Keller, E. (1994). Fundamentals of phonetic science. In E. Keller (ed.), *Fundamentals of Speech Synthesis and Speech Recognition: Basic Concepts, State of the Art, and Future Challenges* (pp. 5-21). Chichester: John Wiley.

Fundamentals of Phonetic Science 1

Eric Keller

Laboratoire d'analyse informatique de la parole (LAIP)
Université de Lausanne, CH-1015 LAUSANNE, Switzerland

The chapter presents basic information about the human production of speech sounds and their manifestations in the speech signal. At the laryngeal level, the vocal cords provide the acoustic vibrations required for voiced and vowel sounds. At the supra-laryngeal level, resonance, plosion and frication in the various speech cavities furnishes distinctive sound quality for the various speech sounds. The relations between production and acoustic representation are examined in some detail for different types of speech sounds (vowels, fricatives, plosives). Finally, a short review of spectral analysis techniques is given.

The purpose of this initial chapter is to prepare the uninitiated reader for the subsequent chapters of this volume by providing a succinct review of how speech sounds are produced and how they appear in the acoustic waveform. A number of major principles will be presented, and the application of each of these principles to its related acoustic manifestation will be illustrated. The present chapter deals with segmental, and the next chapter with suprasegmental aspects of speech.

The Communication Process: Transmission Despite a Noisy Line

Speech is largely used for the purpose of *communication*, i.e., for the transmission of information from person to person. A communication succeeds, if information considered to be important by both speaker and hearer is understood essentially as intended. Speech need not be perfect

— i.e., need not be exactly reproducible — neither in terms of its production, nor in terms of its perception. It suffices that certain crucial *distinctions* be transmitted correctly. For example, the answer to the question “Did you get the sandwich for me?” can be “yes”, “uh-huh”, “sure” or any number of other “yes-or-no” responses. For the communication to succeed, all the questioner needs to know is whether the sense of the answer is yes or no.

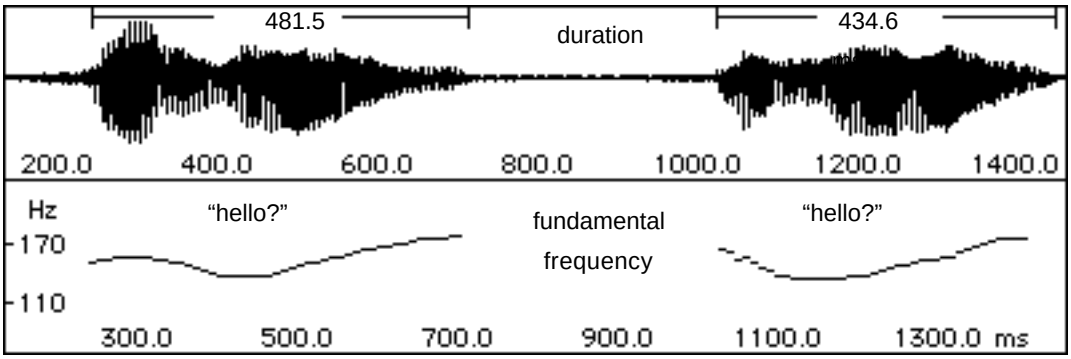

Figure 1. The acoustic signal of the word “hello”, produced twice by the same speaker. Above, signal; below, voice fundamental frequency. It can be seen that duration and fundamental frequency (among other parameters) vary considerably from the first to the second production of the word.

As a consequence, the speech signal is generally quite “noisy”, that is, one representation of a given signal can be — and usually will be — quite different from the next. Also, the signal is not necessarily well-defined with respect to background noise. For example, the top illustration of Figure 1 shows the signal of the word “hello”, pronounced twice in a row by the same speaker. The bottom panel shows the fundamental frequency for the signal. The two signals differ with respect to all three aspects shown here: overall duration, amplitude development, and fundamental frequency. They probably differ also with respect to a number of other physical parameters that are not shown here.

Not all parameters vary to the same degree. For example, the data shown in figure 2 illustrate that durations of word-final syllables and their adjoining pauses are less variable than the fundamental frequency patterns measured in the same word-final syllables. Still, some variation is inevitable, even in the most regular of speech parameters (for more on variability in speech, see Lisker, 1985; Stevens, 1989).

This variation is not simply a matter of “sloppy speech habits”. Even when trained speakers attempt to produce the same utterance repeatedly under laboratory conditions, neural and physiological factors render an exact reproduction impossible. The same is true, of course, when different speakers repeat the same utterance. Patterns also diverge as a result of dialectal and social differences, and as a consequence of age and personal characteristics. In other words, two utterances not only *needn’t* be carbon copies of each other, they simply *can’t*.

