Animal Acoustic Communication Springer-Verlag, 1998, 251-290

CHAPTER 8

Measuring and Modeling Speech Production

P. RUBIN AND E. VATIKIOTIS-BATESON

1 Introduction

In human communication, the speech system is specialized for rapid transfer of information (Liberman et al. 1967; Mattingly and Liberman 1988). Significant events in the acoustic signal can occur in an overlapped or parallel fashion due to the coproduction of speech gestures. One result is that aspects of the signal corresponding to different linguistic units, such as consonants and vowels, often cannot be isolated in the acoustic stream. One way to help tease apart the components of the speech signal is to consider the physical system that gives rise to the acoustic information: The acoustic encoding of phonetic information is viewed in light of the flexibility inherent in the production apparatus, particularly the human supralaryngeal vocal tract, in which individual articulators or groups of articulators can function semi-independently. In this chapter we review this approach. First, we show how the analysis of speech acoustics has benefited by treating the sound production system as one in which the contributions of physical acoustic sources and physiologically determined filters are combined. We then discuss how acoustic diversity has resulted in a desire to find articulatory simplicity. In the process, we review some of the methods used to examine articulatory activity, and also describe in detail a particular attempt at modeling the coordination of the speech articulators. Finally, we consider some recent attempts to explore the links between production, perception, and acoustics in a dynamic-systems approach and in connectionist models. Where possible, recent trends in the field have been exemplified by projects involving ourselves and our colleagues. Although articulation in most animals is simpler than human speech production, the methods we describe are also applicable in this domain.

In order to consider the details of acoustics and production, it is necessary first to briefly describe the system being studied. Figure 1 shows a schematic view of the human sound production system. Note the extent of the anatomy involved in the production process (Ladefoged 1975; Borden and Harris 1984; Lieberman and Blumstein 1988). There are a variety of ways to produce sound. One method involves using the air pressure provided by the lungs to set the elastic vocal folds into vibratory motion. The larynx converts the steady flow of air produced by the subglottal system into a series of puffs, resulting in a quasi-periodic sound wave. Aperiodic sounds are produced by allowing air to pass through the open glottis into the upper airway (the supralaryngeal vocal tract) where localized turbulence can be produced at constrictions in the tract. A third method involves producing transient clicks and pops by rapid release of the articulatory

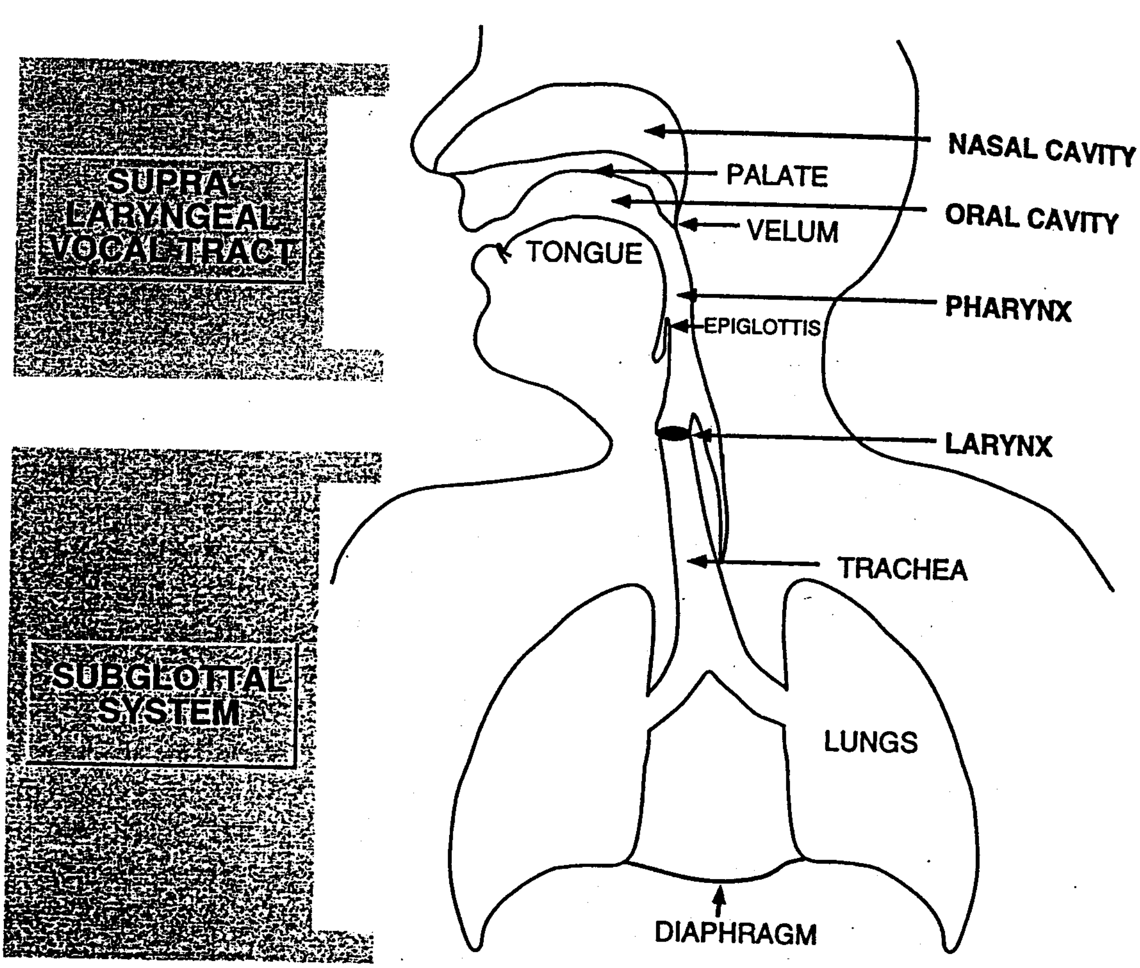


Fig. 1 The human speech production system (see text for details)

closure (Ladefoged 1975). Here the sound sources arise from the local changes in the vocal tract and do not require air pressure from the subglottal system.

2 The Acoustic Theory of Speech Production and Acoustic Analysis

2.1 The Source-Filter Model

Acoustic speech output in humans and many nonhuman species is commonly considered to be the combined outcome of a source of sound energy (e.g., the larynx) modulated by a transfer (filter) function determined by the shape of the supralaryngeal vocal tract. This combination results in a shaped spectrum with broadband energy peaks. This model is often referred to as the "source-filter theory of speech production" and stems from the experiments of Johannes Müller (1848) in which a functional theory of phonation was tested by blowing air through larynges excised from human cadavers.

"Müller...noticed that the sound that came directly from the larynx differed from the sounds of human speech. Speechlike quality could be achieved only when he placed over the vibrating cords a tube whose length was roughly equal to the length of the airways that normally intervene between the larynx and a person's lips. The sound then resembled the vowel [uh], the first vowel in the word *about*..." (Lieberman 1984, p. 131). In this model, the source of acoustic energy is at the larynx — the supralaryngeal vocal tract serves as a variable acoustic filter whose shape determines the phonetic quality of the sound (Fant 1960).

When the larynx serves as a source of sound energy, voiced sounds are produced by a repeating sequence of events. First, the vocal cords are brought together (adduction), which temporarily blocks the flow of air from the lungs and leads to increased subglottal pressure. When the subglottal pressure becomes greater than the resistance offered by the vocal folds, they open again. The folds then close rapidly due to a combination of factors, including their elasticity, laryngeal muscle tension, and the Bernoulli effect. If the process is maintained by a steady supply of pressurized air, the vocal cords continue to open and close in a quasiperiodic fashion. As they open and close, puffs of air flow through the glottal opening. The frequency of these pulses determines the fundamental frequency (F_0) of the laryngeal source and contributes to the perceived pitch of the produced sound. An example of the spectrum of the result of such glottal air flow is plotted at the top left of Figure 2. Note that there is energy at the fundamental frequency ($F_0 = 100$ Hz) and at the harmonics of the fundamental, and that the amplitude of the harmonics falls off gradually. The bottom of Figure 2 shows the comparable case for a fundamental frequency of 200 Hz. The rate at which the vocal folds open and close

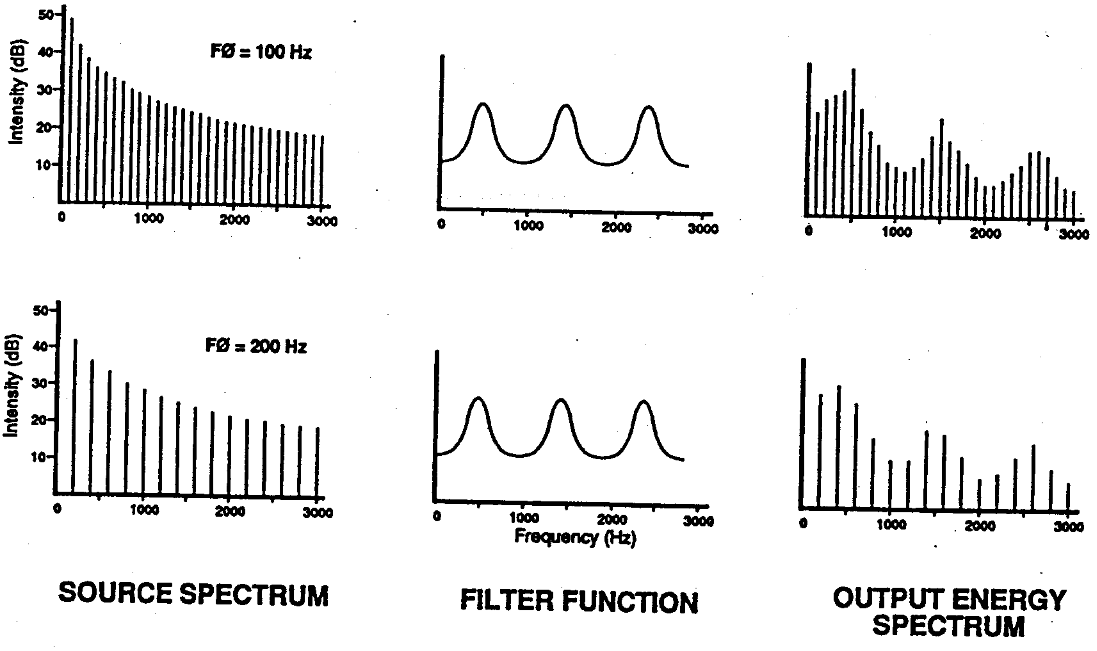


Fig. 2. The source-filter model of speech production. The source spectrum represents the spectrum of typical glottal air flow with a fundamental frequency of 100 Hz. The filter, or transfer, function is for an idealized neutral vowel /ə/, with formant frequencies at approximately 500, 1500 and 2500 Hz.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

during phonation can be varied in a number of ways and is determined by the tension of the laryngeal muscles and the air pressure generated by the lungs. The shape of the spectrum is determined by details of the opening and closing movement, and is partly independent of fundamental frequency. In normal speech, fundamental frequency changes constantly, providing both linguistic information, as in the different intonation patterns associated with questions and statements, and information about emotional content, such as differences in speaker mood. In addition, the fundamental frequency pattern contributes to the naturalness of utterance production. This effect can be illustrated by creating a synthetic version of a natural utterance in which the spectral properties are left largely unchanged while the normally varying fundamental is replaced with one of constant frequency.

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

The spectrum of the glottal air flow, which has energy at the fundamental frequency (100 Hz) and at the harmonics (200, 300 Hz, etc.), is plotted in the top left of Figure 2. The amplitude of the harmonics, which for the purposes of this figure combines the effects of both the source spectrum and radiation, decreases by approximately 6 dB per octave. The top right of the figure shows the spectrum that results from filtering the laryngeal source spectrum in the top left panel with the idealized filter function shown in the center of Figure 2. Note that the laryngeal source energy has been "shaped " by the filter function. Energy is present at all harmonics of the fundamental frequency of the glottal source function, but the amplitude of an individual harmonic is determined by both its source amplitudes and the filter function. The bottom half of Figure 2 shows the effect of using a different source function, while retaining the same filter function. In this case, the fundamental frequency of the glottal source is 200 Hz, with harmonics at integer multiples of the fundamental (400 Hz, 600 Hz, etc.). The spectrum that results from combining this glottal energy with the filter function for an idealized /ə/ has the same overall pattern as that shown above it. However, there are differences in the details. Note, for example, that the lowest formant for /ə/ has a center frequency of 500 Hz. A glottal waveform with a fundamental of 100 Hz has a harmonic at this frequency. A source function with a fundamental of 200 Hz has harmonics that straddle the lowest formant (i.e., at 400 and 600 Hz), as shown in the bottom right of Figure 2. Since the overall shapes are the same, these details do not change the perceived vowel quality,

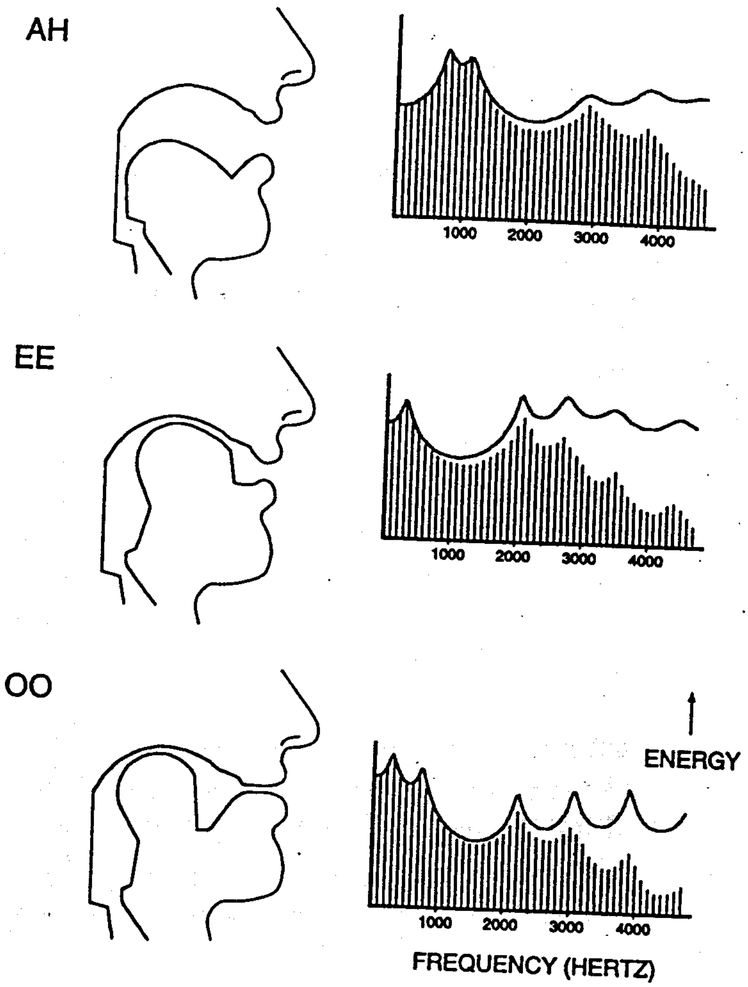


Fig. 3 The source-filter model for selected vowels

which would be that of an /ə/. However, the example shown at the top right of Figure 2 would be perceived to have lower pitch because of its lower fundamental frequency.

