. (After Blacklock, 2004.)

periodic component is stationary, the spectral averaging will not affect it, but will reduce the error in the estimate of the noisy components. If F0 of the periodic component changes noticeably during the interval or across the ensemble averaged, the harmonics will be smeared out, which may be obvious in the averaged power spectrum, or may become clear when that is compared to a spectrogram. In that case, time averaging should be avoided in order to decrease the averaging interval length.

Mixed-source signals can also be decomposed into two parts, harmonic and anharmonic. A wide variety of algorithms exist that accomplish this. After decomposition, each component can be analyzed in the way appropriate to a harmonic signal and a noisy signal, respectively. Jackson and Shadle (2001) reviewed such algorithms and presented their own, which was used to investigate voiced fricatives. Multitaper analysis can also be formulated to identify harmonics mixed with colored noise; a detailed comparison of the two techniques has not yet been made.

Analysis of Noisy Transients

A transient includes nondeterministic noise, is highly nonstationary, and is generally very short. An example from speech is the stop release. Because it is noisy, it requires averaging, but with such a short signal that is difficult to do. Ensemble averaging is possible, but an independent means of aligning the signals in the ensemble would need to be established. Multitaper is also a possibility.

Production Models

We turn now from consideration of analysis techniques appropriate to the type of signal to models of speech production that indicate the parameters we seek from analysis in order to describe and classify sounds. The vast majority of speech production models that are useful for this purpose are source-filter models, with independent source and filter, and linear time-invariant filter. The assumption of independence is flawed – interactions of all sorts have been shown to exist – but it serves well for a first approximation, in part because the models become simple conceptually. The source characteristics can be predicted, and the source spectrum multiplied by the transfer function from that source to an output variable such as the volume velocity at the lips. (If both characteristics are in log form, it is even simpler; they can just be added at each frequency.) While it took years to develop the theory underlying the source characteristics and the tract transfer functions, it is now straightforward to vary a parameter such as F0, a formant frequency, or pharynx cross-sectional area in such a model and see its acoustic effect. It is not so straightforward to analyze the far-field pressure into true source and filter components.

Sources

There are two basic types of sources: the voicing source, generated by vocal fold vibration and nominally located at the glottis, and noise sources, which can be located anywhere in the vocal tract, including at the glottis. In both cases the location of the source is where some of the energy in the airflow is converted into sound. Determining the exact location for noise sources is still a subject of research, and slight differences can affect the predicted radiated pressure significantly.

A number of factors affect the voicing source: subglottal pressure, degree of adduction of the vocal folds, tension of the folds, and supraglottal impedance. They determine, first, whether the vocal folds vibrate and, if so, the frequency at which they vibrate and the mode or register of vibration. The frequency of vibration affects F0 and all its harmonics; the mode of vibration affects the amplitude of all harmonics and also whether noise will also be produced (as in breathy or whispered speech). These differences can be characterized in the time waveform of the glottal volume velocity, $U_g(t)$, or in its spectrum, $U_g(f)$. As a general rule, abrupt 'corners' or changes of slope in the time waveform, which occur for the more adducted registers like modal register or pressed voice, mean there will be more high-frequency energy, i.e., the harmonics will have higher amplitudes compared to falsetto or breathy voice. **Figure 5** shows a typical glottal waveform, with a clear closed phase, and a range of possible spectra; the steeper the slope (e.g., −18 dB/oct), the smoother the time waveform, with a sound quality as in falsetto; the shallower slopes (−12, −6 dB/oct) correspond to a richer, brassier sound. The sound quality is related to the spectral tilt; the spacing of the harmonics that define the spectrum is related to T0, the spacing between glottal pulses (Sundberg, 1987; Titze, 2000).

Noise sources occur when the air becomes turbulent and the turbulence produces turbulence noise. Whether turbulence occurs is determined by the Reynolds number, $Re = VD/\nu = UD/A\nu$, where V = a characteristic velocity, D = a characteristic dimension, typically the cross-dimension where V is measured, U = volume velocity, A = cross-sectional area where D is measured, and ν = kinematic viscosity of the fluid = 0.15 cm^2/s for air. If V increases while D remains the same, or if U stays the same but A decreases, Re will increase. Thus, although the volume velocity must be the same all along the tract (since there is nowhere else for the air to go), the

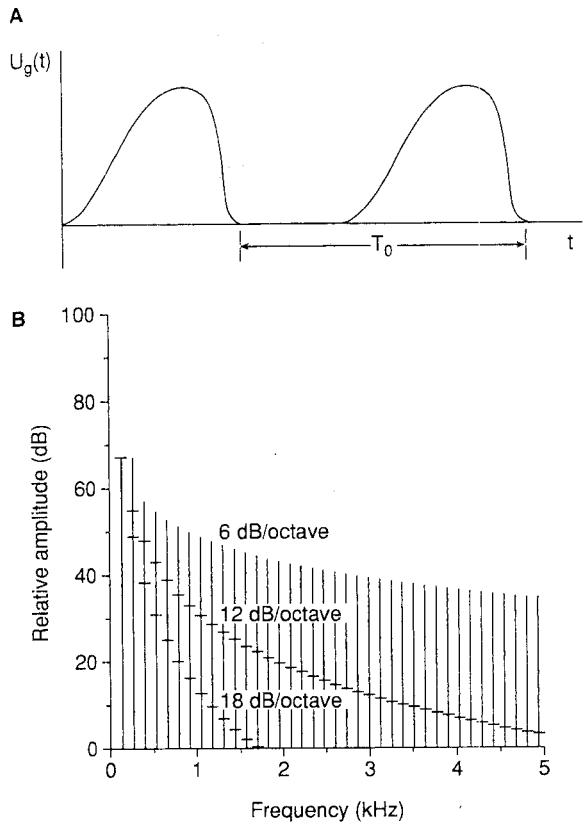

Figure 5 Plots of typical glottal volume velocity, as (A) time waveform $U_g(t)$, and (B) spectrum, $U_g(f)$. (Adapted from Titze (2000) *Principles of voice production* (2nd printing). Iowa City, Iowa: National Center for Voice and Speech, with permission.)