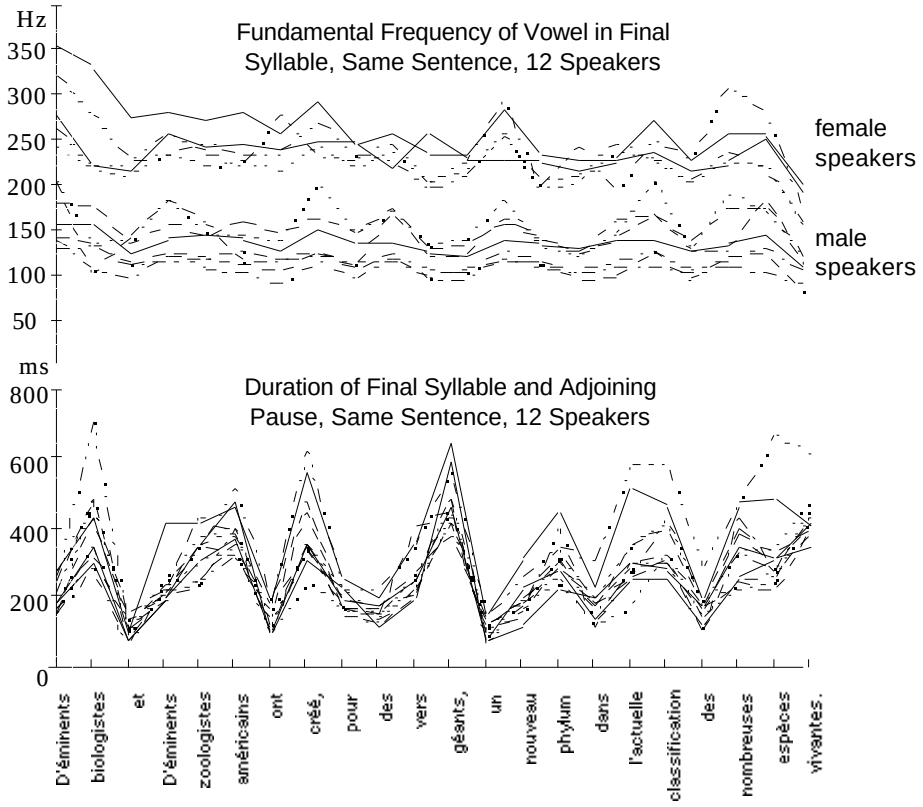

Figure 2. Variability in the frequency and durational domains. The fundamental frequency (Fo) of the final vowel (top panel) varies more than the duration of the corresponding final syllable + pause segment (bottom panel). Fo varies considerably, not only from males to females, but also from speaker to speaker. Measurements performed by Caelen-Haumont, 1991. The sentence translates as "Eminent American biologists and zoologists have created a new phylum for giant worms in the current classification of the numerous living species".

The consequences of this "imperfection" are different for synthesis and recognition. If human speech is never produced in exactly the same way, synthesised speech need not conform 100% to a specific speech model either, though it might help if variations occurring in synthesis resemble those of human speech. Speech recognition devices, on the other hand, must be particularly sophisticated, since they must pick out relevant, but highly variable information from a mass of irrelevant noise. It is an essential tenet of this book that both synthesis and recognition are likely to benefit from closer approximations to human communicative principles. Of particular interest are principles that govern the way that human communicative material or channels can be deficient, or variable, without failing in their central purpose of transmitting information.

Communication Despite Intra- and Inter-Speaker Variability

What are the origins of this variability? To go back to the question "Did you get the sandwich for me?", how does the listener know that the answer is part of the "yes"-class, not part of the "no-" or the "didn't-understand / hear-your-question" class? The answer is rather complex and involves picking out the target signal from background noise and competing speech signals, distinguishing certain sound patterns from other, similar sound patterns, matching the input to the right pre-stored words, and making a host of grammatical and semantic judgements in accordance with learned linguistic patterns[1]. But an excellent place to start the discussion is at the "speech-sound", or the "phonetic", level.

It was said above that transmission depends less on a narrowly defined set of physical parameters, than on a *distinctive* set of parameters. That is, physical parameters need not have specific values, but they must be part of *mutually exclusive envelopes of permissible values*. At the acoustic level, this means that two different speech sounds must belong to distinctive clusters. Example: the vowel /æ/ in "pat" differs only slightly from the vowel /ɛ/ in "pet", but most instances of /æ/ form a cluster distinctive from the cluster formed by the various instances of /ɛ/[2].

The clusters' shapes — and their deformations — are largely determined by the mechanics and the acoustics of the human speech tract. Two speech sounds generally belong to the same cluster, if their production and the resulting acoustic representation are similar[3]. To

---

[1] For more details on some of these issues, see chapters 14 and 15 of this volume.

[2] Throughout this text, three types of transcription are used: Sound symbols ("phonemes") between slashes (e.g. /e/, /s/), "allophones" or "phones" between brackets (e.g. the "back" [ɑ] and the "front" [a], or [r] and [R]), and letters ("graphemes") between quotation marks (e.g. "e", "s", etc). Phonemes, allophones and phones have to do with acoustic sounds, while graphemes are letters as they appear on the written page. More specifically, phonemes are *distinctive* sounds, "allophones" are *non-distinctive* sounds, and "phones" are acoustically coherent *portions* of sounds. Distinctiveness is established by the minimal-pair test: Two sounds are distinctive if they distinguish two words of a language. E.g. in English, /t/ and /d/ are distinctive, since they distinguish, among others, the words "tin" and "din". However, the "back" [ɑ] and the "front" [a] are non-distinctive, since they only distinguish two dialectal variants of the English /a/-sound. The two types of /a/ are thus called "allophones of the phoneme /a/". *Phones* refer to segments of speech sounds that show a certain internal coherence. Diphthongs like /aj/ in "like" or /aw/ in "house", for example, can be seen as being made up of two phones each, [a] and [i], or [a] and [u].

[3] Though sometimes only the acoustic representation is similar, as the rolled [r] and the fricated [R] of French, which belong to the same distinctive *acoustic* envelope, but are *produced* in two entirely different ways.

explain how, we shall now turn to a capsule summary of speech articulation, and its relationship to the associated acoustic waveforms.

A Capsule Summary of Speech Articulation

The Various Ports

Figure 3 provides the overall view of the speech production process. To illustrate the sounds produced in the speech tract, we shall refer to the sounds of English. This will do for an introductory discussion, but it is understood that the production of the sounds of other languages can be quite different from that which is sketched here. Although various languages employ the speech apparatus in essentially similar fashion, some sounds of non-English languages (like clicks and trills, for example) involve uses of the apparatus which are not described here.

The speech production process is initiated by a release of air from the lungs into the vocal tract. At the larynx, the air is either simply *passed through,* or it is *set into vibration* by a rapid, repeated closing action of the vocal cords (see figure 4). In the first case, the resulting sounds become "unvoiced" (typically consonants, like /s/, /t/, or /f/), in the second case, they become "voiced" (vowels or "voiced consonants", such as /m/, /z/, or /g/). The difference between the two types of sounds is quite evident in the resulting acoustic signal (Figure 5).

Above the larynx, there are four major areas ("ports" or "valves") where constrictions in the vocal tract can occur. Each of these areas is associated with a typical set of sounds that owe a major component of their acoustic characteristics to the location of the port.

The *velar port* controls access to the nasal cavity. When it is open, air can escape via this cavity to produce the "nasal consonants" (e.g. /m/ and /n/) and the "nasal vowels" (as in French "fr*an*c", "c*in*q", or in pre-nasal vowels, such as [Ë$in "noon"). When the velar port is closed, air is forced to escape via the oral cavity, giving rise to all remaining sounds.

The next port is the *linguo-palatal port* (Figure 3). It is closed during the production of /k/, /g/ and /ŋ/ "ng". The tongue is a bit removed from the palate, and the port is somewhat opened for the production of /u/ and /o/ vowels, but it is entirely open for the back /a/ (the sound phoneticians write as "closed a" [ɑ], the vowel in the midwestern American pronunciation of "car"). This series of vowels ranging from /u/ to /a/ via /o/ is known as the *back vowel* series.