The flexibility of the human vocal tract, in which the articulators can easily adjust to form a variety of cavity shapes, results in the potential to produce a wide range of sounds. For example, the particular vowel quality of a sound is determined mainly by the shape of the supralaryngeal vocal tract, as reflected in the filter function. Figure 3 illustrates this effect. Three different vocal tract shapes are shown corresponding to the vowels “ah ”(/a/), “ee” (/i/), and “oo” (/u/). In this example, we have used schematized vocal tract shapes from the Haskins Laboratories articulatory synthesizer (see Sect. 9). Plotted in the same graph for each tract shape is the smoothed transfer function that is computationally derived by the synthesizer, as well as the hypothetical energy spectrum that would result from using these functions to filter a glottal source spectrum with a fundamental frequency of 100 Hz. Note that although all three vowels have the same fundamental frequency, their spectra differ according to the filter characteristics of the various vocal tract shapes. Detailed accounts of the acoustic properties of the vocal tract can be found in a number of sources, including Fant (1960), Flanagan (1965), Fry (1979), and Lieberman and Blumstein (1988).

3 A Brief Comparison of Human and Nonhuman Vocal Tract Characteristics

The acoustic theory of speech production (Fant 1960; Flanagan 1965; Lieberman 1975) described in Section 2 relates changes in the vocal tract shape to the resultant acoustics. In an application of this approach, Lieberman and colleagues (Lieberman 1969; Lieberman et al. 1969) studied animal vocalization using acoustic analysis procedures similar to those used for human speech. In particular, spectrographic analysis, waveform measurements, and computer modeling were used to study the vocalic repertoire of a rhesus monkey. Formant frequencies were calculated using area functions derived from plaster castings of the animal's oral cavities and the simulated acoustic properties of the monkey's vocal tract were compared with those of the human. A plot of the lowest two formant frequencies (F1 by F2) revealed that both simulated monkey formant frequencies and values for actual vocalizations lie in an extremely limited acoustic range compared with that of the human vowel space. This limitation was attributed to an anatomical difference between humans and other mammals: While the pharyngeal and oral cavities of the human vocal tract lie at right angles to one another, those of nonhuman vocal tracts are more nearly in a straight line (see Figure 4). Humans can position their tongues in a manner that changes the point of maximum constriction in the tract, allowing differential shaping of the entire structure into two (or three) cavities (tubes). This flexibility in the shaping of the vocal tract's cavity relationships permits the production of the wide variety of sounds observed in speech (Lieberman 1975, 1984). Although Lieberman's simulations are of interest, it should be pointed out that he did not empirically explore the range of sounds that rhesus monkeys are capable of producing (see below).

Comparison of the anatomy of the human larynx with those of other vertebrates suggests that in attaining the power of speech, some of the protective functions provided by more primitive larynges were relinquished (Lieberman 1984; Kirchner 1988). In nonhuman mammals, the larynx and epiglottis are higher in the pharynx than in adult humans. Indeed, for most mammals, especially the herbivores, the larynx is so high that it contacts the soft palate, thus causing a separation of breathing and eating functions. As shown in Figure 4, the epiglottis is in a relatively high position in these animals. It is raised during drinking or eating, sealing off the oral cavity from the nasopharyngeal passage and making it almost impossible for liquid or solid food to pass into the pharynx. This arrangement is also observed in human neonates, and prevents them from suffocating while nursing (Laitman et al. 1977; Sasaki et al. 1977). The adult human is not so fortunate, and is subject to choking while eating. In this case, the epiglottis and larynx are lower and the pharynx serves as a common pathway for air, liquids, and food — increasing the chance that objects may fall into the larynx and block the airway to the lungs (Lieberman 1984).

Recent work (reviewed by Owren and Linker 1995; Hauser 1996) questions the presumption that nonhuman primates lack both the laryngeal control and the flexibility in vocal tract movement needed to produce a variety of meaningful utterances. There

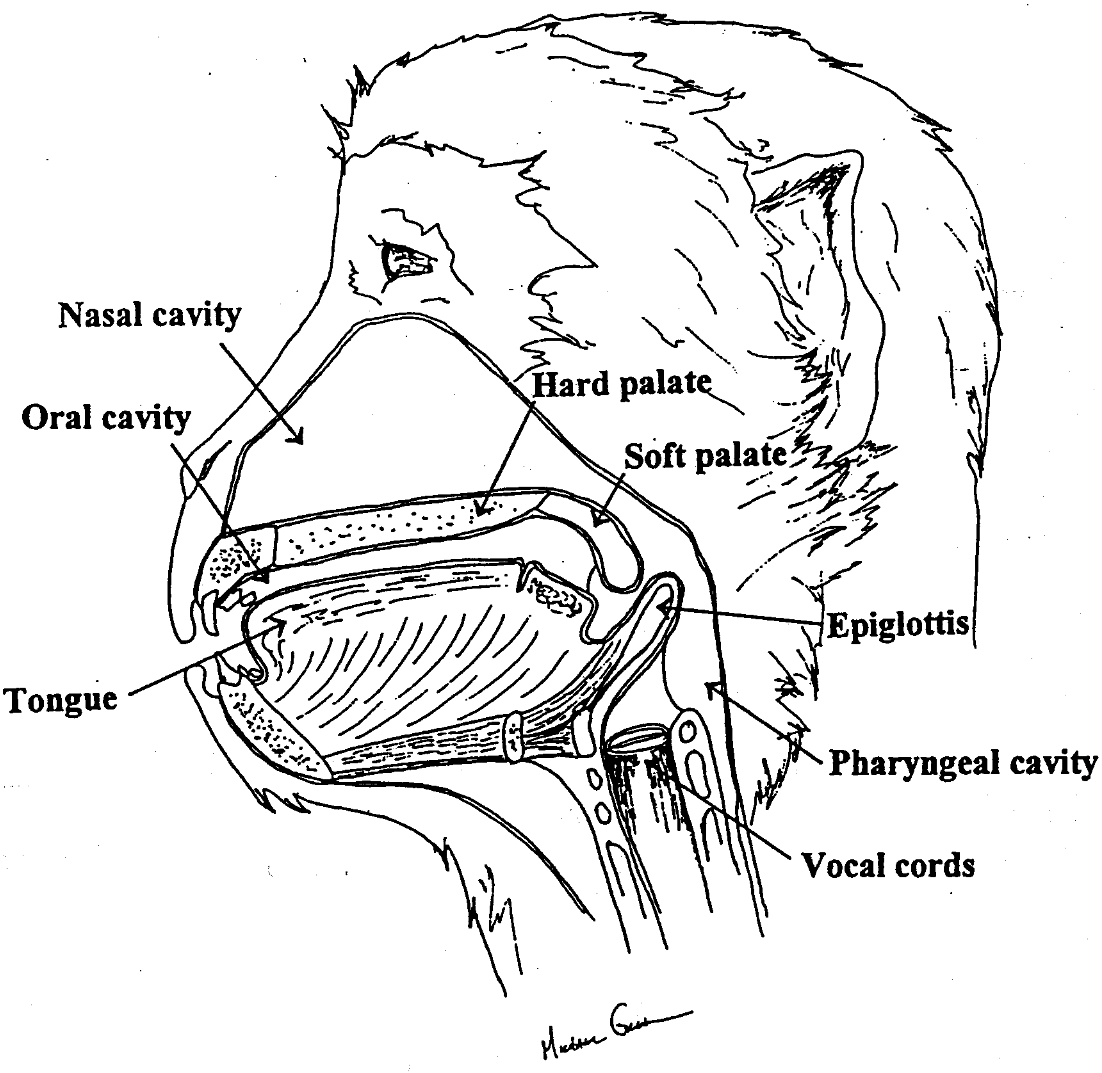


Fig. 4. Midsagittal view of a rhesus monkey vocal tract. (Drawing by Michael Graham)

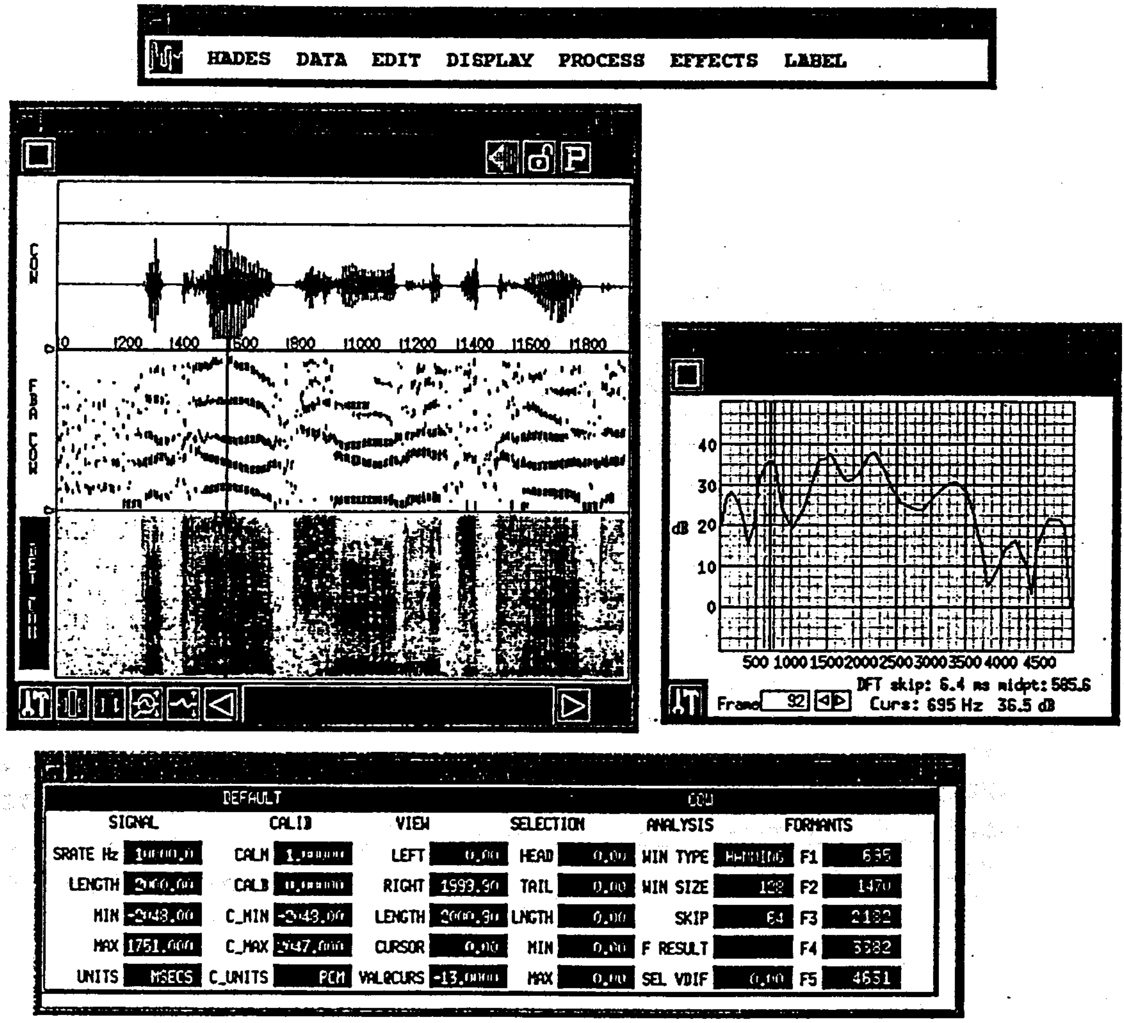
is strong evidence of vocal tract filtering in the productions of a number of nonhuman primates, including baboon *grunts* (Andrew 1976; Owren et al. 1993, 1995), tonal *coo* calls and other vocalizations in macaques (Hauser et al. 1993), semantic alarm calls in vervet monkeys (Owren and Bernacki 1988; Owren 1990), and *double grunts* produced by wild mountain gorillas (Seyfarth et al. 1994). The work of Hauser and colleagues (Hauser 1991, 1992, 1996; Hauser and Fowler 1992; Hauser et al. 1993; Hauser and Schön Ybarra 1994) provides particularly compelling evidence for the ability of monkeys and apes to alter the shape and length of their vocal tracts using a variety of articulatory maneuvers, resulting in a range of resonant frequency patterns. Their findings suggest that spectral properties can be controlled independently of the glottal source. There is also substantial evidence for precise control of vocal fold vibration in monkeys and apes (Andrew 1976; Hauser and Fowler 1992; Hauser et al. 1993; Owren et al. 1995). Additionally, Nowicki and his colleagues (see Gaunt and Nowicki, this Volume) have demonstrated that the vocal tract above the syrinx modifies sounds produced by songbirds. These animals can also modify sounds both through "articulatory" movements of the beak and by elongating the neck, which changes the resonance of the vocal tract. Taken together, these results stand in strong opposition to earlier claims that the source-filter approach applies only in the case of human speech production.