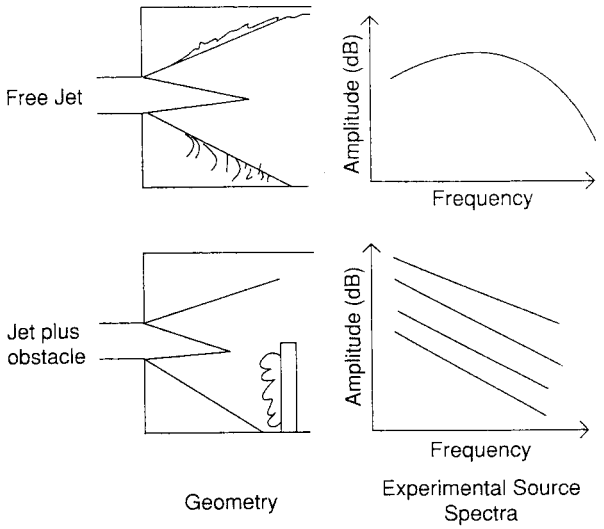

Figure 6 Noise source: shape downstream of constriction and spectrum of noise resulting experimentally, for free jet (top) and jet impinging on an obstacle (bottom).

Reynolds number will be highest at the points of greatest constriction. When Re is greater than a certain critical Reynolds number, Re_{crit}, the jet emerging from the constriction will become turbulent, but where the most noise will be generated depends on the geometry downstream of the constriction.

The simplest model for such turbulence noise is to treat it as completely localized at one place, and place a series pressure source at the equivalent place in the model. The strength of the source, p_s, is related to the parameters that affect the amount of turbulence generated: the pressure drop across the constriction, and the volume velocity and area of the constriction. The spectral characteristic should be broadband noise; sometimes, for convenience, it has been defined as high-pass-filtered white noise (Flanagan and Cherry, 1969). Stevens specified a broad peak characteristic of free jet noise (though free jet noise generation is distributed along the length of the jet) (Stevens, 1971), but some experiments indicate it should have a characteristic with its amplitude highest at low frequencies (Shadle, 1990) (see Figure 6). The location of the source has been experimented with; Flanagan and Cherry (1969) placed it 0.5 cm downstream of the constriction exit; Fant (1970) sought the location generating the best spectral match for each fricative; Stevens (1998) has demonstrated the difference made by placing it at any of three locations downstream. It seems clear that, for some fricatives, a localized source and a characteristic of spoiler in duct is fine, while for others, a distributed source with the broad peak characteristic of a free jet is needed (Shadle, 1990).

Because p_s is related to the pressure drop across the constriction, the amount of noise will change as the constriction area changes (as is needed during a stop release, or in the transitions into and out of a fricative) and as the pressure just upstream of the constriction changes (as when the pressure drop across the glottis changes). Modulation of p_s by the glottal volume velocity is possible in such a model (Flanagan and Cherry, 1969), though the actual mechanism affecting the source in voiced fricatives appears to be somewhat more complex than can be modeled by their synthesizer (Jackson and Shadle, 2000).

Filters

The filtering properties of the vocal tract depend on its shape and size and, to a small extent, on the mechanical properties of its walls. Wherever sound is generated in the tract, sound waves emanate outward from that point. At any acoustic discontinuity (such as a change in cross-sectional area, or encountering a solid boundary), some of the wave may travel onward, and some may reflect. Reflected waves can interfere with sounds emitted from the source later, combining constructively and destructively. At

frequencies in the sound where the interferences recur at the same spatial positions, standing wave patterns will be set up.

Many explanations of standing wave patterns exist (see, for example, Stevens, 1998; Johnson, 2003). It is simple to compute the frequencies at which such patterns will occur for lossless uniform-diameter tubes, where only two cases matter: tube closed at one end and open at the other, so that the boundary conditions differ, and tube open at both ends or closed at both ends, so that the boundary conditions are the same. The first sustains quarter-wavelength resonances, that is, the tube length equals integral multiples of $\lambda/4$, so the resonance frequencies are $f_n = c(2n + 1)/4L$, where $n = 0, 1, 2, 3, \ldots$, and f_n = nth resonance, c = speed of sound, and L = length of tube. The second sustains half-wavelength resonances; the tube length equals integral multiples of $\lambda/2$, so the resonances are $f_n = cn/2L$, where $n = 0, 1, 2, 3, \ldots$.

More complex tract shapes can be approximated by concatenating two or more tubes, each of uniform cross-sectional area. If the number of tubes is low, it is still relatively easy to predict the resonances of the combined system and is, thus, useful conceptually. Analytic solutions can be found for the resonances of the system by solving for the frequencies at which the sum of the admittances at any junction is zero. This was first shown by Flanagan (1972) for a set of two-tube systems approximating vowels. For more than two tubes, it is still possible, but the calculations become so complex that it is preferable to use many more tubes, simulate them as a digital filter, and calculate the resonances by computer. However, if the area changes by a factor of six or more between sections, for instance, with a constricted region between two larger-area sections, one can assume that the cavities are decoupled and compute the resonances for each tube – in this case, three – separately. In this situation, each resonance of the system will have a strong cavity affiliation, with its frequency inversely proportional to the length of that cavity. In other cases, where the area does not change so significantly between sections, the resonances are coupled. An extreme case of a coupled resonance is the Helmholtz resonance, which depends on the interaction of a small-area neck and a large-volume cavity.

All of these resonances result from plane-wave modes of propagation, meaning that the acoustic wavefronts are planar, perpendicular to the duct's longitudinal axis. A point source in the duct will radiate sound in all directions, but below a certain cut-on frequency any sound traveling in directions other than along the duct's axis will die out; these waves are evanescent. The cut-on frequency depends on the cross-dimensions of the duct and its cross-dimensional shape. It is easiest to understand for a duct that has a rectangular cross-dimension, say, L_x by L_y; the cut-on frequency occurs where a half-wavelength fits the larger of L_x and L_y, which we shall call L_{max}. In other words, $f_{co} = c/2L_{max}$. For a duct of circular cross-section, with radius a, $f_{co} = 1.841c/(2a\pi)$.