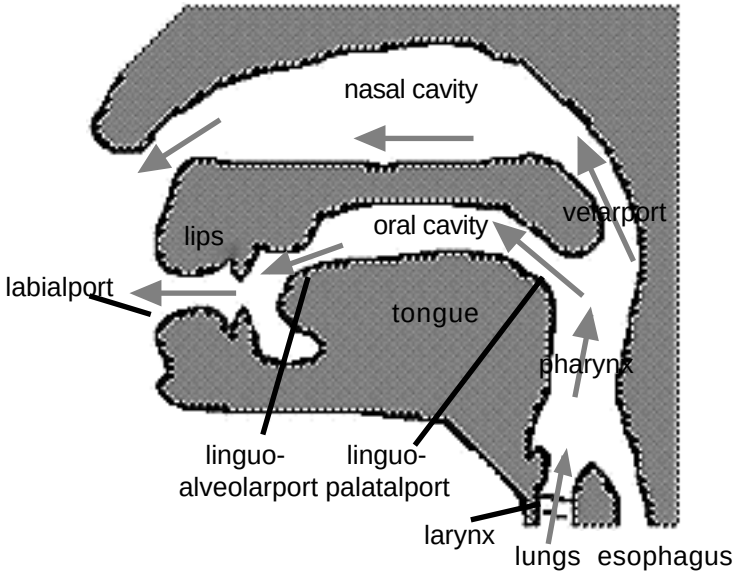

Figure 3. The human vocal tract and the basic speech production process. The air flow generated by the lungs propagates through the larynx, the pharynx and the upper vocal tract. *Oral sounds* (most vowels and consonants) are produced by modifications of the oral cavity. *Nasal sounds* (nasal consonants like /n/ and /m/, and nasal vowels) are produced with air passing through the velar port and the nasal cavity.

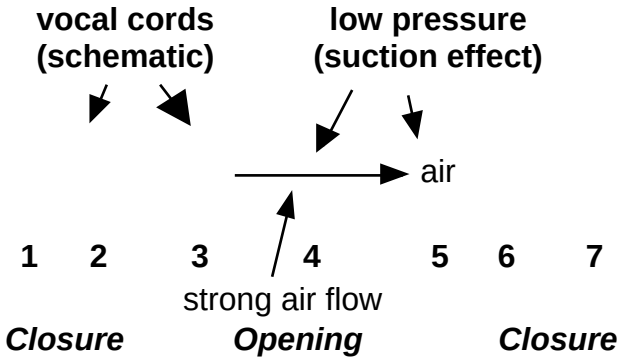

Figure 4. The schematised action of the vocal cords in voiced sounds. During closure, vocal cords force an air pressure build-up. During the opening phase, air escapes rapidly, setting up a suction effect, which leads to a tight closure of the vocal cords.

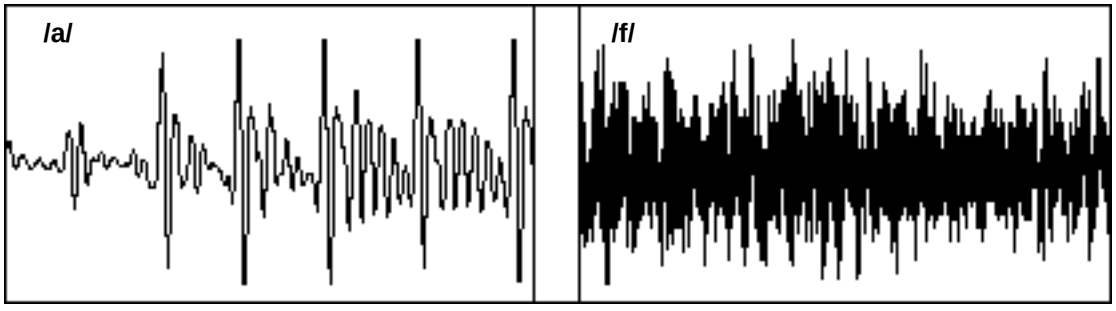

Figure 5. Voiced sounds (left) are easily recognised by their regular fundamental waveform which is absent in unvoiced sounds (right).

The *front vowels* are associated with the next port, the *linguo-alveolar port* (in some languages, the *linguo-dental port*). English vowels distinguished primarily by lingual opening at this location are /i/, /e/, /æ/, and front /a/ (the "open a" [a], the sound that corresponds to a Bostonian or Southern English pronunciation of the vowel in "car"). The linguo-alveolar port is also the location associated with the production of a great number of consonant sounds, such as the so-called *stops* (or *plosives*) /t/ and /d/, the fricatives /s/, /z/, /ʃ/ ("sh"), and /ʒ/ (the sound in "measure").

The *labial port*, finally, is centrally concerned in sounds like /p/, /b/, /f/ and /v/, where the lip opening is a major contributor to their production.

The acoustic function of these ports is to subdivide the vocal tract into various inter-connecting *resonating chambers*. Given the dimensions and wall characteristics of these chambers, acousticians can calculate reasonable approximations of the frequencies and amplitudes of the resonances that are generated in these chambers, and which are superimposed on the fundamental frequency vibration imparted by the vocal cords. In the spectrogram, these resonances are visible as resonance bands, or "formants" (see Figure 6, for example).

It can be appreciated that, given a strong theory on how cavity sizes relate to resonance frequencies, the approximate configuration of the vocal tract can be reconstructed (Fant, 1960, and successors; see also Stevens, 1989). In this way, likely distinctive speech sounds can be deduced from the speech signal, at least in those cases where articulations were clearly executed and ample acoustic energy was available.

The Different Speech Production Modes

In addition to the influence of port location, a second major acoustic component derives from the sounds' various *modes of production*. These modes determine the commonly-used classification of speech sounds into vowels, diphthongs, semi-vowels, stops, fricatives, etc.

Vowels and *diphthongs* are sounds that are produced with vibrating vocal chords and a minimum of obstruction in the upper vocal tract[4]. While both types of sounds are produced with a vocal tract that changes over time, changes are relatively small for vowels which are perceived as single sounds, while such changes are fairly extensive in the case of

[4] For a detailed description of vowel in articulatory and acoustic space, please see the chapter by Boë, Schwartz and Vallée in this volume.

diphthongs, which are perceived as two vowels in sequence. In the acoustic signal, vowels and diphthongs usually show stronger amplitude than consonants, and when viewed in detail, they show a characteristic fundamental frequency pattern (Figure 5, left panel). Detailed studies of *nasal vowels* in languages such as French have documented the frequent presence of supplementary nasal resonance bands arising in the nasal cavity (not shown, O'Shaughnessy, 1982).