4 Acoustic Analysis

A rich set of tools and techniques is available for the detailed study of speech acoustics (Rabiner and Schafer 1978; Witten 1982; Fallside and Woods 1985; O'Shaughnessy 1987, 1995; Kent and Read 1992). The ability to convert the acoustic signal to a digital record means that computers, with an arsenal of numerical and statistical methods, can be used to tease apart the fine structure of a sound both rapidly and reliably. Digital techniques have, in large part, replaced oscilloscopes and analog spectrographic devices. Digital analysis techniques generally provide a greater dynamic range, are more flexible, and use a representation of the signal that can be reanalyzed and reviewed in a variety of ways. The availability of affordable and powerful personal computers, with high-resolution graphical displays and sound digitizing interfaces, allows these tools to be used both in a desktop system, and in the field.

Techniques that examine the time-domain characteristics of the signal can be used to make duration measurements and to provide information about intensity and periodicity. Frequency – domain analysis is used to examine the spectral characteristics of the signal – the underlying formant structure, in particular the details of individual spectral cross-sections. A common approach uses analysis based upon the Fourier transform, which represents signals as sums of weighted sinusoids (Brigham 1974; Fallside 1985). Another common technique that provides information about frequency values is linear prediction coding (LPC) (Atal and Hanauer 1971; Markel and Gray 1976; Atal 1985), a time-domain coding method that is used in analysis, storage, and synthesis. LPC implements a source-filter model for separating the source characteristics of the signal (e.g., fundamental frequency) from the filter (vocal-tract) characteristics. Examples of the use of LPC analysis in the study of animal vocalization are described in more detail in the chapter by Owren and Bernacki (this Volume; see also Carterette et al. 1979, 1984; Owren and Bernacki 1988; Owren 1990; Shipley et al. 1991; Seyfarth et al. 1994).

Figure 5 shows the output of a system used at Haskins Laboratories, known as the Haskins Analysis Display and Experiment System, (HADES; Rubin 1995), that performs such analyses. The large window (on the left in Figure 5) contains three panels that provide both temporal and spectral information about a signal. The top panel is an acoustic waveform (energy vs. time) of the utterance, "The cow chewed its cud". The bottom panel is a spectrogram of the same utterance, calculated using fast Fourier transform (FFT) analysis — a computational method for efficiently calculating the discrete Fourier transform (Cochran et al. 1967). The spectrogram provides a representation of frequency (on the vertical axis) and amplitude (depicted by varying levels of darkness) over time (the horizontal axis). Note the formant structure in this broadband analysis of the signal, as shown by the darker bands of greater acoustic energy. The middle panel provides summary information from an LPC analysis of the signal. The short dark ver



DFT skip: 6.4 ms midpt: 585.6
Frame 92
Curs: 695 Hz 36.5 dB

DEFAULT COW

SIGNAL		CALIB		VIEW		SELECTION		ANALYSIS		FORMANTS	
SRATE Hz	10000.0	CALM	1.000000	LEFT	0.00	HEAD	0.00	WIN TYPE	HAMMING	F1	695
LENGTH	2000.00	CALB	0.000000	RIGHT	1999.90	TAIL	0.00	WIN SIZE	128	F2	1470
MIN	-2048.00	C_MIN	-2048.00	LENGTH	2000.90	LNGTH	0.00	SKIP	64	F3	2182
MAX	1751.000	C_MAX	1747.000	CURSOR	0.00	MIN	0.00	F RESULT		F4	3382
UNITS	MSECS	C_UNITS	PCM	VAL@CURS	-13.0000	MAX	0.00	SEL VDIF	0.00	F5	4651

Fig. 5. HADES spectrogram and waveform

tical bars correspond to the peaks of the smoothed LPC spectra. An example of such a spectrum is shown in the window on the right Figure 5. This window shows a spectral cross-section for a portion of the vowel in the word "cow", at the point indicated by the vertical bar in the spectrogram. The "control panel" window at the bottom of the figure provides additional information about the signal, including the "formant" peaks (for the selected spectrum) automatically obtained from the LPC analysis (these are actually the peaks of the LPC spectrum for a selected frame). This sort of analysis is the standard approach for displaying and quantifying spectral and temporal aspects of both speech and non-speech signals. As mentioned above, tools for producing such analyses are now readily available on desktop and portable computers. Other chapters in this Volume provide additional details about speech analysis techniques.

Synthesis provides another tool for studying speech acoustics. In this technique, the acoustic waveform is created electronically, through the use of specialized circuitry or by combinations of computer hardware and software. A variety of methods can be used to provide an acoustic description. These include replaying stored samples of segments, words, and phrases; copying and regenerating aspects of the acoustic information from its spectrographic representation; specifying segmental descriptions (phonemes, diphones, syllables, etc.), concatenating them, and then deriving their acoustic properties; and using text-to-speech systems that include phonetically based rules to automatically convert strings of text into acoustic patterns. This methodology provides a practical means for experimentally testing the perceptual significance of particular aspects of the acoustic information in the speech signal (Flanagan and Rabiner 1973; Klatt 1980, 1987). For example, details of both the source and filter can be provided and systematically manipulated and evaluated. Overviews of speech perception, analysis, coding, and synthesis are available in Flanagan (1965), Witten (1982), and O'Shaughnessy (1987, 1995).

Signal synthesis has frequently been used to simulate animal vocalizations, although usually involving editing of natural sounds or relatively simple waveform generation. Perhaps the earliest application of speech-related techniques to nonhuman vocalization synthesis was Capranica's (1966) formative study of bullfrog mating calls. In this work, analog equipment was configured in a source-filter fashion with periodic pulse trains and white noise providing input to a parallel combination of resonant circuits. Digitally based linear-predictive synthesis was later used by Carterette and his colleagues (Carterette et al. 1984) in studying cat vocalizations (see also Shipley et al. 1991). Signal processing analyses of alarm calls of vervet monkeys by Owren and Bernacki (1988) provided the basis for evaluating species-typical perceptual processing using LPC-based synthetic stimuli (Owren 1990). Monkey calls generated specifically by speech synthesizers have been used in a number of studies of perceptual processing in Japanese macaques, including work by Petersen (1981), May et al. (1988, 1989), and Hopp et al. (1992).

5
Measuring and Analyzing Speech Production

The configuration of the human vocal tract, which "shapes" speech acoustics, depends on the position of the speech articulators (e.g., the tongue, lips, jaw, velum, and larynx). Furthermore, because the acoustics are continually changing during speech, it is the behavior of the speech articulators over time — the *changes* of articulatory configuration and their acoustic consequences — that must be analyzed. Recently, more adequate tools for observing speech production have been developed resulting in renewed interest in considering speech articulation and acoustics together. This Section provides a brief sketch of the history of acoustic and articulatory research and examines some major experimental and analytical techniques used to study speech articulation. Of particular relevance to research with animals, we suggest, are those techniques focusing on the dynamic behavior of individual articulators (e.g., jaw, tongue), the functional or task-specific spatiotemporal coordination among articulators (e.g., lip—jaw, tongue—jaw, lip—larynx), and the specification of an articulatory-acoustic relation from which anatomical and behavioral constraints may evolve.

6 Different Limitations of Human and Animal Research

Research using both human and animal subjects may endanger them. This has limited invasive experimentation, particularly in humans. Therefore, much of our detailed physiological and anatomical knowledge of sound production and perception systems has come from studies of other animals, often chosen because of presumed structural and/or functional similarities with humans — for example, hearing in chinchillas and cats (Neff 1964), laryngeal structure and innervation in dogs and monkeys (e.g., Larson 1988; Alipour-Haghighi and Titze 1991), and perceptual processing in nonhuman primates (Stebbins and Summers 1992). Direct knowledge of humans comes from *post mortem* anatomical and histological studies (Galaburda 1984), from studies of accidental trauma due to war, disease, injury, and surgery (e.g., Walsh 1957; Luria 1975; Weismer 1983), or from congenital abnormalities such as cleft-palate (Warren 1986) and deafness (Rubin 1983). The great advantage of using human subjects in behavioral studies is their ability to understand instructions and exert adequate control at various levels of perception and production. Behavioral research on other animals often must be conducted indirectly via discrimination tasks, which usually entail a considerable investment of resources for training (e.g., Norris and Møhl 1983; Petersen et al. 1984). However, at least until recently, research involving animals has been fairly free to combine systematically well controlled behavioral and physiological techniques – techniques which tend to be dangerously invasive (for overviews, see: Neff et al. 1975; Simmons and Grinnell 1988).

7 Overview of the Shifting Articulatory and Acoustic Emphases in Speech Research

Before high-resolution data transduction and recording techniques became available in the second half of this century, speech was ephemeral and its record impressionistic. The best a transcriber could do was to write down what was said in a notation that would allow a rough reconstruction of what had been heard. By the 1880s, a standard orthography for phonetic transcription had emerged (the International Phonetic Alphabet, or IPA), which scholars hoped would be rich enough for precise transcription of utterances in any of the world's languages. This development was grounded in the use of symbols for minimally distinctive sound segments (e.g., the phonemes /d/ and /t/ in 'hid' vs. 'hit') augmented by diacritic marking of context-specific phonetic differences. However, the method was still susceptible to the biases or misperceptions of the listener/transcriber.

Early methods of in vivo investigation were restricted to what could be seen (e.g., movement of the lips and jaw), felt (e.g., vibration of the larynx, gross tongue position), or learned from practiced introspection of articulator position during production (e.g., Bell 1867). However, when the results of these methods were combined with those from anatomical and mechanical studies of cadavers, a great deal was correctly surmised about the relation between vocal tract shape and the resultant acoustics. This knowledge had practical applications such as teaching the deaf to speak, and provided the basic scheme for modeling articulation e.g., the vowel space, pitch dependence on pressure, elastic tension of the vocal folds, and height of the larynx.

By the end of the nineteenth century, the development of mechanical transduction techniques for slow moving events such as rhythmic motion of the jaw, thorax (respiration), and other rhythmically entrained structures (e.g., finger- and foot-tapping during speech) made fairly detailed kinematic analysis possible (Sears 1902; Stetson 1905). Physiological studies began as well, for instance, examining transduction of neuromuscular events (Stetson 1928). By modern standards, analysis of data from these studies was quite basic (e.g., measures of duration and observationally inferred estimates of articulator speed and impulse force). Nonethless, during this period many interesting claims and comprehensive hypotheses were advanced concerning the organization and control of learned voluntary behaviors (see Boring 1950 for a review). The culmination of this epoch, in which the basic research paradigm primarily entailed inference of articulatory events during production of minimally contrastive phonetic events, was the development of X-ray photography. Finally, the configuration of the entire vocal tract could be captured — first statically and later dynamically (cineradiography) — during speech production. Unfortunately, the characteristic events revealed by analysis of these data were very coarse-grained and difficult to quantify.

The rapid development of acoustic recording techniques and the sudden awareness of the dangers of X-ray exposure in the late 1920s helped drive the shift in interest from speech articulation to acoustics. During the 1930s and 1940s, recording, display, and analysis systems such as wire-recorders, the oscilloscope, and the sound spectrograph (Koenig et al. 1946), respectively, made it possible to study speech acoustic events in greater detail and revealed phoneme-specific information in the acoustic patterns. In particular, vowel formants and consonant-dependent formant transitions were recognized as key components to phoneme identity, and their patterning alone was shown to be sufficient for synthesis of acceptable and distinct syllables such as /ba/, /da/, /ga/ (Cooper et al. 1951; Liberman et al. 1959). Given the ability to synthesize speech from acoustic patterns, it seemed possible to conduct meaningful research in the acoustic domain alone, regardless of the underlying articulatory configurations.

Another contributor to the shift in focus from articulation to acoustics was the emergence of *distinctive-feature* theory. Heavily influenced by information theory (e.g., Fano 1949), this approach proposed that phonemes are composed of distinctive features whose binary values (+/-) enable minimal phonemic and, therefore, informational contrasts (Jakobson et al. 1963). Thus, consonant pairs such as /p,b/, /t,d/, and /k,g/ can be distinguished by the value of the voicing feature alone. Although feature detection was not restricted to the acoustic domain, the distinctive feature's role as the critical information-bearing element in the process of perceiving speaker intention required that the underlying message be encoded in the acoustic properties of the signal. This connection led quite naturally to the assumption that the medium of communicative interaction was strictly acoustic.

These developments were followed by an intense effort to decompose acoustic signals into minimally contrastive cues to "phonetic intent" (e.g., Lisker 1957; Abramson and Lisker 1965; Liberman et al. 1967; Liberman and Studdert-Kennedy 1978; Repp 1983, 1988; O'Shaughnessy 1987). However, despite demonstration of the perceptibility of a variety of acoustic cues, a number of problems became apparent. Taken together, these persistent difficulties suggest that consideration of acoustics alone may not reveal how listeners actually arrive at a given speech percept. One problem is that speech acoustics are highly variable both within and across speakers. Not only can the same phoneme have different acoustic properties when produced by different speakers, but also no two utterances by the same speaker are acoustically the same. Another problem is the difficulty in finding acoustic cues that persist across the range of phonetic contexts. Thus, the relatively long time between the release of a voiceless stop consonant (e.g., /p,t,k/) and the onset of voicing for a following vowel (as in /pa/) clearly distinguishes this kind of sound from its voiced counterpart (/b,d,g/, respectively) — if the syllable is stressed. In other contexts, however, this voice-onset time (VOT) dimension is not as strong a cue to voicing, or may be absent entirely (e.g., before a pause). The search for persistent cues led to a third problem, that multiple cues to the same phonetic feature may overlap or trade off, depending on the context, speaker, or other factors. For example, vowel quality and duration can both provide cues to consonant voicing (Lisker and Abramson 1967). The inability to identify unique cues for specific features suggested the possibility of multiple mappings between phonetic categories and acoustic distinctions. Finally, recent research has shown that both visual and acoustic information can be useful in speech perception (Benoît et al. 1992, 1994; Massaro et al. 1993; Sekiyama and Tohkura 1993). Indeed, the two modalities appear to complement each other in that some of the cues that are more unstable acoustically (for instance, cues to place of articulation that often occur very briefly in the fine structure of the spectrum) are visually the most consistently perceived, while the opposite is true for acoustically more stable cues such as nasality (for review see Summerfield 1987, 1991).