Above the cut-on frequency, cross-modes will propagate. These modes are also dispersive, meaning that higher frequencies travel faster (Pierce, 1981). Many of the assumptions underlying the basic model used in speech become progressively less true.

For vocal-tract-sized cross-dimensions, what are the cut-on frequencies? If the duct is rectangular, with $L_{max} = 2.5$ cm, $f_{co} = 7.2$ kHz; $L_{max} = 4.0$ cm gives $f_{co} = 4.41$ kHz.

If the duct is circular, a diameter $2a = 2.5$ cm gives $f_{co} = 8.42$ kHz; $2a = 4.0$ cm gives $f_{co} = 5.26$ kHz. The maximum cross-sectional areas in these cases are, respectively, 6.2 and 16 cm^2 for rectangular duct, and 4.9 and 12.6 cm^2 for the circular duct. (We use $c = 35{,}900$ cm/s as the speed of sound at body temperature, 37 °C, and for completely saturated air.) Obviously the vocal tract is never precisely rectangular or circular in cross-section. But in comparing to Fant's data, for instance (1970), we can estimate that the cut-off frequencies for the six vowels of his subject ranged from 4.6 to 9.0 kHz (assuming a rectangular cross-section) or 4.8 to 9.3 kHz (assuming circular cross-section). For a smaller subject, and where cross-dimensions are given (Beautemps *et al.*, 1995), the largest cross-dimension in the front cavity is 1.79 cm (for /a/), giving $f_{co} = 10.0$ to 11.8 kHz; the largest back-cavity cross-dimension is 2.4 cm (for /i/), giving $f_{co} = 7.5$ to 8.8 kHz. For formant estimation for vowels, then, the lumped-parameter models considering only plane-wave propagation are based on reasonable assumptions. For fricatives, there may well be significant energy above the cut-off frequency, where these models become increasingly inaccurate, but in the absence of articulatory data good enough to support more complex high-frequency models, plane-wave propagation models are often pressed into service.

There are several sources of loss in the vocal tract that have the effect of altering resonance frequencies and bandwidths. The most significant is radiation loss, especially occurring at the lip opening, but also present to a lesser extent wherever a section with small cross-sectional area exits into a region of much larger area. The main effect is to tilt the spectrum up at high frequencies. If resonances have been computed assuming no loss, their predicted frequencies will be higher than actually occur, and the difference is bigger at higher frequencies. The larger the area of

the mouth opening relative to the front-cavity volume, the greater the radiation loss. If there is a small constriction such that front and back cavities are decoupled, back-cavity resonances will have little radiation loss and so will have sharper peaks (lower bandwidths) than the front-cavity resonances.

Viscosity describes the loss that occurs because of the friction of the air along the walls of the tract; heat conduction describes the thermal loss into the walls. Both increase when the surface area of the tract is higher relative to the cross-sectional area and increase with frequency. Though not as big sources of loss as radiation, they contribute to the increased bandwidths of higher resonances. Finally, the walls of the tract are not rigid; when modeled as yielding, the bandwidths of low-frequency resonances are predicted to increase (Rabiner and Schafer, 1978).

Any sound source excites the resonances of the vocal tract, and those resonances can be calculated, approximately or more precisely, by the methods outlined above. There may also be antiresonances, when the tract is branched and/or when the source is intermediate in the tract. The antiresonances vary according to the position and type of source; for each source possibility, a different transfer function can be computed. The transfer function is a function of frequency and is the ratio of output to input. Thus, multiplying the transfer function for a particular source by the source's spectral characteristic yields the predicted output spectrum. At frequencies where the transfer function equals zero, the output will be zero no matter what input is applied; at frequencies where the transfer function has a high amplitude, any energy in the input at that frequency will appear in the output, scaled by the amplitude of the transfer function.

It is worth remembering that the resonances and antiresonances are properties of the actual air in the tract, duct, tube system. Poles and zeros are attributes of the transfer function, where the analytical expression goes to infinity (at a root of the denominator) or to zero (at a root of the numerator). A spectrum of actual speech is best described as having peaks and troughs; according to the particular set of approximations used, these may be modeled as corresponding to poles and zeros. A given spectral peak may be produced by more than one resonance, modeled by more than one set of poles; a pole-zero pair near each other in frequency may effectively cancel, producing neither peak nor trough.

Methods of Classification

Vowels

Peterson and Barney (1952), in their classic study, determined the range of variation in the first two formants for 10 vowels, thus demonstrating not only the usefulness of those two parameters but also their average values for men, women, and children. Although they measured formants from spectral slices, having determined the best place to compute the slice from a spectrogram, that is only one of several techniques available now. One can locate the vowel using only the time waveform and compute the LPC spectral envelope and determine the frequencies of the peaks in that envelope. One can run an LPC-based formant tracker on the entire utterance, which computes the peak frequencies directly. Since LPC can occasionally fail to identify closely spaced formants separately, as a safeguard one can compute either a single DFT or the entire spectrogram, respectively, and superimpose the LPC spectral envelope or formant tracks on top for a quick visual check of the LPC performance. The window used for either DFT or LPC analysis should be at least as long as a single pitch period; the LPC order should be chosen to allow for the expected number of formants within the frequency range, or adjusted and recomputed after an initial analysis.