Semi-vowels or *semi-consonants* (like /w/ and /j/) also show strong changes over time (that is why they are sometimes called "glides"). They either lead into vowels (e.g. /*w*æn/ "wan" or /*j*æk/ "yak"), or they terminate vowels (as in /paw/ "pow", /paj/ "pie", where they are part of diphthongs). These sounds are best thought of as transitions. The [j] or [w] phases of such transitions are produced with narrower port constrictions than the adjoining vocalic phases. Acoustically, the transition between the two phases is primarily reflected in the resonance patterns seen in spectrographic representations[5]. For example, the transition is prominently reflected in the time course of formants 1 and 2 (Figure 6).

Fricatives are sounds produced with a close approximation between two articulators, that is, with strong occlusions of the port in question. As a result, air turbulences are created that are seen as high-frequency noise in the signal (Figure 5, right panel).

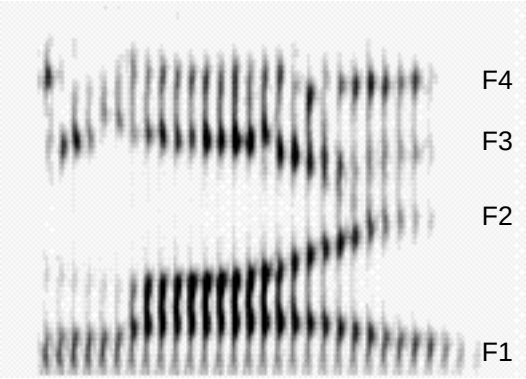

Figure 6. Spectrogram of the diphthong /aj/ in English. Four formants can easily be identified. The transition from [a] to [j] is associated with an increase in the gap between formants 1 and 2 (F1 and F2).

Stops or *plosives* are complex sounds that are made up of two distinct initial phases, followed by a transition phase (Figure 7). Initially, while the port is closed, the signal shows silence. At the opening of the port, there is usually a fairly sharp release of the air pressure that has built up in the

[5] A spectrogram (or "sonogram") represents the three physical dimensions of speech sounds in the x, y and z axes. The x axis shows time, the y axis shows frequency, and the z axis (typically captured as levels of darkness) shows the amplitude or energy level.

oral tract during the closure phase. This gives rise to a so-called "burst" and to a short period of frication. In the last stage, the frication merges into the characteristics of the succeeding sound (often a vowel).

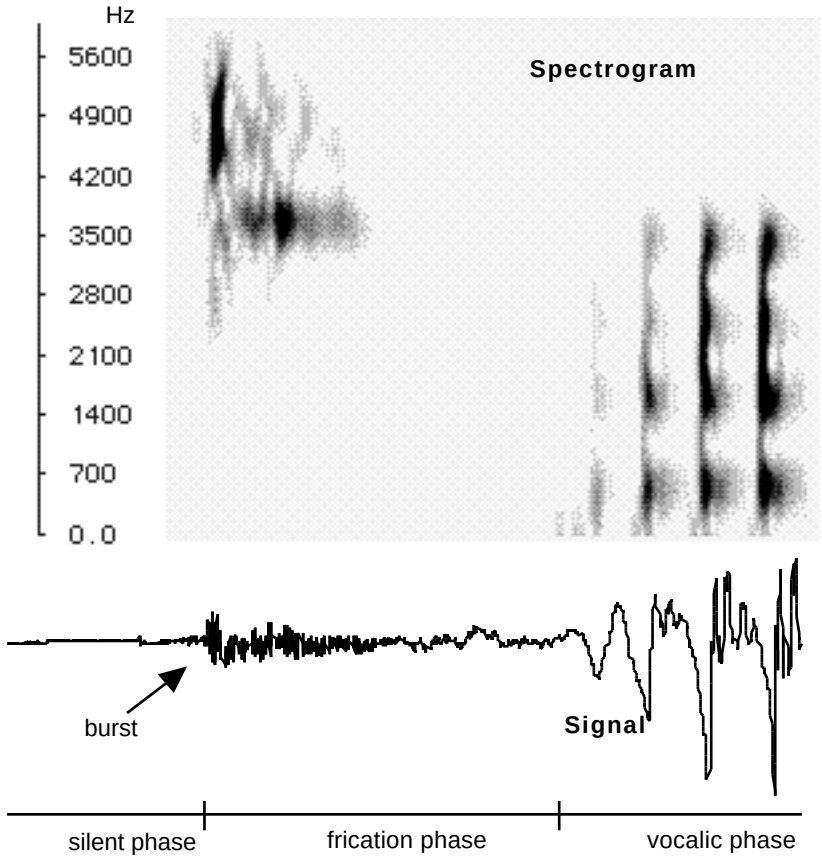

Figure 7. Spectrogram (top) and signal (bottom) of the three phases of the English stop /t/ in the word "test". The silent phase corresponds to the period where the tongue rests against the palate. The frication phase begins with an identifiable burst, as the tongue separates from the palate and an airstream begins rushing through the opening. The vocalic phase corresponds to the onset of vocal fold activity, which imparts a strong cyclic variation to the acoustic waveform.

Nasal consonants involve a complete, sustained occlusion of the oral vocal tract, with the air escaping through the nasal cavity. Acoustically, these sounds can be identified by prominent vocal chord vibrations, a lessened amplitude with respect to that of adjoining vowels, and a reduced presence of higher formants (Figure 8).

The /l/-sounds or *liquids* are produced with a partial and central occlusion of the linguo-alveolar port, whereby air is allowed to escape to either or both sides of the tongue. Acoustically, these sounds are similar to nasal consonants.

Aspirated consonants (such as /h/) arise through general frication in the vocal tract, with a strong component originating at the *glottis*, or at the laryngeal passage. Acoustically, such consonants appear as relatively

weak frication noise showing formant components associated with the surrounding vowels.