In articulation research, recognition of the mismatch between variable acoustic events and phonetic categories coincided with improvements in articulatory transduction techniques (e.g., greater subject safety, increased measurement accuracy, and reduced cost) and a growing corpus of physiological data from studies of other biological movement systems that suggested lawful constraints on the organization and production of biological behavior. Coupled with the notion that communicative intent must go through the production structure before any acoustic "encoding" can occur, these developments have led to an emphasis on considering phonetic abstractions in terms of underlying articulatory behavior.

As an example, consider production of the bilabial /b/ between two vowels (e.g., /aba/). Acoustically, no two productions are identical, even in the same speaker. Yet, the articulatory event is relatively simple. Following the first vowel, the lips come together, stopping the airflow from the vocal tract, and then are released for the next vowel. Vocal fold vibration, necessary for each of the two vocalic segments, may or may not stop during the closure period. Such stopping depends on whether or not the lips are closed long enough for the air pressure above the glottis to become the same as the pressure below the glottis, at which point the vocal folds will cease to vibrate. Thus, at least some of the acoustic variability can be explained by observing the timing of lip closure and release relative to aerodynamic factors such as air flow and intra-oral pressure.

The difference between an oral /d/ and a nasal /n/ provides another example of the articulatory perspective. Both of these sounds are produced by placing the tip of the tongue against the alveolar (maxillary) ridge and/or front teeth. Articulatorily, they are distinguished by whether or not the velar port is open. If it is, the result is a nasal sound. Acoustically, however, the difference is quite complex. In early articulatory synthesis, the degree of velar port opening was systematically varied in ordered steps — producing a stimulus continuum that ranged from acoustically perceptible /da/ to /na/ (Rubin et al. 1981) — simply by increasing velar opening from completely closed to the degree of opening necessary for an acceptable /na/. Producing a similar effect with acoustically based synthesis requires simultaneous control over a variety of parameters, including frequencies and bandwidths for three oral resonances, as well as the frequencies and bandwidths of nasal resonances and antiresonances. Another example of how simple variations in articulatory movement can result in complex changes in both acoustic and phonetic detail is provided in Section 8.

8
A Paradigm for Speech Research

A recurrent belief among speech researchers is that the listener extracts information about the production process itself from the speech signal — a process that, despite small differences in detail, is anatomically and physiologically constrained the same way for the entire species (Liberman et al. 1967; Mattingly and Liberman 1969; Fowler et al. 1980; Browman et al. 1984; Liberman and Mattingly 1985; Mattingly and Liberman 1988; Browman and Goldstein 1992; Fowler 1995). Thus, if the invariant aspects of production can be detected using analytical techniques, they can be used to determine the cognitive and neurophysiological underpinnings of speech behavior, as well as the mapping between the articulatory-to-acoustic domains. Furthermore, it seems increasingly likely that the anatomical and physiological constraints on speech production, while allowing extremely complex coordinated behaviors and acoustic output, may be quite similar in form to constraints affecting other biological behaviors. For example, the attempt to model the motion of the speech articulators may benefit from efforts to model other movements, such as locomotion, arm movement, and posture control.

Articulatory research of the past two decades has sought to describe the correspondence between phonetic units and the spatiotemporal behavior of various speech articulators, including attendant muscle activity, regulation of air supply, airflow through the larynx, and, to a lesser extent, airflow throughout the vocal tract. Whatever the actual subject of study, experiments have generally been designed to examine either the behavior of an articulator, or a set of articulators, over time and across different perturbing contexts (e.g., different vowel and consonant combinations, different speaking rates and intonation patterns, or experimentally controlled mechanical perturbations to the articulators). Measures in such studies have included characterizations of muscular activity, articulator motion and configuration, and resulting acoustic energy. The specific articulatory structures and combinations observed have been primarily determined by technological limitations on data acquisition and analysis. However, recent developments have facilitated simultaneous data collection from many sources, as outlined in Figure 6 (see also Borden and Harris 1984; Baken 1987; Fujimura 1988, 1990; Hardcastle and Marchal 1990; Kent et al. 1991; Bell-Berti and Raphael 1995).

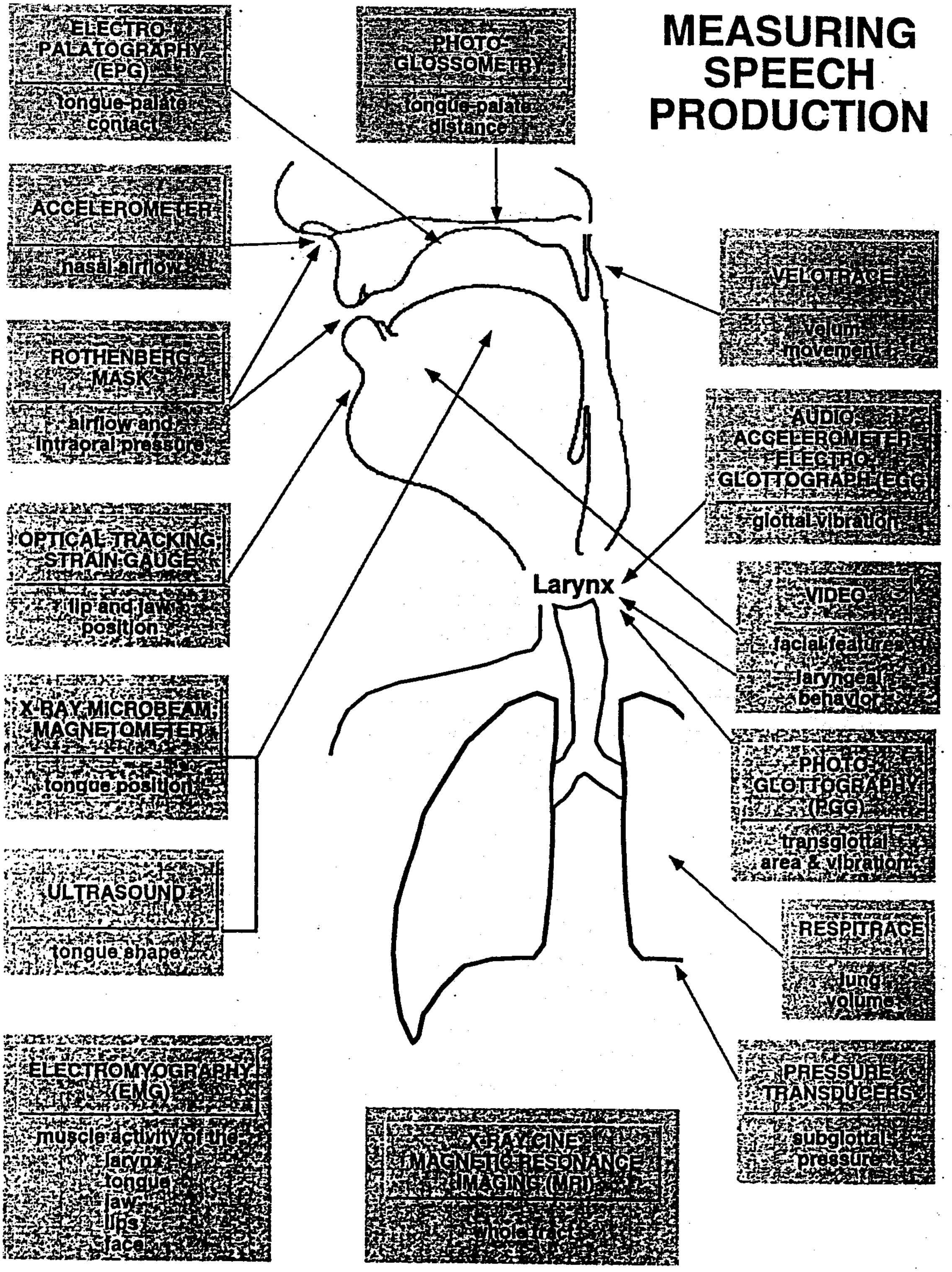


Fig. 6. Measuring speech production

Although by no means exhaustive, Figure 6 provides an overview of the various transduction devices and techniques that have been used to investigate vocal tract structures during speech production. Only a few of the structures involved in speech and nonspeech vocal production are readily accessible to noninvasive external view. Motions of the lips and jaw can be transduced optoelectronically (Sonoda and Wanishi 1982; Harrington et al. 1995; Vatikiotis-Bateson and Ostry 1995) or using strain gauges (Abbs and Gilbert 1973). Measurements of the lips and jaw, as well as other facial features, can also be made from video or film sources — examples in nonhuman vocal production can be found in the work of Hauser and colleagues (Hauser et al. 1993; Hauser and Schön Ybarra 1994), where frame-by-frame-video analysis was used to explore the role of mandibular position and lip configuration in rhesus monkey call production.

In human speech production, a number of other noninvasive techniques have been used. Air flow at the external boundaries of the nasal and oral cavities may be recorded using a Rothenberg mask (Rothenberg 1977). Glottal waveforms can be externally sensed using an electroglottograph, which measures impedance changes as the vocal folds open and close (Fourcin 1974, 1981; Kelman 1981; Rothenberg 1981), or an accelerometer, a transducer used to measure vibrations on the body surface (Askenfelt et al. 1980). In both cases, the device is strapped onto the neck in the vicinity of the thyroid cartilage. Lung volume can be measured using a spirometer (Beckett 1971) or a body plethysmograph (Hixon 1972) while the contributions of the ribcage and abdominal cavities to lung volume change can be evaluated using magnetometers (Mead et al. 1967), mercury strain gauges (Baken and Matz 1973), and inductive plethysmographs (Sackner 1980).

All other techniques included in the figure are to some extent invasive and require cooperation of the subject, particularly in the placement of transduction devices and sensors. Using a flexible fiber-optic endoscope, the larynx can be illuminated for video and photoglottographic recording of the laryngeal structures and transglottal areas (Sawashima et al. 1970; Fujimura 1977; Sawashima 1977; Fujimura et al. 1979). It is also possible to place miniature pressure transducers above and below the glottis for measurement of supra- and sub-glottal pressure (Cranen and Boves 1985; Gelfer et al. 1987).

Observation of the tongue, the most versatile and complex speech articulator, has been the most difficult of all. Optoelectronics, electromagnetic inductance, ultrasound, and X-ray imaging are currently available methods used to measure various aspects of tongue movement. For example, photoglossometry is an optoelectronic technique that uses reflection to measure the distance between the tongue surface and points on the hard palate (Chuang and Wang 1975). Electropalatography (Hardcastle 1972; Fletcher et al. 1975; Recasens 1984; Hardcastle et al. 1991) measures the pattern of contact between the tongue and the hard palate. The X-ray microbeam system tracks sagittal position of radio-opaque pellets on the surface of the tongue, lips, and jaw (Kiritani et al. 1975; Nadler and Abbs 1988; Westbury 1994). Electromagnetic techniques (e.g., magnetometers) can be used to recover similar information through transduction of field fluctuations at multiple points on the various articulator surfaces (Hixon 1971a,b; Schönle et al. 1987; Perkell et al. 1988, 1992; Tuller et al. 1990; Löfqvist et al. 1993; Löfqvist and Gracco 1994). Ultrasound has been used to acquire point-specific tongue data as well (Keller and Ostry 1983; Kaburagi and Honda 1994), but is used primarily to provide dynamic views of the tongue surface and other soft tissue structures (Morrish et al. 1985; Stone et al. 1988). While no system has yet surpassed the high resolution, sagittal view of the entire vocal apparatus provided by cineradiography (Perkell 1969; Subtelny et al. 1972; Wood 1979), the recently developed magnetic resonance imaging (MRI) technique is very promising (Baer et al. 1991; Moore 1992; Tiede 1993; Dang et al. 1994; Rubin et al. 1995). Although MRI applications have been limited to imaging the static vocal tract, improvements in scan rates and image-enhancement techniques may soon allow highly detailed, three-dimensional images of the vocal tract and surrounding structures during active speech production.

The ability to record various signal combinations in synchrony compensates substantially for the individual limitations of the available time-varying measurement techniques. There is a basic tradeoff between techniques that make rapid and accurate 'fleshpoint' measures of particular vocal tract structures (for instance, using tiny pellets placed on the tongue surface or mandible in the case of the X-ray microbeam, or markers placed on the lips and jaw in the case of optoelectronic position sensing devices), and those that provide more global views of the vocal tract (but with poorer spatiotemporal resolution), such as ultrasound or MRI. For example, while ultrasound has scan times that are short enough (approximately 35 ms) to allow tracking of the relevant motions of the tongue body, it cannot capture tongue-tip motion during production of /d,t,n/, where the tongue may contact the maxillary arch for as little as 15 ms. Ultrasound transduction of tongue-tip gestures is further complicated by the requirement that there be only one air—tissue boundary between the externally mounted transmitter/transducer and the articulator surface of interest. When raised to the maxillary arch, the tongue tip usually causes an additional air cavity to appear between the underside of the tongue and the mouth floor. However, when ultrasound is combined with fast transduction systems, such as the X-ray microbeam or magnetometer, a much better picture of the tongue's activity emerges (e.g., Stone 1990, 1991). Used together, then, the wide variety of transduction devices currently available makes it possible to assess the dynamic interaction of laryngeal and supralaryngeal structures at biomechanical, neurophysiological, and acoustic levels of analyses.