To understand formant patterns, it is useful to consider vocal tract shapes as departures from a uniform tube that is closed at one end, the glottis, and open at the lips. For a length of 17.5 cm, assuming no losses, resonances are predicted at 500, 1500, 2500, ... Hz. Shortening the uniform tube raises all frequencies. Decreasing the area at the lip end only, akin to rounding, lowers all frequencies and reduces the bandwidths. To consider vowels other than schwa, we need at least a two-tube model. If the tongue is high, the pharyngeal area becomes large, and the oral cavity area becomes small. The lowest formant is best modeled as a Helmholtz resonance and moves down from 500 Hz; upper formants shift, depending mainly on the lengths of the two cavities, and partly depending on the area ratio. The rule of thumb, often quoted, is that increasing tongue height brings down F1, and increasing tongue frontness brings up F2. This rule works roughly, even though /u/ cannot really be modeled by a two-tube combination. The extreme, cardinal vowels /i/ and /a/ do fit. For /i/, the tongue is high and front, F1 is low, and F2 is high. For /a/, the tongue is low and back, F1 is high, and F2 is low. For these vowels with large area differences from pharynx to oral cavity, the tubes can be treated as decoupled, leading to the observation that cavity affiliation of each formant occurs in a different order. For /i/, F1 is a Helmholtz resonance, and F2, F3, and F4 are, respectively, back, front, and back-cavity resonances. For /a/, F1 to F4 are, respectively, back, front, back, front-cavity resonances. This means that, in a transition from one to the other, as occurs in the diphthong

/aI/, the formants do not smoothly change frequencies from one vowel to the other.

These models help us to understand, but real speech is seldom so clean. **Figure 7** shows waveforms and spectrograms of two sentences, one spoken by an adult female (**Figure 7A**), "Don't feed that vicious hawk," and one by an adult male (**Figure 7B**), "You should be there on time," both British speakers. We will be referring to these spectrograms throughout this section. Note that the vowel in 'You' at the start of **Figure 7B** has a low F1, but high F2 inconsistent with /u/; /ju/ has apparently been realized as [jI]. The vowel in 'should' is very short, but still has three steady formants visible. The vowel in 'be,' after the initial formant transitions, has a classic pattern for /i/; note the differences between this [i] and that in **Figure 7A**, 'feed.' The words 'there on' show a fairly gradual lowering of F3 for [r], followed by a more sudden lowering of F2 for [a]. The formants in 'time' do change from F1 and F2, from being near each other to a wider separation, as expected, but F2 does not rise very far.

The simple models also allow one to understand how vowels vary with gender and age. As children grow, their pharynxes lengthen more than do their oral cavities; the vocal tract length differences between adult men and women are due more to differences in pharyngeal than in oral cavity length. Thus, the formant space does not scale uniformly by vocal tract length. The higher F2 in the female subject's [i] agrees with this explanation.

To a first approximation, the voicing source and the vocal tract filter are independent. We can therefore think of the transfer function from glottis to lips as a spectral envelope that is sampled by the fundamental and its harmonics. If the vocal tract remains the same shape, leaving the formants at the same frequencies, the harmonics sample it more coarsely at higher values of F0. On average, speakers with smaller larynges also have shorter vocal tracts, so that, as the range of F0 values possible moves up, the range of formant frequencies increases too. However, as Peterson and Barney's data show (1952), they do not increase at the same rate; women's F0 is 1.7 times higher than men's, while their formant frequencies are only 1.15 times higher, on average. This means that F0 is much more likely to approach F1 in women than in men, and formants may be difficult to resolve. An example of this occurs in **Figure 7A** in 'feed,' where F0 is 273 Hz.

Finally, sometimes the properties of the voicing source are of more interest than the filter properties of vowels. It is possible to inverse filter the speech signal and arrive at an estimate of the glottal volume velocity. In order to inverse filter, one must estimate what the filter was, invert that, and multiply it by the speech spectrum. Clearly, the estimate of the glottal volume velocity is only as good as the estimate of the tract filter function, but the technique has led to detailed explanations of voice quality differences, including source differences between men and women. One of the better-known techniques uses the Rothenberg mask to measure volume velocity at the lips and inverse filter that signal rather than the far-field pressure. This provides information about the mean flow of air through the vocal tract, including the degree of breathiness in the glottal volume velocity.

Nasals

In order to produce a nasal, the velum is lowered, and complete closure is effected in the oral cavity. The oral cavity becomes a side branch that contributes antiresonances inversely related to the length from pharynx to place of closure; the resonances arise from the pharynx and nasal cavities. The nasal cavities are convoluted in shape, uniquely so for each individual; the length of the effective tract is thus longer than that of pharynx plus oral cavity, with a correspondingly lower first formant. Bandwidths of all resonances are also larger because there are more surfaces to absorb sound.

As with fricatives, the radiated spectrum is a mixture of peaks and troughs that are not always easy to map to particular cavities. The nasal formants are packed more closely in frequency than nonnazalized vowel formants, but they may not all be apparent because of the antiresonances. Some of the series of antiresonances may appear as deep spectral troughs, but where they coincide with nasal formants, they will cancel, or nearly so, and neither will be apparent. Because of the cancellation and the wide bandwidths, the spectrum of a nasal will overall have lower amplitude than an adjacent vowel. While the antiresonances that could provide a place cue may not be strikingly apparent, particularly if there is background noise, the formant transitions in adjacent vowels will also provide place cues; briefly, all formants will decrease before a bilabial nasal, F1 and F3 will decrease and F2 will increase before a velar nasal, and F1 and F3 will decrease and F2 will decrease or increase depending on the vowel before an alveolar nasal (Kent and Read, 1992; Johnson, 2003).

The clearest example of such transitions occurs in **Figure 7A** in 'don't,' where F2 clearly rises during the nasalized portion. In **Figure 7B**, no transitions are obvious in the vowel of 'on,' though F2 and the amplitude both decrease abruptly at the start of the nasal. In 'time,' a slight F1 transition is observed, and F3 appears to drop abruptly, though it is not well-defined in the spectrogram.