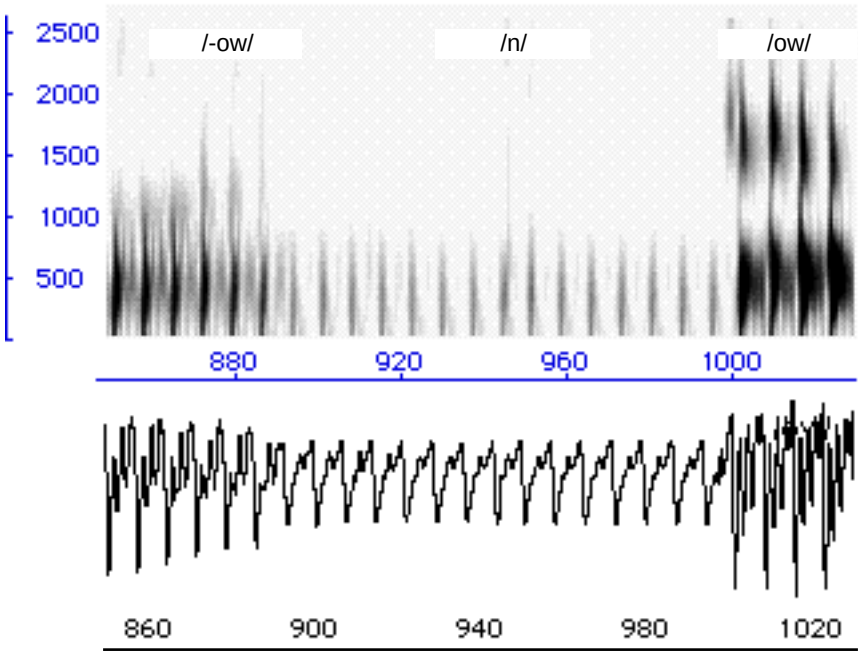

Figure 8. Spectrogram and signal of the nasal consonant /n/ between the two vowels /o/ of "no-*n*o".

Speech as an Integrated String of Speech Sounds

It is of great importance to understand that unlike written text, and despite any perceptual impression, speech is *not a simple, linear sequence of events in time*. Rather, speech is an *integrated* string of speech sounds.

This integration has often been characterised as a simple "fading" from one set of characteristics to the next set of characteristics, occurring at the transition between two speech sounds (one form of this "fading" is known as "feature spreading"). Unfortunately, things are not quite as simple. Some speech sounds (such as vowels) have much greater acoustic impact ("salience") than do other sounds (such as aspirated or nasal consonants). For some other sounds, such as stops, fairly complicated transition patterns exist. Furthermore, transition patterns are not the same inside and between words, nor even inside and between syllables. Speech synthesis devices ignore these complexities at their own peril: devices that do not integrate a relatively sophisticated transitional logic are condemned to sound quite unnatural, and may even produce unusual clicking noises at sound transitions. On the other hand, a great deal of naturalness in synthetic speech can be gained by a good reproduction of human transition patterns.

Again, it is illustrative to examine the speech production process and its reflection in the acoustic waveform. Let us take the case of stops preceding vowels. For example, the sounds /k/ and /t/ are articulated quite differently in front of a /u/ (as in "'coon" or "too") than in front of /i/ (as in "keen" or "tea"). Consequently, the acoustic transition for /ku/ is rather different from that for /ki/, and the acoustic transition for /tu/ is different from that for /ti/.

Initially, this is surprising since to the human ear, the two k's are quite similar to each other, as are the two t's. However, in terms of the parameters that a computer examines, all four sounds are so different from each other that a simple pattern recogniser would be unlikely to assign the two types of /t/ to a single group. More likely, a straightforward classifying algorithm would group the /t/ and the /k/ in front of the /u/ into one group, and the /t/ and the /k/ in the context of the /i/ into another. To understand why, let us turn to how these sounds are produced in the human vocal tract (figures 9 and 10).