A multitude of articulatory studies have been conducted using these various techniques. Although many different issues have been addressed in this body of work (see Levelt 1989 for a review), the invariance issue has been fundamental. Many studies have attempted to identify the articulatory characteristics of different phonemes using experimental designs that manipulate stress and speaking rate (e.g., Kuehn and Moll 1976; Gay 1981), or phonetic context (e.g., Sussman et al. 1973). Stress-rate studies have been used to identify articulatory attributes of the phoneme that are independent of the changes induced when non-phonetic factors are varied. Experiments that vary the phonetic context (e.g., /aba, ibi, ubu/) are similar in that they allow the perturbing effects of the context to be distinguished from the inherent, and presumably stable, characteristics of the target phoneme (/b/ in this example). Another direction taken in the search for invariance has been to demonstrate that the articulators act in a flexible but highly coordinated fashion in achieving specific phonetic goals. These studies have used mechanically induced perturbations, either to reduce the number of articulators involved in a task (for instance, in bite-block and braking studies in which mandible position is fixed; e.g., Folkins and Abbs 1975; Lindblom et al. 1979), or to severely limit an articulator's contribution (for instance, through dynamic perturbation of the lips or jaw; e.g., Abbs and Gracco 1983; Kelso et al. 1984; Gracco and Abbs 1985; Saltzman et al. 1995).

Normal production processes always involve multiple articulators, even if some of the structures may not be moving. Thus, although the relative contributions of the two lips and the jaw may vary in bilabial productions (e.g., the final /b/ in baeb), the acoustic result is roughly the same in each case. Perturbing the system, for instance by removing the contribution of an articulator such as the jaw and examining kinematic and physiological effects, demonstrates that flexible and rapid compensatory effects act to prevent any loss of intelligibility in the speech signal. Because recording can be synchronized in the data channels of interest (including the perturbation delivery signal), the timing of articulatory events at both neuromuscular and kinematic levels of observation can be precisely examined (e.g., Tuller et al. 1982, 1983).

Although factors such as stress and speaking rate have fairly consistent articulatory laryngeal and supralaryngeal correlates, variability both within and across speakers is nonetheless quite high. One way to work around much of this variability is to focus on the relations among kinematic variables instead of on the individual measures themselves (e.g., Ostry et al. 1983; Kelso et al. 1985). For example, the relation between peak velocity and movement amplitude for a given articulator is quite stable and linear across most of its range of motion. At the same time, changes of stress and speaking rate are clearly marked by local changes in the relationship between these measures. Borrowing from the study of other biological movement systems, such as limb motion, researchers have begun to model such patterns of behavior as second-order mechanical systems. In this technique, the dynamic parameters of such systems (e.g., mass, stiffness, and viscosity) can be inferred from the relations among kinematic observables. Among other things, the approach promises the possibility of adducing stable (invariant) values of dynamic parameters from variable kinematic measures — values that can be compared across articulatory structures and many speaking contexts, including different languages (Vatikiotis-Bateson and Kelso 1993).

Figure 7 illustrates a qualitative method for discerning structure in the continuous kinematics of recurrent behavior. In this example, position and instantaneous velocity are plotted as trajectories in *phase space*, which provides a graphical description of the relation of the state (or phase) variables of dynamic systems. The result is a two-dimensional space where time is implicit in the continuity of consecutive data points but is not shown on a separate axis (see Abraham and Shaw 1982, 1987). Furthermore, such qualitative assessment can be used to direct subsequent quantitative analysis. Thus, even though gross effects of stress can be seen in the left panel of the figure as the roughly alternating sequence of larger and smaller movements, it is quite difficult to interpret the significance of these individual patterns of articulator position and instantaneous velocity, much less any relationships between them. But, when the two variables are plotted in phase space, as in the right panel of the figure, their continuous correlation can be seen in the stability of the trajectory shapes associated with the repetitive syllable sequence. Certain aspects of their variability also become readily apparent. For example, the phase portraits show motion of the articulators to be less variable during production of the consonant (top right) than of the vowel (bottom right). It can also be seen that the correlation between velocity and position is different for the two articulators, particularly during the closing phase of the movement cycle. The phase portraits reveal that there is greater tendency for covariation of movement amplitude and peak velocity for the jaw alone than for the lower lip — whose motion includes that of the jaw.

The phase-space representation also provides a means for considering distinctive differences in interarticulator timing. In Figure 8, for example, the relative timing of upper-lip movement toward closure for the second /b/ in bapab and the vowel—vowel movement cycle of the jaw is expressed in phase angle (Tuller and Kelso 1984; cf. Nittrouer et al. 1988). Across a range of available data (i.e., including changes in medial consonant identity, syllable stress, and overall speaking rate), the latency of upper-lip movement onset for the medial consonant has been found to be linearly proportional (Kelso et al. 1986a, 1986b), or nearly so (Nittrouer et al. 1988), to the period of the jaw-motion cycle associated primarily with production of the preceding vowel. Furthermore, phase-angle analysis of these data has demonstrated reliable differences for the medial consonants tested (Kelso et al. 1986a).

Thus, phase-angle analysis provides a precise means of transforming an articulatory database involving many dimensions of variability into a more functionally relevant form with fewer extraneous influences. Similar approaches are applicable across all articulator systems (Saltzman and Munhall 1989; Löfqvist 1990; Tuller and Kelso 1995). Examples include the coordination of laryngeal and supralaryngeal structures, such as the lips (Munhall et al. 1986; Munhall and Löfqvist 1992) or the tongue (Manuel and Vatikiotis-Bateson 1988), and functional coordination among various supralaryngeal articulators, such as coupling of the tongue and lip (Faber 1989) or the tongue and jaw (Stone and Vatikiotis-Bateson 1995).

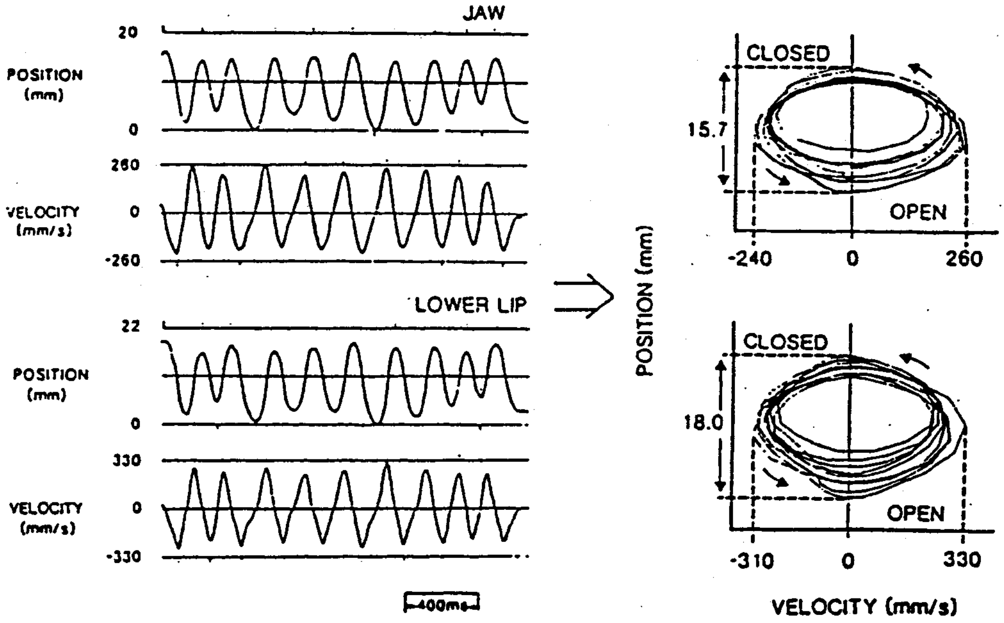


Fig. 7. *Left* Position and velocity over time of jaw and lower lip light-emitting diodes (LEDs) for sentences produced with reiterant /ba/ at a normal rate. *Right* Corresponding phase plane trajectories. *CLOSED* denotes the highest position achieved for /b/ and *OPEN* the lowest position for /a/. (From Kelso et al. 1985, ©1985 Acoustical Society of America)

9
A Model of the Human Vocal Tract

Although there is great interest in studying the speech production process, some of the methods discussed in the Section 8 place practical limits on the amount of data that can be gathered and analyzed. In addition, speakers cannot exercise the degree of control over their articulators needed for certain studies of the contributions of individual articulators. Paralleling the method for studying speech production and perception that uses speech synthesized from acoustic parameters as a fundamental tool, we use an articulatory synthesis (ASY) system at Haskins Laboratories that synthesizes speech through control of articulatory instead of acoustic variables (Mermelstein 1973; Rubin et al. 1981). ASY is designed for studying the linguistically and perceptually significant aspects of articulatory events. It allows quick modification of a limited set of key parameters that control the positions of the major articulators: the lips, jaw, tongue body, tongue tip, velum, and hyoid bone (whose position determines larynx height and pharynx width). Any particular set of parameter values provides a description of vocal tract shape that is adequate for research purposes, and that incorporates both individual

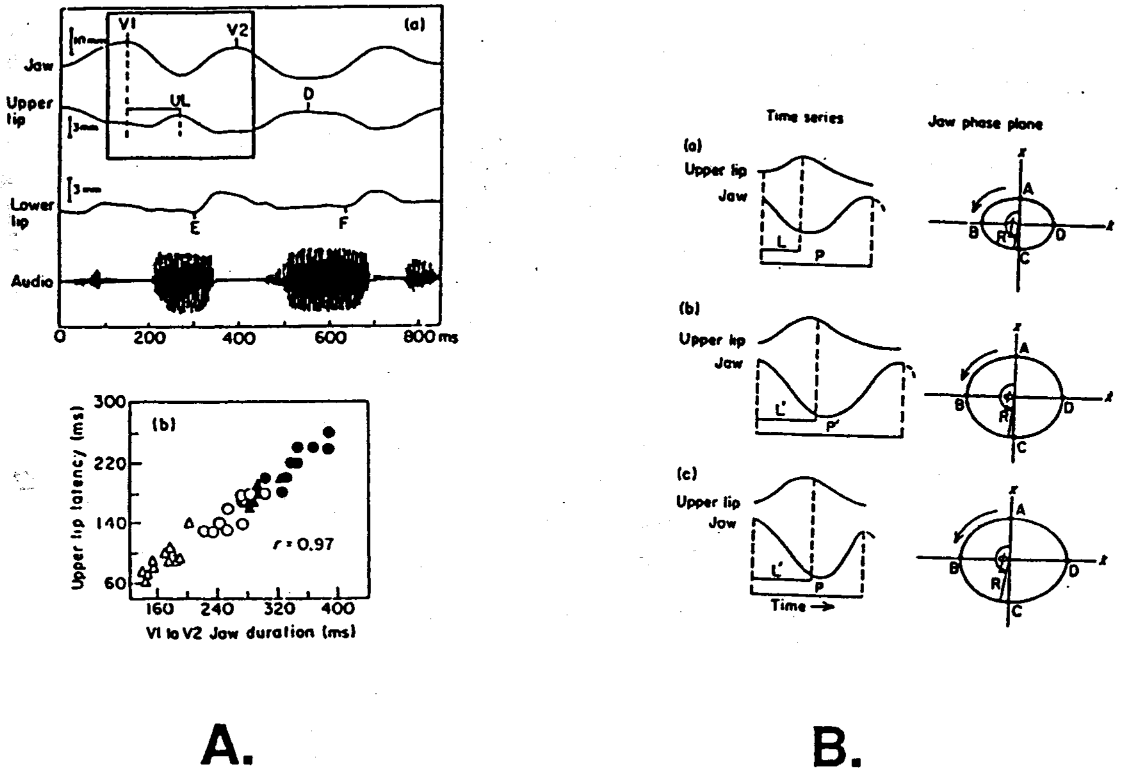


Fig. 8. A. a Movements of the jaw, upper lip, and lower lip corrected for jaw movement, and the acoustical signal, for one token of /ba'pab/. Articulator position (*y axis*) is shown as a function of time. Onsets of jaw and lip movements are indicated (empirically determined from zero crossings in the velocity records). b Timing of upper lip associated with /p/ production as a function of the period between successive jaw lowerings for the flanking vowels for one subject's productions of /ba#pab/●, Slow rate, first syllable stressed; ○, slow rate, second syllable stressed; ▲, fast rate, first syllable stressed; △, fast rate, second syllable stressed B. Left: time series representations of idealized utterances. Right: corresponding jaw motions, characterized as a simple mass spring and displayed on the "functional" phase plane (i.e. position on the vertical axis and velocity on the horizontal axis). (a), (b) and (c) represent three tokens with vowel-to-vowel periods (*P and P'*) and consonan latencies (*L and L'*) that are not linearly related. Phase position of upper lip movement onset relative to the jaw cycle is indicated. (From Kelso et al. 1986a, used with permission)

articulatory control and links among articulators. Additional input parameters include excitation (the sound source) and movement-timing information. An important aspect of this model's design is that speech sounds used in perceptual tests can be generated through by varying the timing or position parameters. Another very important aspect of the system is that the synthesis procedure is fast enough to make interactive on-line research practicable.

Figure 9 shows a midsagittal view of the ASY vocal tract in which the six key articulators are labeled. These articulators can be grouped into two major categories: those whose movements are independent of the movements of other articulators (the jaw, velum, and hyoid bone), and those whose movements are dependent on the movements of other articulators (the tongue body, tongue tip, and lips). The articulators in the second group normally move when the jaw moves. In addition, the tongue tip can move relative to the tongue body. Individual gestures can thus be separated into components arising from the combined movement of several articulators. For example, the lip-closing gesture used in the production of the utterance /aba/ is a combined movement of the jaw and lips. Movements of the jaw and velum have one degree of freedom, while all others have two degrees of freedom. Movement of the velum has two effects. It alters the shape of the oral branch of the vocal tract and, in addition, changes the size of the coupling port to the nasal tract, which is fixed.