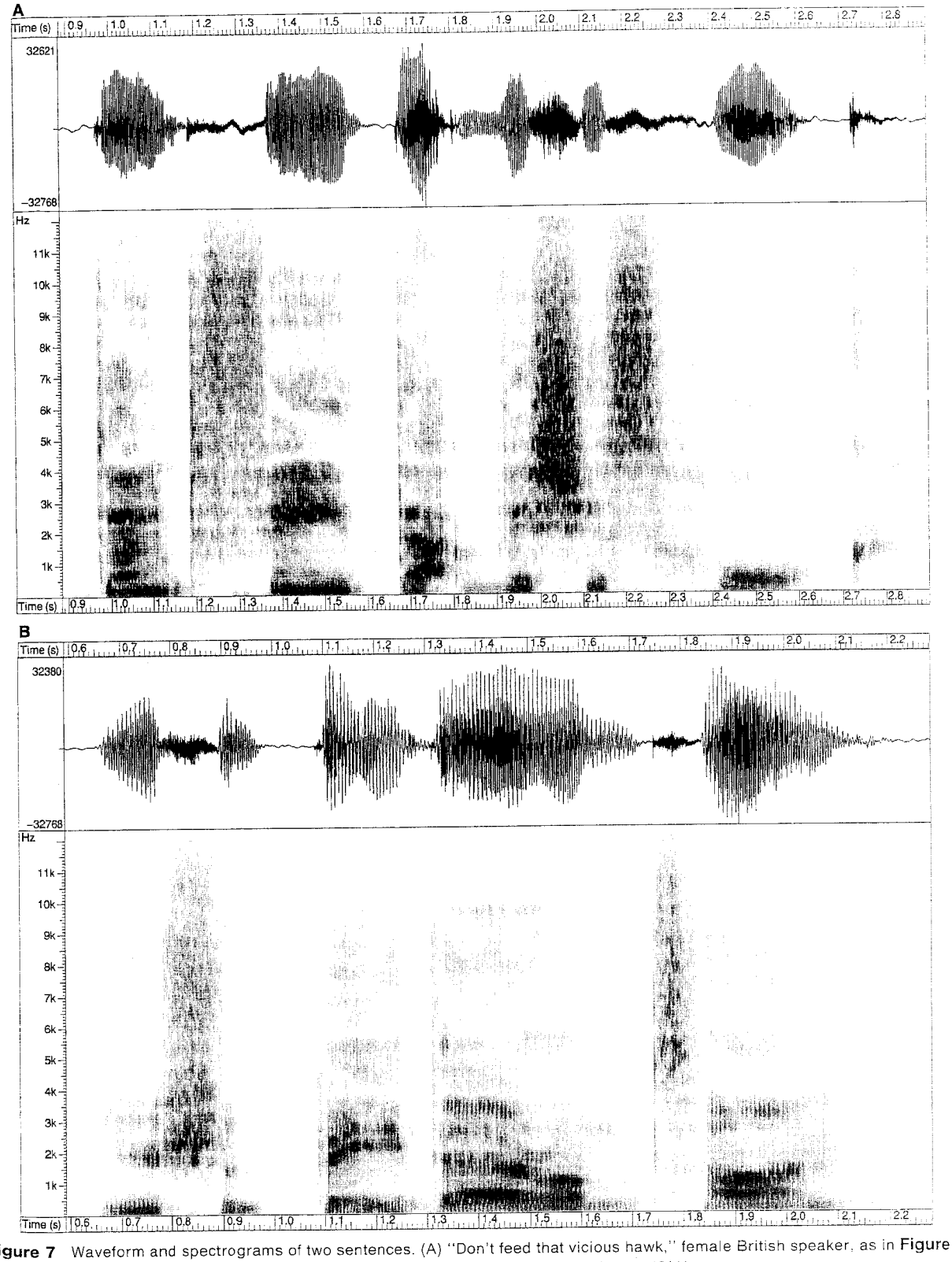

Figure 7 Waveform and spectrograms of two sentences. (A) "Don't feed that vicious hawk," female British speaker, as in **Figure 1**; (B) "You should be there on time," male British speaker. Note spectrograms extend up to 12 kHz.

In nasalized vowels the velum is down, but the oral cavity is not closed. The presence of two distinct paths still allows for interference effects, but the antiresonances will be at different frequencies than for nasals, and these frequencies will depend on the area of the velo-pharyngeal port. The resonances will correspond to those of the vowel alone (pharynx plus oral cavity) and the nasal formants (pharynx plus velo-pharyngeal port plus nasal cavities); the antiresonances may cancel some of these, or may show up as spectral troughs, but it is likely that the lowest nasal formant will be the highest-amplitude peak.

Fricatives

Many different sets of parameters for fricatives have been explored, but none are yet sufficient to classify them. Theoretically it seems straightforward; when the constriction is small, as during the 'steady-state' portion of a fricative, the back-cavity resonances are essentially cancelled. The noise source excites the front-cavity resonances, and antiresonances – zeros – appear at low frequencies and at higher frequencies inversely related to the distance between source location and constriction exit. If the source is not well localized, these higher-frequency antiresonances may smear out and not be readily apparent. The frequency at which the energy appears in the spectrum thus should differentiate fricatives by place, with longer front cavities for palatals and velars corresponding to energy at lower frequencies. However, the frequency ranges used for different fricatives overlap extensively across subjects. Further, interdentals seem to be highly variable even within subject, with, sometimes, barely discernable noise. For instance, in **Figure 7A**, [f] has significant energy from 1200 Hz to 11 kHz (and possibly higher; the antialiasing filter begins to act there), though clearly not as high amplitude as the [s] or [ʃ] in the same sentence, and lasting for 150 ms. The [v], however, appears to consist of a voicebar lasting 100 ms and weak noise, albeit at roughly the same frequency range, for only 10–20 ms. Note also that [ʃ] differs slightly in **Figure 7A** and **7B**, with the frequency of the lower edge of the high-amplitude region occurring at approximately 2.0 kHz for the female, 1.5 kHz for the male subject. This may be due to a difference in length of the front cavity, or, more likely, to the influence of the vowel context, with the higher cut-on frequency corresponding to the high unrounded vowel.