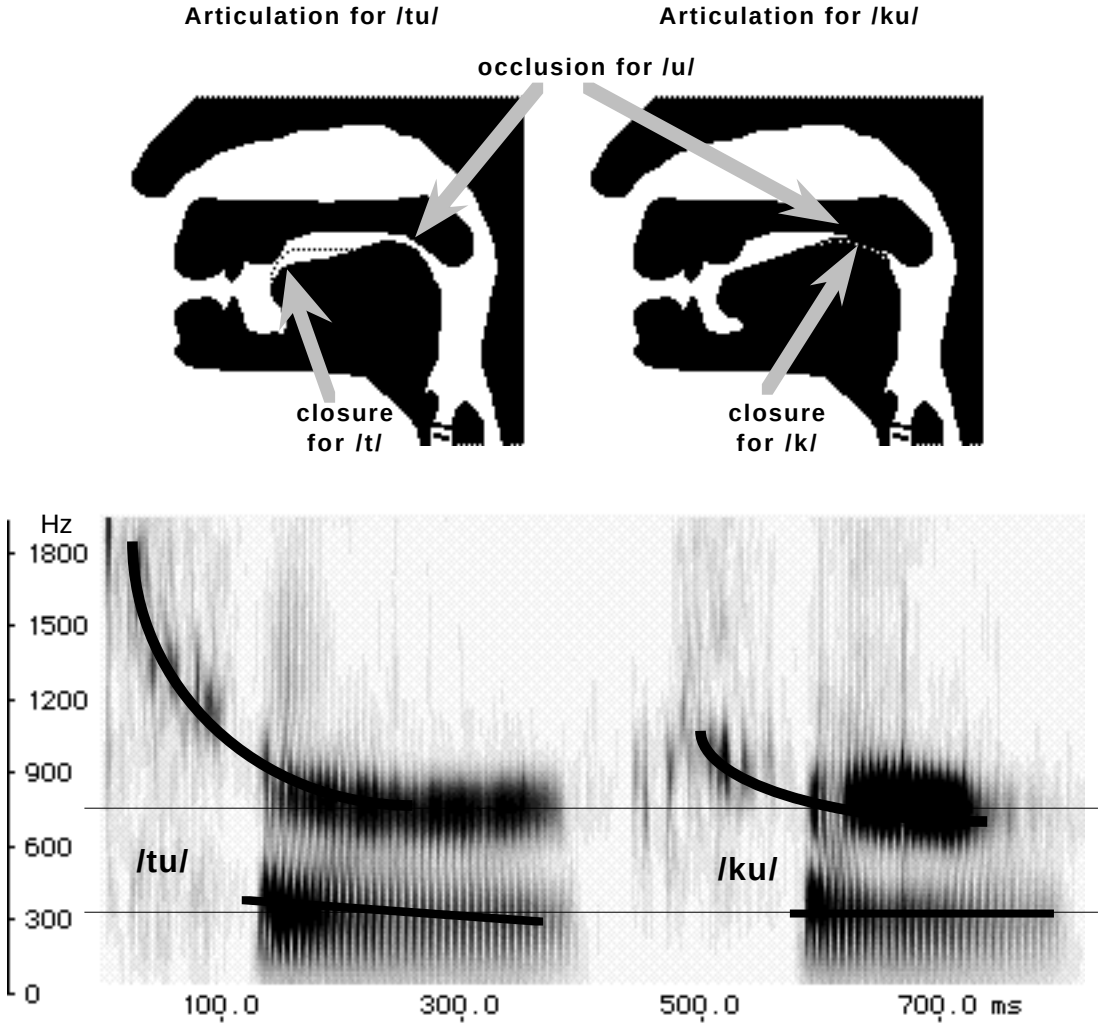

Figure 9. The articulation points for /tu/ and /ku/ and typical spectrograms associated with the two syllables when spoken in isolation. Acoustically /tu/ and /ku/ are quite similar.

As illustrated in Figures 9 and 10, the production of /u/ requires a strong narrowing of the linguo-pharyngeal port, and the production of /i/ requires a similar narrowing of the linguo-alveolar port. Articulatorily, these are the dominant phenomena, since the tongue has to remain in those positions for the entire duration of the vowel in question. By contrast, the preceding stop closure is of short duration, it is a simple flap. Consequently, the location of the flap adjusts to that of the vocalic occlusion. In the case of /k/, the closure occurs in the rear when it precedes /u/, and it is more anterior when it precedes /i/. Similarly, the closure for /t/ is produced in more posterior position when it precedes an /u/, and it generally occurs more anterior when it precedes /i/.

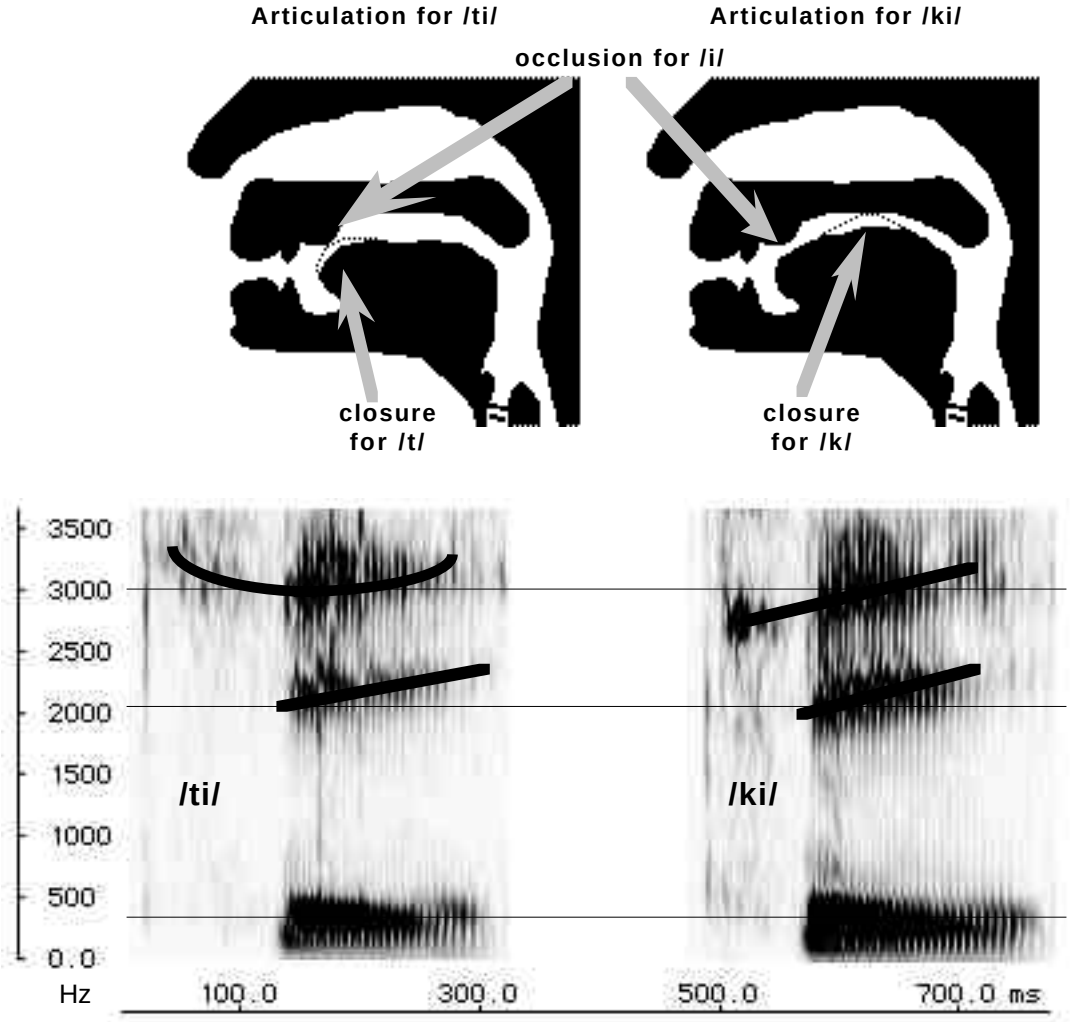

Figure 10. The articulation points for /ti/ and /ki/ and typical spectrograms associated with the two syllables when spoken in isolation. Again, /ti/ and /ki/ are quite similar, and collectively are quite dissimilar from /tu/ and /ku/.

The acoustic effects of these adjustments can be examined in the accompanying spectrograms. The formants that are discernible in the stops preceding /u/ can be seen to "meld into" /u/ while those of the stops preceding /i/ merge with the formants of /i/. The difference

between /t/ and /k/ are minor. In fact, a case can be made for the hypothesis that the human ear is more strongly guided by vocalic than by consonantal differences when distinguishing the two stop consonants[6]. In this view, the main differential information is encoded, not in the stop itself (it is acoustically weak and largely silent anyway), but in the stop-vowel transition and in the adjoining vowel. That is where the ear apparently focuses when it makes its identification of the stop consonant (Fowler, 1979; Pompino-Marschall *et al.*, 1987), and that is where major physical differences can be found.

While the case of the stops is an illustrative extreme, the same principle also applies to a somewhat lesser extent to other sounds. Nasal and fricative consonants, for example, are best modelled in their vocalic context, whereby the type and form of contextual influence depends largely on the consonant's position within the syllable[7].