An overview of the steps involved in articulatory synthesis using our model is provided in Figure 10. Once the articulator positions have been specified (see below), the midsagittal outline is determined and can be displayed. Cross-sectional areas are cal-

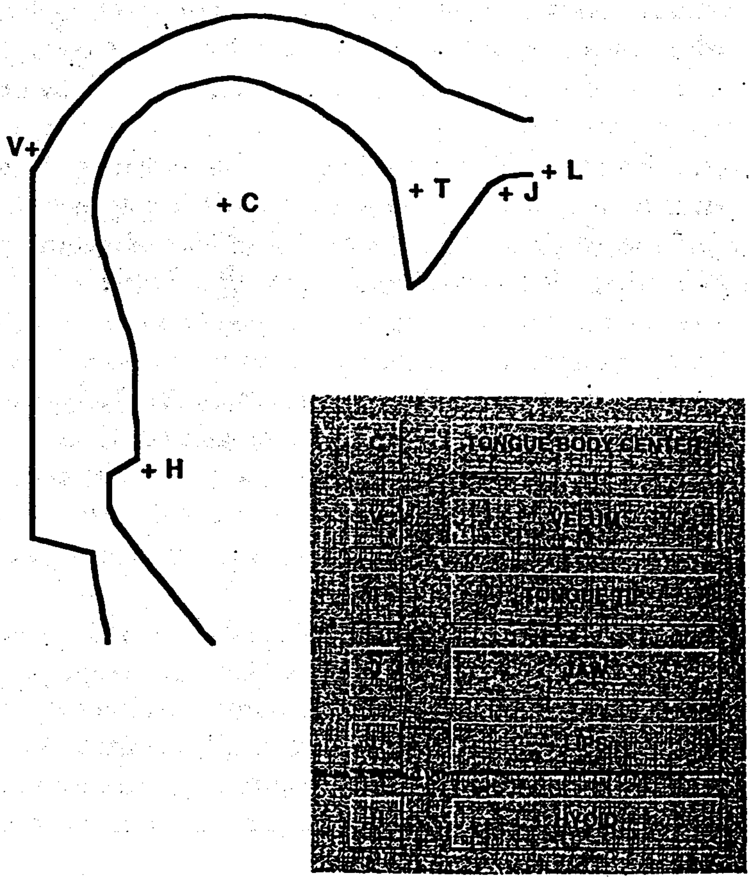


Fig. 9. Haskins Laboratories ASY vocal tract outline with key parameters labeled

culated by superimposing a grid structure, aligned with the maxilla in the outline, and computing the points of intersection of the outline and the grid lines. The resolution of this grid is variable, within certain limits. In general, parallel grid lines are set 0.25 cm apart and radial lines occur at 5° intervals. Sagittal cross-dimensions are calculated and converted to cross-sectional areas, using different formulas for estimating the shape in the pharyngeal, oral, and labial regions. These area values are then smoothed and approximated by a sequence of uniform tubes of fixed length (0.875 cm). The number of area values is variable because the overall length of the tract varies with both hyoid height and degree of lip protrusion. Improvements in the representation of the third dimension of vocal tract shape are eventually expected, and will be based on data from MRI of the vocal tract (Baer et al. 1991).

Once the area values have been obtained, the corresponding acoustic transfer function is calculated using a technique based on the model of Kelly and Lochbaum (1962), which specifies frequency-independent propagation losses within sections and reflections at section boundaries. Nonideal terminations at the glottis, lips, and nostrils are accurately modeled. However, the effects of other variables, such as tissue characteristics of the vocal-tract walls, are accounted for by introducing lumped-parameter elements at the glottis and within the nasal section.

In the interest of computational efficiency, and because the synthesizer was designed as a research tool to provide rapid feedback about changes in articulatory configuration, a number of compromises have been made in the details of the model. For example, acoustic excitation of the vocal tract transfer function is most commonly specified as an acoustic waveform, rather than through simulation of the physiological and aerodynamic factors of phonation. In this approach, control over the shape of individual glottal waveform pulses is limited to two parameters (Rosenberg 1971): the *open quotient* (i.e., the *duty cycle*, or relative durations of the open and closed portions of the glottal cycle) and the *speed quotient* (i.e., the ratio of rise-time to fall-time during the open portion). The fricative source is simulated by inserting shaped random noise anterior to the place of maximum constriction in the vocal tract. Acoustic output is obtained by supplying the glottal or fricative excitation as input to the appropriate acoustic transfer function, implemented as a digital filter.

Greater accuracy is achieved in the phonatory model through using a fully aerodynamic simulation of speech production that explicitly accounts for the propagation of sound along the tract (McGowan 1987, 1988). Such an approach provides a number of benefits, including more accurate simulation of voiced and fricative excitation, interaction of source and tract effects, and the effects of side branches and energy losses. However, it can result in slower overall calculation times. Because of such practical considerations, a choice of methods for calculating acoustic output has been implemented in the model.

Specification of the particular values for the key articulators can be provided in a number of ways. In the simplest approach, a display of the midsagittal outline of the vocal tract can be directly manipulated on-screen by moving one or more of the key articulators. The tract is then redrawn and areas and spectral values are calculated. This method of manual graphical modification continues until an appropriate shape is achieved. This shape can then be used for static synthesis of a vowel sound. Alterna-

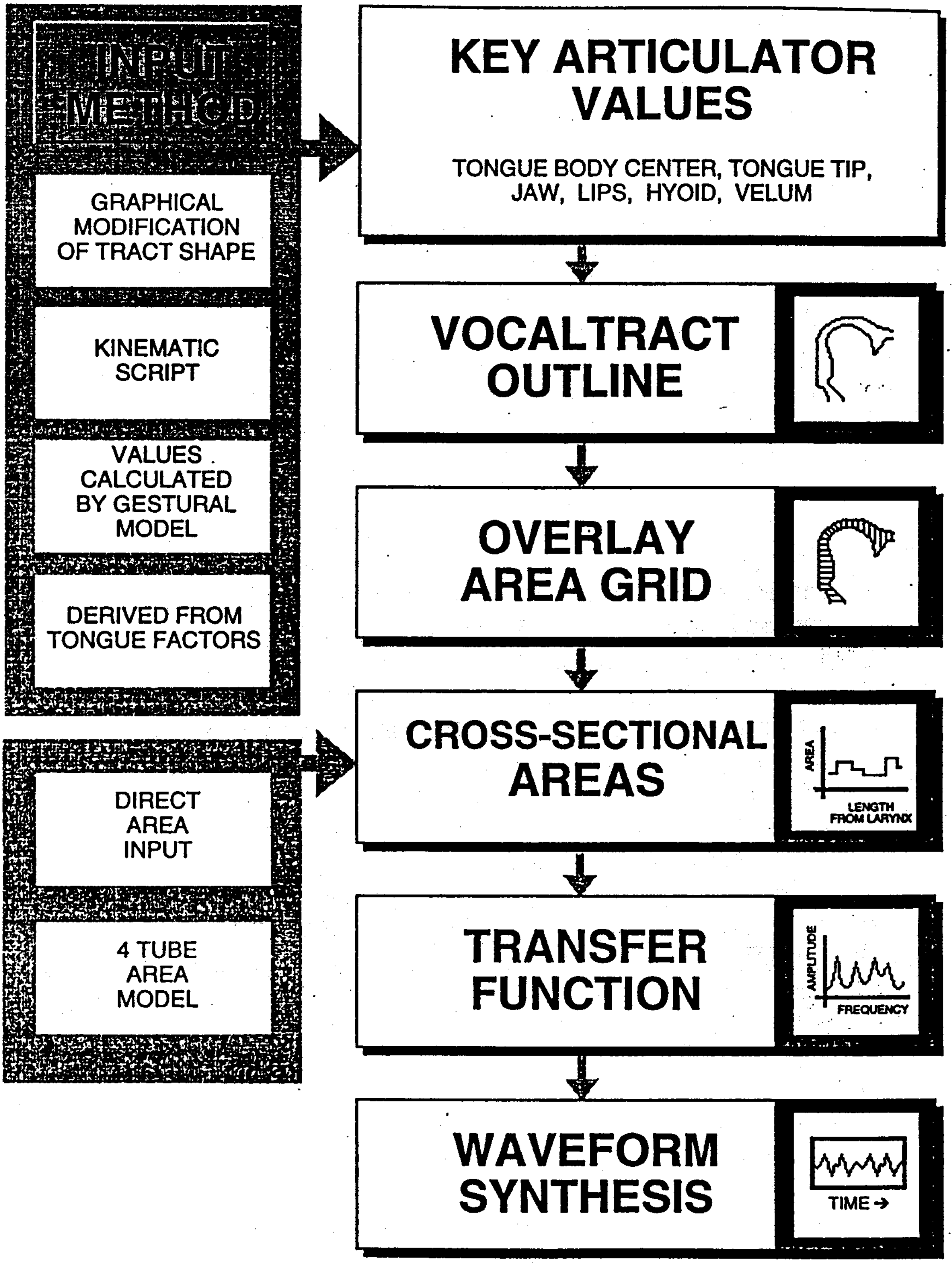


Fig. 10. The steps in articulatory synthesis

tively, individual shapes can be deposited in a table that will later be used as the basis for a "script" that specifies the kinematics of a particular utterance.

Using the method of kinematic specification, the complex acoustic effects of simple articulatory changes can be illustrated. The top of Figure 11 shows midsagittal vocal tract outlines for the major transition points (key frames) in the simulation of the articulations of four utterances: /bænænə/, /bændænə/, /bædnænə/, and /bædd ætə/ (i.e., "banana", "bandana", "bad nana", and "bad data", respectively). In this contrived example, the articulations of the four utterances are produced in very similar ways. The only parameter that varies is the timing of velar movement. The bottom of Figure 11 shows the degree of velar port size opening and the contrasting patterns of velar timing for the four different utterances. For utterance /bænnæə/ the velum is closed at the start, opens rapidly, and stays open throughout the rest of the utterance. In /bændænə/ the pattern of velar opening is similar, except that the velum closes and opens rapidly in the middle of the utterance, during the movement from /n/ to /d/. In /bædn ænə/ the velum stays relatively closed at the beginning of the utterance, opens during the movement from /d/ to /n/, and stays open throughout the rest of the utterance. Finally, in /bæddætə/ the velum stays relatively closed throughout the utterance.

All of these utterances have the same general form: C-V-CC-V-C-V, in which the initial consonant is /b/, and the vowel pattern is / æ,æ, ə /. With the exception of the velum, the articulators move in the same manner in each case. Note that the simple change in timing of velar opening in these four tokens results in considerable differences in the identities of the consonants that occur in the middle and near the end of the utterances. Simple changes in articulatory timing can also result in complex acoustic changes, as illustrated by the pseudo-spectrograms of the four utterances shown in Figure 11. These displays show only formant peaks, automatically extracted from the transfer functions of each utterance, where formant amplitude is indicated by height of the corresponding bar. Although these displays show little detail, a wide variety of acoustic differences can be seen in the four utterances.

This example illustrates one method, albeit a very schematized one, that can be used with the articulatory synthesis model to provide kinematic specifications. In real speech, production of such utterances is more variable and detailed. If desired, one can attempt to simulate these details by varying the model's input parameters on a pitch-pulse by pitch-pulse basis. Alternatively, specifications for vocal tract shape and kinematic trajectories can be calculated using an underlying dynamic model (see Sect. 10). In general, the latter approach is the technique most commonly used in our present simulations.

As mentioned above, the vocal tract model needs improvement in a number of ways. In addition to enhancements already achieved in its aeroacoustic simulation, changes are being made in its articulatory representation. The choice of particular key articulators described above has proven to be too limited. For example, additional tongue shape parameters are needed to simulate tongue bunching and for more accurate control of shape in the pharyngeal region. It would also be desirable to be able to fit the ASY model to a variety of head shapes, including those of females, infants, and, potentially, nonhuman primates. For this reason, a configurable version of ASY (CASY) is being developed that has increased flexibility in the model's internal linkages, the potential for adding new parameters, and a method for fitting vocal tract shape to actual X-ray or MRI data with concomitant adjustment of its internal fixed parameters.

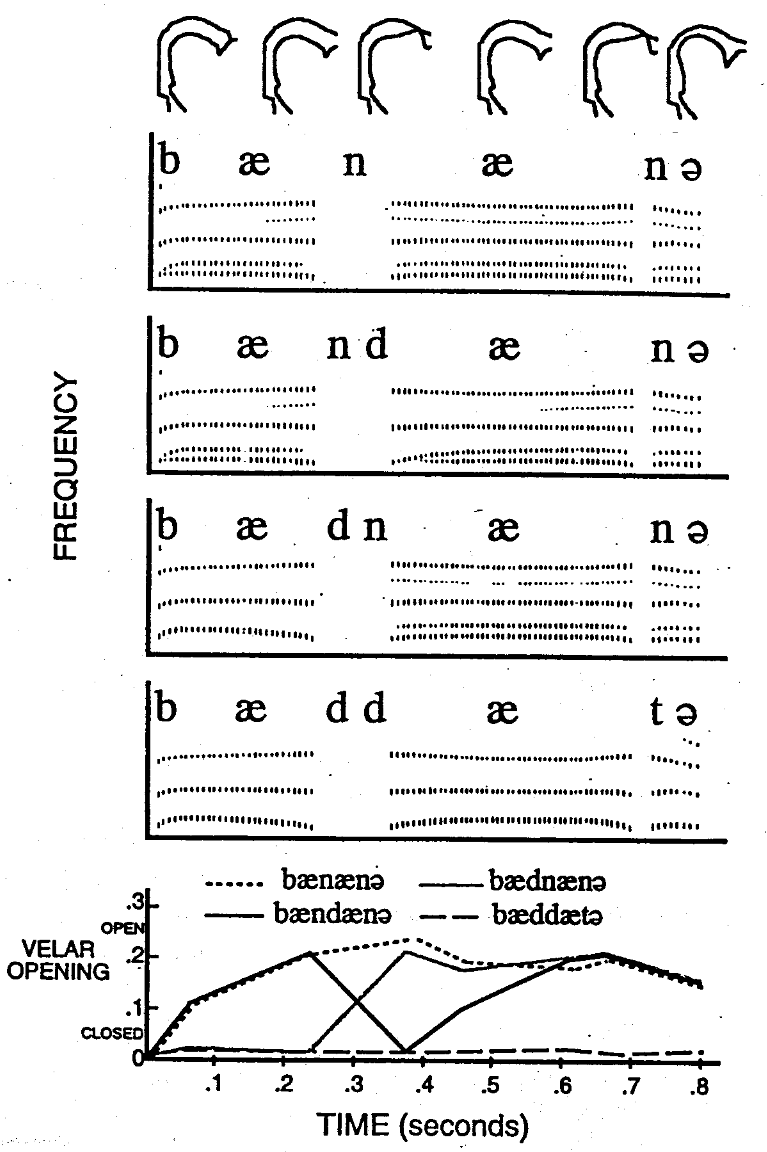


Fig. 11. Articulatory synthesis of /bæn ænə/, /bænd ænə/, /bædn ænə/, and /bædd ætə/

Finally, research is also underway to provide a more complete three-dimensional representation of vocal tract shape. A guiding principle of this approach is that actual physiological measurements and comparisons be used as the basis of improvements in both static and dynamic aspects of the simulation of speech production. In addition to our own interest in achieving more accurate physiological modeling, other researchers have focused on areas such as the control and dimensionality of jaw motion (Flanagan et al. 1990; Ostry and Munhall 1994; Vatikiotis-Bateson and Ostry 1995), as well as modeling soft-tissue structures, such as the tongue (Wilhelms-Tricarico 1995) and lips (Abry and Boë 1986; Badin et al. 1994; Benoît et al. 1994; Guiard-Marigny et al., in press).