It was thought at one point that identification of interdentals depends on transitions, while that of /s, ʃ/ depends only on steady-state characteristics. An obvious difference in articulation tends to support this theory; /θ/ requires the tongue tip to be in contact with the teeth, unlike in /f/. However, careful manipulation of speech signals shows that transitions as well as steady-state characteristics are important for /s, ʃ/ (Whalen, 1991).

In the transition from a vowel to a fricative several things happen, and not always in the same order. Formants shift as the constriction becomes smaller, noise begins to be produced, and the formants as well as antiresonances begin to be excited. Back-cavity resonances can be prominent for a time until the constriction area decreases sufficiently for them to be cancelled. As the noise increases, the rate at which it increases depends on the fricative; stridents appear to have the most efficient noise sources, in that the noise produced increases at a greater rate proportional to the flow velocity through the constriction. Both spectral tilt and overall spectral amplitude are affected. Within a given place and for a given subject, the spectral tilt can be thought of as occurring in a family of curves; if the same fricative is produced with greater effort, the spectrum tends to have higher amplitudes overall and a less negative slope. Voiced fricatives with the same place will have a set of curves with a similar relationship of spectral tilt to effort that is less than, but overlapping with, the range for their voiceless versions. However, these differences, while predictable from an understanding of flow noise sources, do not sufficiently distinguish fricatives (Jesus and Shadle, 2002). Finally, voicing changes during the transition for both voiced and voiceless fricatives, presumably to allow sufficient pressure drop across the constriction to support frication.

Many researchers have pursued methods of characterizing fricative spectra by statistical moments, as if they were probability distributions. Recently Forrest *et al.* (1988) described their calculation of spectral moments, indicating that these were sufficient to distinguish stops, but applied to fricatives, distinguished /s, ʃ/ from each other and from the interdentals /f, θ/, but did not distinguish the interdentals at all. More recent studies have used methods of computing the moments that showed that certain moments of the English voiceless fricatives were statistically significantly different, but the differences were not enough to allow for categorization.

Spectral moments capture the gross distribution of energy over the chosen frequency range, but ignore particular features that we can attribute to particular production methods, such as back-cavity formants appearing in the transition regions, or the salience and frequency of spectral troughs. In addition, the gross parameters captured depend greatly on the particular spectral representation from which the moments were calculated. Ideally, a low-variance spectral estimate would be used, but this has not

typically been done. Computing the moments does some spectral smoothing as with frequency interpolation, but with more bias, and amplitude thresholding and frequency range can affect the results dramatically. One question was whether using a better spectral estimate before computing moments might improve results. It appears that starting with a good spectral estimate helps, but only marginally; new parameters are needed for significant gains (Blacklock, 2004).

The best parameters appear to be based on multitaper spectra, with frequency range to 20 kHz, amplitude threshold carefully controlled, and identical recording conditions across subjects, as shown in Figures 4 and 8. Figure 8 shows examples of multitaper spectrograms of the voiceless fricatives uttered in the same vowel context, by the same subject. While differences between the fricatives are apparent in these examples, the problem is to find characteristics that hold up across tokens, contexts, and subjects. Men and women require different parameters; an examination of the variance of the mean power spectral density in 12 subjects indicated that /s, ʃ/ can best be distinguished from each other at 2.5 kHz for men, 3.0 kHz for women. The main spectral peak in /f/ occurred at 2, 4, or 7 kHz for men, but most often at 2 kHz; for women the peak occurred at 2, 4, or 8 kHz, and was more evenly distributed among these frequencies. Spectral variation within particular tokens was also examined, with somewhat inconclusive results. Clearly multitaper analysis is a powerful tool that bears further investigation (Blacklock, 2004).

Finally, voiced fricatives often devoice, with the amount somewhat dependent on language (studies on English, French, and Portuguese are cited in Jesus and Shadle, 2002) as well as with fricative place (posterior fricatives devoice more often) and position within the phrase (end of sentence devoices more often). Devoicing allows more air pressure to be dropped across the supraglottal constriction, thus strengthening the noise source. However, it appears that in some cases the fricative 'denoises' instead,