Speech Signal Analysis Techniques

Most discussion of speech presupposes some knowledge of speech signal analysis techniques. While a detailed review is well beyond the scope of this volume, it may serve to provide a general overview of these techniques, and to refer the reader to some excellent contemporary surveys of this field (e.g. Cooke *et al.*, 1993; Rosen and Howell, 1991).

The basic process of signal analysis is as illustrated in Figure 11. The analysis proceeds by a series of transformations, each of which prepares the signal in some way to highlight specific, speech-related features. Before performing a given transformation, the signal may have to be adequately prepared, e.g. it may have to be filtered or subsampled. Analyses are typically performed either in the temporal or the frequency domain.

Temporal domain: Features in the temporal domain are either evident in the raw signal, or they become (more) evident by deriving a secondary signal that has particular properties. The major time-domain parameters of interest are duration and amplitude. Durations of pauses, syllables, segments, etc., are typically measured directly in the raw signal, or are calculated on the basis of an amplitude envelope (see "RMS", Figure 11). The amplitude is obtained by averaging signal values over a moving time window. Values are squared, so that positive and negative values contribute equally to the amplitude, the mean is taken, and optionally, the

---

6 This point of view is not universally shared.

7 For a more detailed characterization of the principles governing these interactions, please refer to John Local's chapter in Section 3 of this volume.

square root is extracted. The final value plotted over time provides the amplitude curve.

Frequency domain: Features in the frequency domain are identified by spectral analysis. There are a number of techniques that calculate the spectrum from a signal, such as FFT, LPC, Wigner-Ville and cone kernel techniques. Of these, the FFT (Fast Fourier Transform) is the most common. It provides a measure of the frequencies found in a given segment of a signal by decomposing it into its sine components. Another set of particularly useful techniques are the cone kernels (Loughlin *et al.*, 1993). These analyses enhance and isolate the peaked formations in the spectrogram, and thus help identify formants in speech sounds like /a/, where they are too closely spaced for a reliable distinction on the basis of FFT analysis (Figure 12).

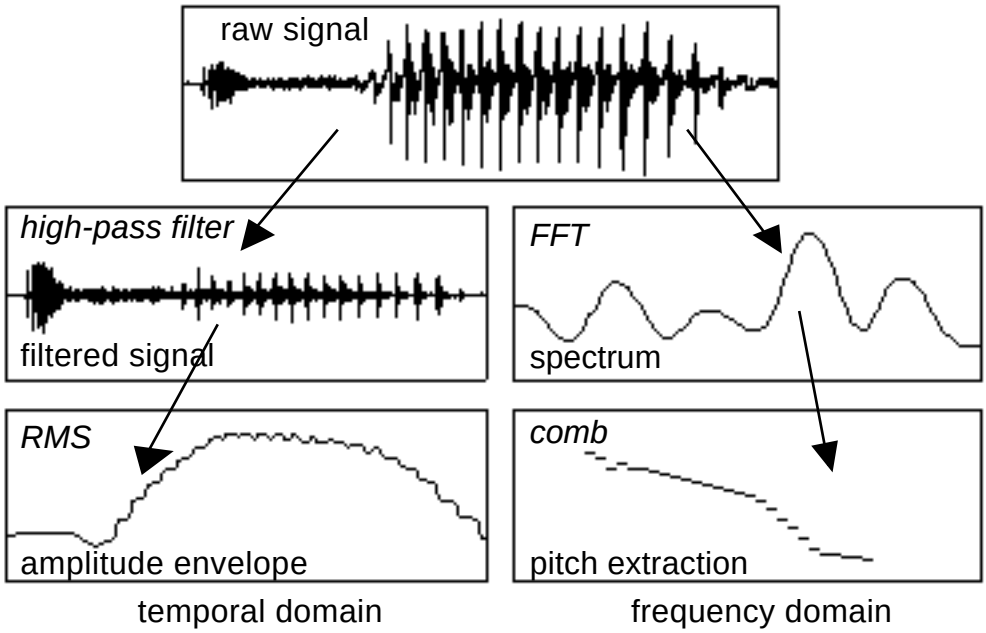

Figure 11. The basic design of speech signal analysis. Work proceeds from the raw signal (top) to a primary decoding level (middle), and/or to secondary and tertiary decoding levels (bottom). These decoding techniques can transform the time-domain primary signal either into a secondary signal, which is displayed in the time domain (left) or in the frequency domain (left). The analysis techniques used for this illustration are named in italics.

The next step in spectral analysis are tertiary signals extracted from spectral information. A typical example are *pitch extractors*, techniques that derive the fundamental frequency from a series of spectra (Hermes, 1993; Hess, 1983). Other common tertiary techniques are the *cepstrum* (a spectrum taken of a spectrum) and the *LPC* ("Linear Predictive Coding", which is often calculated by taking a spectrum taken of an autocorrelation). These algorithms help in removing a person's individual

speech characteristics from the signal, and thus facilitate the identification of fundamental frequency and formants.

Detectors: A special class of analysis techniques is the detector class. These are algorithms designed to identify one particular aspect of speech in the on-going signal. For example, a simple short-term count of how many times the signal crosses the baseline every 10 ms provides a good first estimate of the location of fricative consonants[8] (*cf.* Figure 5). It is possible to write detectors for a number of aspects of speech, such as the silence-speech distinction, the presence of voicing, or the vowel-consonant distinction (Styger *et al.*, in press). Furthermore, artificial neural networks can be programmed to perform such detector functions on selected features of speech (see also article by Torkkola, this volume).