10 Gestural Modeling

Modeling the speech production process requires a detailed consideration not only of the static anatomic and physiological aspects of the system, but also of how it changes over time. The speech articulators are continually in motion, producing a varying acoustic stream. The perceiver, in turn, is sensitive to both the local details of the resulting acoustic pattern and the global characteristics of change (Remez et al. 1981). In general, a greater emphasis has been placed on studying the static rather than the time-varying aspects of speech events. However, at Haskins Laboratories there has been a long-standing interest in the gestural basis of speech production and its relationship to perception (Liberman et al. 1967; Mattingly and Liberman 1969; Liberman and Mattingly 1985; Fowler 1995). Some of the techniques described in Section 8 are useful for examining the kinematics of the speech articulators. Theoretical approaches to studying action systems have also pointed out the necessity and desirability of examining the dynamic system that underlies these kinematic patterns (Bernstein 1967; Fowler 1977, 1984; Turvey 1977; Kelso et al. 1986a; Tuller and Kelso 1995).

Over the past several years, a computational model has been developed at Haskins Laboratories that combines these intersecting concerns in the form of a tool for representing and testing a variety of theoretical hypotheses about the dynamics of speech gestures and their coordination (Browman et al. 1984; Browman and Goldstein 1985). This approach merges a phonological model, based on gestural structures (Browman and Goldstein 1986, 1989, 1990, 1992), with an approach called *task dynamics* (see below) that characterizes speech gestures as coordinated patterns of goal-directed articulator movements. At the heart of both of these approaches is the notion of a gesture, which is considered in this context to be the formation of a constriction in the vocal tract by the organized activity of an articulator or set of articulators. The choice of gestural primitives is based upon observations of functional units in actual production. These models attempt to reconcile the linguistic hypothesis that speech involves an underlying sequence of abstract, context-independent units with "the empirical observation of context-dependent interleaving of articulatory movements" (Saltzman and Munhall 1989, p 333). The focus in this case is on discovering the regularities of gestural patterning and how they can be specified (see also Perrier et al. 1991; Shirai 1993 Kröger et al. 1995).

The computational model has three major components. First, a gesturally based phonological component (the linguistic-gestural model) provides, for a given utterance, a "gestural score" which consists of specifications for dynamic parameters for the set of speech gestures corresponding to the input phonetic string (Browman et al. 1986), and a temporal activation interval for each gesture, indicating its onset and offset times. These intervals are computed from the gesture's dynamic parameters in combination with a set of phasing principles that serves to specify the temporal patterning among the gestural set (Browman and Goldstein 1990). Second, the task-dynamic model is used to compute coordinated articulator movements from the gestural score in terms that are appropriate for the ASY vocal tract model. This model, in turn, allows computation of the speech waveform from these articulatory movements. An example of such a gestural score, for the utterance [p^{h}am], can be seen in Figure 12. This figure shows the periods of gestural activation (filled boxes) and trajectories generated during simulations (solid lines) for the four *tract variables* (see below) that are controlled in the production of this utterance: velic aperture, tongue body constriction degree, lip aperture, and glottal aperture.

The task-dynamic model used in this computational system has proved useful for describing the sensorimotor control and coordination of skilled activities of the limbs, as well as the speech articulators (Kelso et al. 1985; Saltzman 1986; Saltzman and Kelso 1987; Saltzman et al. 1987; Saltzman et al. 1988a, 1988b; Saltzman and Munhall 1989; Fowler and Saltzman 1993). For a particular given gesture, the goal is specified in terms of independent task dimensions, called *tract variables*. Each tract variable is associated with the specific set of articulators whose movements determine the value of that variable. For example, one such tract variable is lip aperture (LA), which corresponds to the vertical distance between the lips. Three articulators can contribute to changing LA: the jaw, the upper lip, and the lower lip. The standard set of tract variables in the computational model, and their associated articulators, can be seen in Figure 13. Recently, this set has been extended by incorporating aerodynamic and laryngeal components,

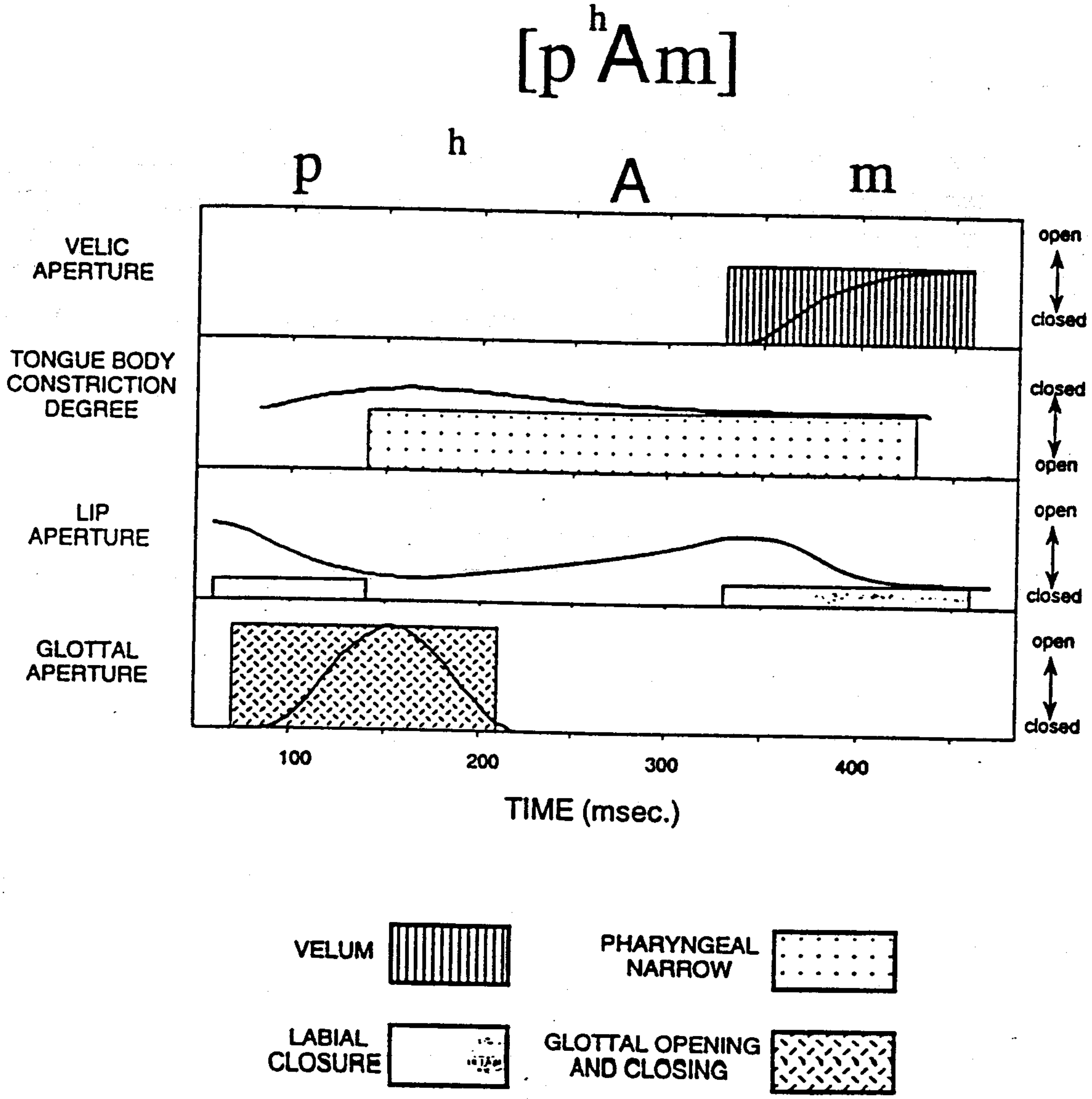


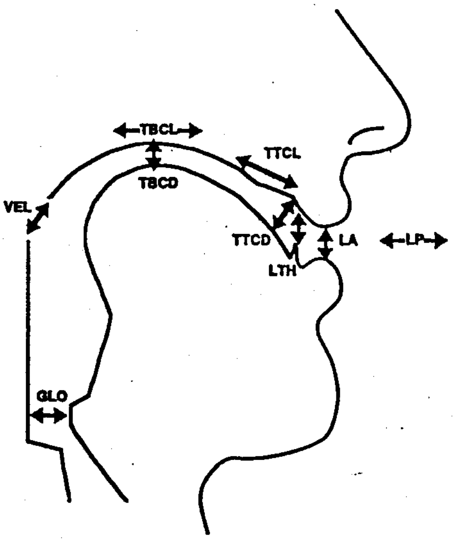
Fig. 12. Gestural score for the utterance [p^{h}am]

producing a more realistic model of source-related factors (McGowan and Saltzman 1995). Tract variables and articulators compose two sets of coordinates for gestural control in the model. In addition, each gesture is associated with its own *activation* coordinate, whose value reflects the strength with which the associated gesture "attempts" to shape vocal tract movements at any given point in time. Invariant gestural units are posited in the form of context-independent sets of dynamic parameters (e.g., protrusion target, stiffness, and damping coefficients, respectively, for the lips), and are associated with corresponding subsets of all three coordinate systems. Thus, the tract-variable and model articulator coordinates of each unit specify, respectively, the particular vocal tract constriction (e.g., bilabial) and the articulatory synergy that is affected directly by the associated unit's activation. Currently the model offers an intrinsically dynamic account of interarticulator coordination within the time span of single and temporally overlapping (coproduced) gestures, under normal conditions as well as in response to mechanical perturbations delivered to the articulators.

At the present stage of development, the task-dynamic model does not provide a dynamic account of intergestural timing patterns, even for simple speech sequences. Current simulations rely on explicit gestural scores to provide the timing patterns for gestural activation intervals in simulated utterances. While such explicitness facilitates research by enabling us to model and test current hypotheses of linguistically significant gestural coordination, an approach in which temporally ordered activation patterns are derived as implicit consequences of intrinsic *serial dynamics* would provide an important step in modeling processes of intergestural timing. Recent computational modeling of connectionist dynamic systems has investigated the control of sequences (e.g. Grossberg 1986; Tank and Hopfield 1987; Jordan 1989, 1990; Kawato 1989, 1991; Kawato et al. 1990). This serial-dynamics approach is well-suited for orchestrating the temporal activation patterns of gestural units in a dynamical model of speech production.

Connectionist models are also being applied to mapping relationships between acoustics and articulation. Investigators have long attempted to derive information about underlying articulation directly from the acoustic signal (Schroeder 1967; Atal et al. 1978; Wakita 1979; Levinson and Schmidt 1983; Kuc et al. 1985; Shirai and Kobayashi 1986; Sondhi and Schroeter 1987; Larar et al. 1988; Boë et al. 1992; Schroeter and Sondhi 1992; Badin et al. 1995; Beautemps et al. 1995), and a variety of connectionist or neural network methods are now being used for such mappings (e.g., Kawato 1989; Jordan 1989, 1990; Rahim and Goodyear 1990; Bailly et al. 1990,1991; Shirai and Kobayashi 1991; Papçun et al. 1992). In one example, Rahim and Goodyear (1990) trained a multilayer perceptron to map relationships between the power spectra of vowels and consonants and the parameters of a vocal tract model whose shape was specified by the areas of a fixed number of acoustic tube sections. For each member of the training set, input acoustic data consisted of 34 samples of the log power spectrum (from 100 to 4000 Hz). In an analysis-by-synthesis approach, a first-order gradient descent optimization procedure was used to minimize the spectral error between the target and synthesized spectra by adjusting the area values to reduce the acoustic mismatch. This mapping technique provided an efficient method for deriving vocal tract synthesis values directly from acoustic data.

In a related approach, Bailly and colleagues (Bailly et al. 1990;1991) proposed a method of control for an articulatory synthesis model (Maeda 1979) based on the opti-



TRACT VARIABLES		ARTICULATORS INVOLVED
LP	lip protrusion	upper and lower lips, jaw
LA	lip aperture	upper and lower lips, jaw
LTH	lip-teeth height	jaw
TTCL	tongue tip constriction location	tongue tip, tongue body, jaw
TTCD	tongue tip constriction degree	tongue tip, tongue body, jaw
TBCL	tongue body constriction location	tongue body, jaw
TBCD	tongue body constriction degree	tongue body, jaw
VEL	velic aperture	velum
GLO	glottal aperture	glottis

Fig. 13. Task dynamic tract variables

mization approach for motor skill learning developed by Jordan (1988, 1989, 1990; see also Rahim et al. 1993). This modified version of Jordan's sequential network operated under certain constraints arising from the kinematic properties of the biological system being controlled and from the phonological task being simulated. Specifically, these constraints restricted the possible solutions that the feedforward multilayered perceptron could use to model the mapping between the production of vocalic gestures and the trajectories of the first three formants. An additional feature of the modeling approach was that it could generalize its movement pattern by interpolating new trajectories based on existing learned trajectories.