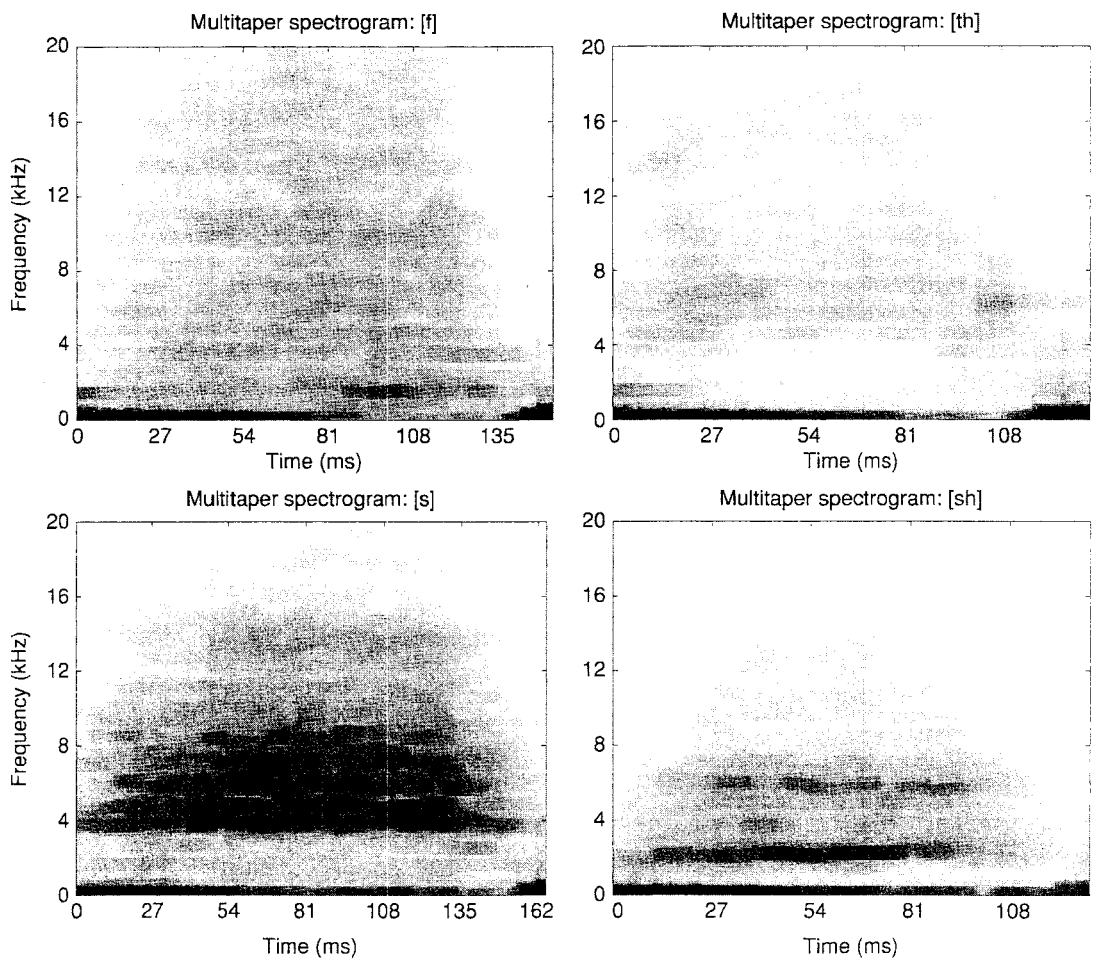

Figure 8 Multitaper spectrograms of [f] from 'buffoon,' [θ] from 'Methuselah,' [s] from 'bassoon,' and [ʃ] from 'cashew,' same British male speaker as in **Figure 4**. (After Blacklock, 2004.)

with additional pressure drop being used across the glottis, strengthening the voicing source. Voiced fricatives are shorter in duration than their voiceless equivalents in all languages studied. The modulation of the noise source by the voicing source indicates that the phase of the modulation changes rapidly in the transition into and out of the fricative (Jackson and Shadle, 2000). This may be a feature that humans notice and use in identification; further studies await.

Stops and Affricates

Stops are a relatively well-understood class. The manner in which they are articulated is related to the temporal events that are observable in the time waveform; the place at which they are articulated is related mainly to spectral cues. Before the stop begins, articulators are moving toward closure; if the stop occurs postvocalically, formant transitions will occur that offer place cues. For the stop itself, first is the period of closure, during which no air exits the vocal tract; voicing may continue briefly but no other sounds are produced. When closure is released, there may be the release burst, followed by brief frication as the articulators move apart, followed by aspiration and, finally, by voice onset. After voice onset the formants are more strongly excited, and transitions characteristic of the stop's place will again be observable.

Not all of these stages occur with every stop. If the stop is preceded by /s/, it has a closure period but no burst release. Syllable-final stops are often not released. The frication period is not always present and distinguishable from aspiration. Both frication and aspiration may be missing in voiced stops; they tend to be present in voiceless stops, but formant transitions are less obvious in the vowel occurring after the stop.

These latter two points are related to one of the stronger cues to voicing of a stop, the voice onset time (VOT). The VOT is the time between stop release and voice onset. In voiced stops, although voicing may well cease during closure as the pressure builds up in the vocal tract, the vocal folds remain adducted; when the supraglottal pressure suddenly drops following release, phonation begins again quickly, leading to a short VOT. In voiceless stops, the vocal folds are abducted and take time to be adducted for the following voiced segment, leading to a long VOT. Aspiration noise is produced near the glottis because the glottis, while narrowing, provides a constriction small enough to generate turbulence noise.

Experiments in which the VOT has been varied in synthetic stimuli have shown that VOT alone produces a categorical discrimination between voiced and voiceless stops, with a threshold value of 20–30 ms. However, VOT varies to a smaller extent by place, with velar stops having longer VOT than bilabial stops; this difference is as much as 20 ms. Finally, VOT varies with speech rate, with values shortening at higher rates.

The main spectral cues in stops are the burst spectral shape and the formant transitions in adjacent vowels. Additional cues lie in the spectral shape of the frication interval, but this is so brief, relatively weak, and time-varying that it is much less easy to analyze. The spectral shape of all three is related to the movement of the articulators toward closure for the stop. It can be shown that any narrowing in the anterior half of the vocal tract will cause the first formant to drop in frequency. The direction of frequency change in F2 and F3 depends on the place of the target constriction (of the stop) and the position of the tongue before the movement began (the vowel front- or backness). As demonstrated initially by Delattre *et al.* (1955) and cited in numerous references since, for bilabial stops all formants decrease in frequency when moving toward the stop (i.e., whether observing formant transitions pre- or poststop); a clear example of this is seen for 'be' in **Figure 7B**. For velar stops, F1 and F3 decrease; F2 increases when moving toward the stop. For alveolar stops, F1 and F3 decrease; F2 increases for back vowels and decreases for front vowels. But note that in **Figure 7A**, the vowel formants in 'hawk' do not change noticeably near the closure.

The burst spectra follow related patterns, since they are produced by an impulse excitation of the vocal tract just after closure is released. For bilabials, the spectrum has its highest amplitude at low frequencies and falls off with frequency. Alveolars are high amplitude at 3–5 kHz, and velar bursts are highest amplitude at 1–3 kHz. Though these are referred to, respectively, as having shapes of falling, rising, and indeterminate or compact or midfrequency, these terms are relative to a frequency range of 0 to, at most, 5 kHz. The [t] in 'time' in **Figure 7B** shows a striking burst, frication, aspiration sequence, which extends up to 12 kHz. The theoretical burst spectral shapes are roughly similar to those of fricatives at each place, as we would expect, since all back-cavity resonances should be cancelled immediately postrelease, and the front-cavity resonances are excited.