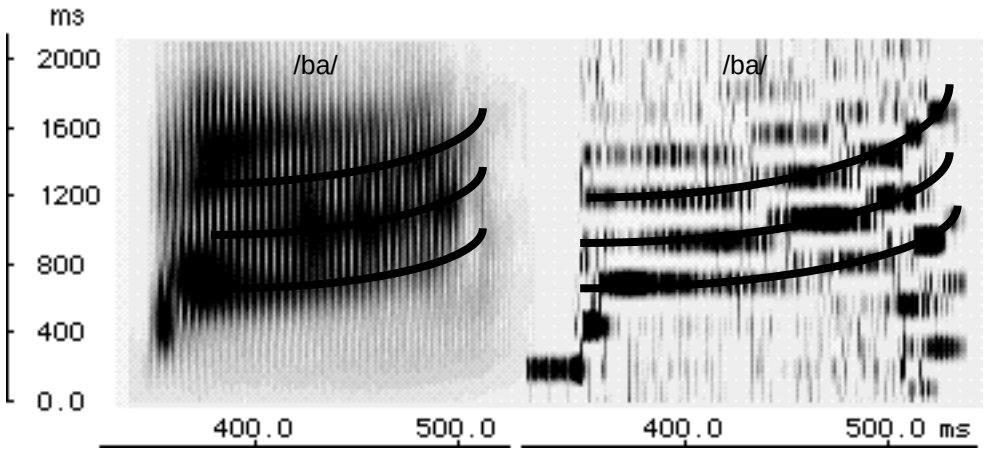

Figure 12. An FFT-based wide-band spectrogram for /ba/ (left) and a wide-band cone kernel spectrogram of the same syllable (right). Cone kernel techniques help isolate formant and fundamental frequency information in cases where resonance bands are closely spaced together.

Finally, *auditory models* take a particularly prominent place in this context[9]. When the speech signal is transformed very much in the way it is modified by the human ear, it becomes easier to separate from background noise. This has major implications for automatic speech recognition. For example, Hunt and Lefèbvre (1989) examined the performance of a recognition system which was given various transformations of a noisy raw signal as input. Transformations that incorporated concepts derived from auditory processing were much more resistent to noise than the standard "cepstrum"-type of signal transform.

[8] Simple zero-crossing counts do well with unvoiced fricatives, but tend to fail with strongly voiced fricatives whose signals often deviate substantially from the zero line. More reliable algorithms detect high-frequency components in the signal, no matter whether they cross the zero line or not.

[9] See the article by Summerfield and Culling, in Section 3 of this volume.

Depending on which measure was used, the authors' own model performed anywhere from 10 to 100 times better than the standard model. Their model works with a spectral representation based on linear discriminant analysis, a representation that had evolved from work on auditory modelling. As can be seen, knowledge about the human auditory capacity can go some distance in charting future directions in speech engineering.

Altogether, the field of speech analysis techniques has been very lively over the past 20 years. New techniques do a much better job at pinpointing speech-related features. Also, how the signal is interpreted has much to do with how it is viewed. In the future, we may view the speech signal in somewhat different ways than we do now, which no doubt will colour our understanding of how machines might better use and reproduce speech signal information. As techniques are developed to improve the identification of specific aspects of speech, recognition devices can be directed to either listen for, or to disregard, selected components of the speech signal. By the same token, synthesis techniques can be custom-tailored to employ such information in order to enhance speech in particular fashion.

Conclusion

Speech communication is based on the principle of distinction. At the speech sound level, classes of distinctive sounds are generally formed on the basis of articulatory organising principles. This is of particularly great importance with respect to stop consonants whose acoustic characteristics are directly related to articulatory execution. These concepts can be exploited for improving the naturalness of speech synthesis devices. Automatic speech recognition may also profit from these concepts by refining techniques of preparing the speech signal for higher-level analyses. A number of new speech analysis techniques go some distance in augmenting recognition capacities in continuous and noisy speech by identifying speech-related parameters against the usual background of irrelevant information.

References

Caelen-Haumont, G. (1991). *Stratégies des locuteurs et consignes de lecture d'un texte: Analyse des interactions entre modèles syntaxiques, sémantiques, pragmatique et paramètres prosodiques*, Thèse d'Etat, Aix-en-Provence.

Fant, G. (1960). *Acoustic theory of speech production*. The Hague: Mouton.

Fowler, C.A. (1979). Perceptual centers. *Speech Production and Perception: Perception & Psychophysics, 25,* 375-388.

Hermes, D.J. (1993). Pitch Analysis. In M. Cooke , S. Beet, & M. Crawford (Eds.), *Visual representations of speech signals* (pp. 3-25). Chichester: John Wiley & Sons.

Hess, W. (1983). *Pitch determination of speech signals (algorithms and devices).* Springer-Verlag.

Hunt, M.J., & Lefèbvre, C. (1989). A comparison of several acoustic representations for speech recognition with degraded and undegraded speech. *Proceedings IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP-89,* Glasgow, Scotland, 262-265.

Lisker, L. (1985). The pursuit of invariance in speech signals. *Journal of the Acoustical Society of America , 77,* 1199-202.

Loughlin, P.J., Atlas, L.E., & Pitton, J.W. (1993). Advanced time-frequency representations for speech processing. In M. Cooke, S. Beet, & M. Crawford (Eds.), *Visual representations of speech signals* (pp. 27-53). Chichester: John Wiley & Sons.

O'Shaughnessy, D.A. (1982). A study of French spectral patterns for synthesis. *Journal of Phonetics, 10,* 377-399.

Pompino-Marschall, B., Tillmann, H.G., & Kühnert, B. (1987). P-centers and the perception of "momentary tempo". *Proceedings of the 11th ICPhS. Vol. 4* (pp.94-97). Tallinn.

Stevens K.N. (1989). On the quantal nature of speech. *Journal of Phonetics, 17,* 3-45.

Styger, T., Gabioud, B., and Keller, E. (in press). Méthodes informatiques pour l'analyse de paramètres primaires en parole pathologique. In J.-P. Goudailler (Ed.), *Les faits intonatifs dans l'acquisition et la pathologie du langage.* CALAP, no. 11.

Introductions and Reviews

Cooke, M., Beet, S. & Crawford, M., (Eds). (1993). *Visual representations of speech signals.* Chichester: John Wiley & Sons.

Hardcastle, W.J. & Marchal, A., (Eds). (1990). *Speech production and speech modeling.* Dordrecht: Kluwer Academic Publishers.

Ladefoged, P. (1975). *A course in phonetics.* New York: Harcourt Brace Jovanovich.

MacNeilage, P.F., Studdert-Kennedy, M.G., and Lindblom, B. (1985). Planning and production of speech: An overview. *Journal of the American Speech and Hearing Association, 15,* 15-21.

MacNeilage, P.F. , (Ed.). (1983). *The production of speech.* Berlin: Springer Verlag.

Rosen, S. & Howell, P. (1991). *Signals and systems for speech and hearing.* London: Academic Press.

Tohkura, Y., Vatikiotis-Bateson, E., & Sagisaka, Y. , (Eds.). (1992). *Speech perception, production and linguistic structures.* Amsterdam: IOS Press.