Affricates can be thought of as a combination of a stop and a fricative, but with some important differences in timing and place from either. The closure and release of a stop are evident, but the frication period is long for a stop and short for a fricative. Aerodynamic data indicate that the constriction opens more slowly for /tʃ/ than for /t/, directly supporting the longer frication duration for the affricate compared to the

stop (Mair, 1994). The rise time for the frication noise for /tʃ/ is significantly shorter than for /ʃ/ (Howell and Rosen, 1983).

Conclusion

We have surveyed some aspects of acoustics, recording equipment, and techniques, so that appropriate choices can be made. It is possible to compare speech analysis results using recordings that were not made in the same way, provided that information such as type of microphone and its position relative to the speaker have been noted, ambient noise has been recorded, and so on.

By the same token, signal processing principles and techniques have been reviewed so that the techniques can be chosen appropriately for both the signal type (whether periodic, noisy, or a combination) and the information sought (absolute level, formant frequencies, properties of the voice source, etc.). Some parameters must be estimated and the analysis done twice or more, iterating. Others must be done correctly the first time, such as antialiasing before sampling a signal. Each of the different methods of spectral analysis has its place; the choice of which is best depends not only on the type of speech sound being studied, but also on the speaker.

Finally, the basic manner classes of speech have been reviewed and parameters that can be used for classification discussed.

See also: Phonetics, Articulatory; Voice Quality.

Bibliography

Beautemps D, Badin P & Laboissiere R (1995). 'Deriving vocal-tract area functions from midsagittal profiles and formant frequencies: a new model for vowels and fricative consonants based on experimental data.' *Speech Communication 16,* 27–47.

Bendat J S & Piersol A G (2000). *Random data: analysis and measurement procedures* (3rd edn.). New York: John Wiley and Sons, Inc.

Beranek L (1954). *Acoustics.* New York: McGraw-Hill Book Co. Reprinted (1986). New York: Acoustical Society of America/American Institute of Physics.

Blacklock O (2004). *Characteristics of variation in production of normal and disordered fricatives, using reduced-variance spectral methods.* Ph.D. thesis, School of Electronics and Computer Science. UK: University of Southampton.

Catford J C (1977). *Fundamental problems in phonetics.* Bloomington, IN: Indiana University Press.

Crystal D (1991). *A dictionary of linguistics and phonetics* (3rd edn.). Oxford: Blackwell Publishers Inc.

Delattre P C, Liberman A M & Cooper F S (1955). 'Acoustic loci and transitional cues for consonants.' *Journal of the Acoustical Society of America 27,* 769–773.

Fant C G M (1962). 'Sound spectrography.' *Proceedings of the 4th International Congress of Phonetic Sciences.* The Hague: Mouton. 14–33. Reprinted in Baken R J & Daniloff R G (eds.) *Readings in clinical spectrography of speech.* San Diego, CA: Singular Publishing Group and Pine Brook. NJ: Kay Elemetrics Corp.

Fant G (1970). *Acoustic theory of speech production.* The Hague: Mouton.

Flanagan J L (1972). *Speech analysis synthesis and perception.* 2nd edn. New York: Springer Verlag.

Flanagan J L & Cherry L (1969). 'Excitation of vocal tract synthesizers.' *Journal of the Acoustical Society of America 45,* 764–769.

Forrest K, Weismer G, Milenkovic P & Dougall R N (1988). 'Statistical analysis of word – initial voiceless obstruents: preliminary data.' *Journal of the Acoustical Society of America 84(1),* 115–123.

Gold B & Morgan N (2000). *Speech and audio signal processing.* New York: John Wiley & Sons, Inc.

Howell P & Rosen S (1983). 'Production and perception of rise time in the voiceless affricate/fricative distinction.' *Journal of the Acoustical Society of America 93,* 976–984.

Jackson P J B & Shadle C H (2000). 'Frication noise modulated by voicing, as revealed by pitch-scaled decomposition.' *Journal of the Acoustical Society America 108(4),* 1421–1434.

Jackson P J B & Shadle C H (2001). 'Pitch-scaled estimation of simultaneous voiced and turbulence-noise components in speech.' *IEEE Transactions on Speech and Audio Processing, 9(7),* 713–726.

Jesus L M T & Shadle C H (2002). 'A parametric study of the spectral characteristics of European Portuguese fricatives.' *Journal of Phonetics 30,* 437–464.

Johnson K (2003). *Acoustic and auditory phonetics* (2nd edn.). Oxford: Blackwell Publishers.

Kent R D & Read C (1992). *The acoustic analysis of speech.* San Diego: Singular Publishing Group.

Ladefoged P (2001). *Vowels and consonants.* Oxford: Blackwell Publishing.

Mair S (1994). *Analysis and modelling of English /t/ and /tsh/ in VCV sequences.* Ph.D. thesis, Dept. of Linguistics and Phonetics. UK: University of Leeds.

McClellan J H, Schafer R W & Yoder M A (1998). *DSP first: A multimedia approach.* Upper Saddle River, NJ: Prentice Hall.

Olive J P, Greenwood A & Coleman J (1993). *Acoustics of American English speech: a dynamic approach.* New York: Springer-Verlag.

Peterson G E & Barney H L (1952). 'Control methods used in a study of the vowels.' *Journal of the Acoustical Society of America 24,* 175–184.

Pierce A D (1981). *Acoustics.* New York: McGraw-Hill Book Co.

Rabiner L R & Schafer R W (1978). *Digital processing of speech signals.* Englewood Cliffs, NJ: Prentice-Hall, Inc.

Dordrecht: Kluwer Academic Publishers.
Stevens K N (1971). 'Airflow and turbulence noise for fricative and stop consonants: static considerations.' *Journal of the Acoustical Society of America 50,* 1180–1192.

Speech.
Whalen D H (1991). 'Perception of the English /s/-/ʃ/ distinction relies on fricative noises and transitions, not on brief spectral slices.' *Journal of the Acoustical Society of America 90(4:1),* 1776–1785.