Similarly, a number of approaches are being used at Haskins Laboratories to model the development of the connection between articulation and acoustics. The interest is in studying how a neural network model comes to constrain the potential movements of a dynamically changing vocal tract as it "learns" the relationship between acoustics and vocal tract variables and/or gestural scores — a process that may be similar to the exploratory activities found in infants during speech development. Examples include work by McGowan (1994, 1995) and Hogden (Hogden et al. 1993;1996). McGowan has used genetic algorithms (which simulate crossover, mutation, and selection processes) in conjunction with the task-dynamic model to recover articulatory movments from formant frequency trajectories. The result is an analysis-by-synthesis optimization procedure in which the fitness of each gestural score is based on how well its corresponding formants match those of the original signal to recover articulatory movements from formant frequency trajectories. Hogden's approach uses continuity constraints in the process of recovering the relative positions of simulated articulators from speech signals generated through articulatory synthesis.

Finally, connectionist models may be particularly well-suited to directly examining the dynamic properties of the musculo-skeletal system, where previous efforts to characterize the mapping between motor commands to muscles and the resulting behavior of speech articulators have been severely hampered in several ways. For instance, the muscles associated with speech articulation are typically either small and highly interconnected (e.g., tongue muscles), or are hard to monitor safely (e.g., the masseter — the large jaw-raising muscle). Thus, it is difficult to ascertain the muscle sources of electromyographic (EMG) records, which are themselves very complex. Despite the use of signal conditioning and numerical techniques, such as signal rectification and integration, smoothing (low-pass filtering), and ensemble-averaging over multiple trials, identification of "key" events has been restricted to visually observable landmarks in the signal, such as the onset or peak of EMG activity. Interpretation of this restricted set of events has relied primarily on statistical analysis of highly variable mean values, which must then be reliably correlated with other arbitrarily chosen, discrete events in the articulator movement behavior.

In contrast, artifical neural networks, for example, have been used to obtain the forward mapping between muscle activity and resulting articulator motion. Such muscle-based models are inherently dynamic because they estimate the muscle forces required to move the articulators. They also enable the entire EMG to be used as the "motor command input", rather than just those events that stand out visually on a display screen. Hirayama and colleagues (Hirayama et al. 1992, 1993, 1994; Vatikiotis-Bateson et al. 1991) have used real physiological data — articulator movements and EMG from muscle activity — to develop a preliminary model of speech production based on the articulatory system's dynamic properties. Using these EMG data, a neural network learned the forward-dynamics model of the articulators , i.e., mapping from current input (EMG information) and current state (position and velocity) to the next state. After training, the acquired model was incorporated into a recurrent network (i.e, one with feedback loops) that was found to successfully predict continuous articulator trajectories using the EMG signals as the motor command input. Simulations of articulator perturbation were then used to assess the properties of the acquired model.

This kind of modeling implicitly assumes a causal link between muscle activity and movement, rather than taking the more traditional and difficult approach of attempting to reject the implausible null hypothesis of the absence of a connection. Because the goal of the network is to formalize or "learn" that link, any degree of correlation between muscle and articulator behavior is useful in determining the proper coefficients or "weights" of the model equation. In a different, but related, approach, Wada and colleagues (Wada and Kawato 1995; Wada et al. 1995) have demonstrated that a tight coupling exists between formation of movement patterns and recognition of such patterns.

Using examples from cursive handwriting and estimation of phonetic timing in natural speech, they have developed a computational theory of movement pattern recognition that is based on a theory for optimal movement pattern generation and that may be widely applicable across movement systems.

11 Summary

The recent period of rapid evolution in the methods for studying speech has seen a growing interest in a variety of areas, including modeling of the speech production process, methods for examining the kinematics of the speech articulators (often with reference to underlying dynamic models), and initial steps exploring the relationship between acoustics and articulation using both connectionist and other kinds of models. Historically, the use of the sound spectrograph established a research paradigm in which invariant cues to phonetic segments were sought in the acoustic signal. In the last decade, this form of analysis has been supplemented by computer-based analysis systems using both frequency- and time-domain techniques to segregate the signal source from subsequent filtering effects, and to provide detailed information about filter characteristics.

In spite of this significant emphasis on understanding the physical signal, however, invariant links between acoustic and phonetic aspects of speech have proved elusive. Due both to these difficulties and for a variety of theoretical reasons, researchers have increasingly turned to the sound production process itself in attempting to explain the stability and efficiency of speech. In addition, the momentary and punctate cue-based acoustic approach is expanding to include event-based analysis techniques. As these theoretical approaches evolve, computer simulations are increasingly being used as vehicles to explore models of production and coordination, while the availability and use of transduction equipment is providing a means for improving the inherent realism of these models. Overall, a variety of approaches suggest that the future holds promise for successful study of speech production. For instance, both new and existing technologies are becoming more accessible, including X-ray microbeam, alternating field magnetic tracking, and imaging based on magnetic resonance and ultrasound techniques. The utility of dynamic approaches, including serial dynamics, is increasing and will continue to be driven by the intense interest that exists in neural network modeling and other optimization techniques. Advances in these areas will in turn be occurring in a context of increasing power and ease in acoustic and statistical analyses of signals, using desktop workstations or personal computers that are directly accessible both in the laboratory and in the field, rather than being located in remote mainframe computing centers.

It has been proposed that speech perception is a highly specialized process that can be differentiated from other forms of auditory perception (Whalen and Liberman 1987; Mattingly and Liberman 1988). A major factor in that specialization is the degree to which the speech perception and production systems are structurally and behaviorally linked. For example, the fact that articulatory gestures encode information in parallel provides an efficient means for overcoming inherent temporal constraints on resolution in both the auditory and the articulatory systems. Thus, the gestural/articulatory models and analyses discussed here are intended to illuminate at least some of the organizational principles underlying the intimate connection between perception and production in human speech systems. While little information is available as to whether analogous links may be involved in the acoustic communication processes of other species, production and perception processes are also clearly biologically specialized in nonhumans. Thus, while the choice of analysis, description, and modeling methods must entail careful consideration of *what* is important to a given animal in a particular ecological niche, the speech-related approaches discussed here may prove useful in providing both applicable technological innovations and theoretical models that can be adapted to the perception and production processes of nonhumans as well.

Acknowledgments. Preparation of this chapter was supported, in part, by NIH grants HD-01994 and DC-00121 and NSF grant BNS-8820099 to Haskins Laboratories. We thank Catherine Browman, Carol Fowler, Louis Goldstein, Vincent Gracco, Katherine S. Harris, Ignatius Mattingly, Richard McGowan, Robert Remez, Elliot Saltzman, Michael Studdert-Kennedy, and D. H. Whalen for their comments.

References

Abbs JH, Gilbert BN (1973) A strain gage transduction system for lip and jaw motion in two dimensions: design criteria and calibration data. J Speech Hear Res 16: 248–256

Abbs JH, Gracco VL (1983) Sensorimotor actions in the control of multimovement speech gestures. Trends Neurosci 6: 391–395

Abraham R, Shaw C (1982) Dynamics – the geometry of behavior, part I. Periodic behavior. Aerial Press, Santa Cruz,

Abraham R, Shaw C (1987) Dynamics: a visual introduction. In: Yates FE (ed) Self-organizing systems: the emergence of order. Plenum, New York, p 543

Abramson AS, Lisker L (1965) Voice onset time in stop consonants: acoustic analysis and synthesis. Proc 5th Int Congr of Acoustics, Liege

Abry C, Boë LJ (1986) Laws for lips. Speech Commun 5: 97–193

Alipour-Haghighi F, Titze IR (1991) Elastic models of vocal fold tissues. J Acoust Soc Am 90: 1326–1331

Andrew RJ (1976) Use of formants in the grunts of baboons and other nonhuman primates. In: Harnad SR, Steklis HD, Lancaster J (eds) Origins and evolution of language and speech. Ann NY Acad Sci 280: 673–693

Askenfelt A, Gauffin J, Sundberg J, Kitzing P (1980) A comparison of contact microphone and electroglottograph for the measurement of vocal fundamental frequency. J Speech Hear Res 23: 258–273

Atal BS (1985) Linear predictive coding of speech. In: Fallside F, Woods WA (eds) Computer speech processing. Prentice-Hall, London, pp 81–124

Atal BS, Hanauer SL (1971) Speech analysis and synthesis by linear prediction of the acoustic wave. J Acoust Soc Am 50: 37–655

Atal BS, Chang JJ, Mathews MV, Tukey JW (1978) Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer sorting technique. J Acoust Soc Am 63: 1535–1555

Badin P, Motoki K, Miki N, Ritterhaus D, Lallouache M-T (1994) Some geometric and acoustic properties of the lip horn. J Acoust Soc Jpn E15: 243–253

Badin P, Beautemps D, Laboissiere R, Schwartz JL(1995) Recovery of vocal tract geometry from formants for vowels and fricative consonants using a midsagittal-to-area function conversion model. J Phonet 23: 221–229

Baer T, Gore JC, Gracco LC, Nye P (1991) Analysis of vocal tract shape and dimensions using magnetic resonance imaging: vowels. J Acoust Soc Am 90: 799–828

Bailly G, Jordan M, Mantakas M, Schwartz JL, Bach M, Olesen M (1990) Simulation of vocalic gestures using an articulatory model driven by a sequential neural network. J Acoust Soc Am 87: S105

Bailly G, Laboissière R, Schwartz JL (1991) Formant trajectories as audible gestures: an alternative for speech synthesis. J Phonetics 19: 9–23

Baken RJ (1987) Clinical measurement of speech and voice. Little, Brown, Boston,

Baken RJ, Matz BJ (1973) A portable impedance pneumograph. Hum Commun 2: 28–35

Beautemps D, Badin P, Laboissière 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 Commun 16: 27–47

Beckett RL (1971) The respirometer as a diagnostic and clinical tool in the speech clinic. J Speech Hear Disord 36: 235–241

Bell AM (1867) Visible speech or self-interpreting physiological letters for the writing of all languages in one alphabet. Simpkin and Marshall, London

Bell-Berti F, Raphael LJ (eds) (1995) Producing speech: contemporary issues. For Katherine Safford Harris. AIP Press, New York

Benoît C, Lallouache T, Mohamadi T, Abry C (1992) A set of French visemes for visual speech synthesis. In: Bailly G, Benoît C (eds) Talking machines: theories, models and applications, Elsevier, Amsterdam, p 485

Benoît C, Mohamadi T, Kandel S (1994) Audio-visual intelligibility of French speech in noise. J Speech Hear Res 37: 1195–1203

Bernstein NA (1967) The coordination and regulation of movements. Pergamon, London

Boë LJ, Perrier P, Bailly G (1992) The geometric vocal tract variables controlled for vowel production: proposals for constraining acoustic-to-articulatory conversion. J Phonet 20:27–38

Borden GJ, Harris KS (1984) Speech science primer. Williams & Wilkins, Baltimore

Boring EG (1950) A history of experimental psychology, 2nd edn. Appleton-Century-Crofts, New York

Brigham EO (1974) The fast Fourier transform. Prentice Hall, Englewood Cliffs

Browman CP, Goldstein L (1985) Dynamic modeling of phonetic structure. In: Fromkin VA (ed) Phonetic linguistics. Essays in honor of Peter Ladefoged. Academic Press, New York, pp 35–53

Browman CP, Goldstein L (1986) Towards an articulatory phonology. Phonol Year 3: 219–252

Browman CP, Goldstein L (1989) Articulatory gestures as phonological units. Phonology 6: 201–251

Browman CP, Goldstein L (1990) Tiers in articulatory phonology, with some implications for casual speech. In: Kingston J, Beckman M (eds) Papers in laboratory phonology: I. Between the grammar and the physics of speech. Cambridge University Press, Cambridge, England, pp 341–376

Browman CP and Goldstein L (1992) Articulatory phonology: an overview. Phonetica 49: 222–234

Browman CP, Goldstein L, Kelso JAS, Rubin P, Saltzman E (1984) Articulatory synthesis from underlying dynamics. J Acoust Soc Am 75: S22–S23

Browman CP, Goldstein L, Saltzman E, Smith C (1986) GEST: a computational model for speech production using dynamically defined articulatory gestures. J Acoust Soc Am 80: S97

Capranica RR (1966) Vocal response of the bullfrog to natural and synthetic mating calls. J Acoust Soc Am 40: 1131–1139

Carterette E, Shipley C, Buchwald J (1979) Linear prediction theory of vocalization in cat and kitten. In: Lindblom B, Ohman S (eds) Frontiers of speech communication research, Academic Press, New York, pp 245–257

Carterette E, Shipley C, Buchwald J (1984) The speech of animals. In: Bristow G (ed) Electronic speech synthesis. Techniques, technology and applications. McGraw Hill, New York, pp 292–302

Chuang C, Wang W (1975) A distance-sensing device for tracking tongue configuration. J Acoust Soc Am Suppl 1 59: S11

Cochran WT, Cooley JW, Favin DL, Helms HD, Kaenel RA, Lang WW, Maling GC Jr, Nelson DE, Rader CM, Welch PD (1967) What is the fast Fourier transform? IEEE Trans Audio Electroacoust AU-15: 45–55

Cooper FS, Liberman AM, Borst JM (1951) The interconversion of audible and visible patterns as a basis for research in the perception of speech. Proc Natl Acad Sci USA 37: 318–325

Cranen B, Boves L (1985) Pressure measurements during speech production using semiconductor miniature pressure transducers: impact on models for speech production. J Acoust Soc Am 77: 1543–1551

Dang J, Honda K, Suzuki H (1994) Morphological and acoustical analysis of the nasal and the paranasal cavities. J Acoust Soc Am 96: 2088–2100

Faber A (1989) Lip protrusion in sibilant production. J Acoust Soc Am 86: S113

Fallside F (1985) Frequency-domain analysis of speech. In: Fallside F, Woods WA (eds) Computer speech processing. Prentice-Hall, London, pp 41–80

Fallside F, Woods WA (eds) (1985) Computer speech processing. Prentice-Hall, London

Fano RM (1949) The transmission of information. MIT RLE Tech Rep 65, Cambridge

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

Fay RR (1988) Hearing in vertebrates: a psychophysics databook. Hill-Fay, Winnetka, Illinois

Flanagan JL (1965) Speech analysis, synthesis, and perception. Springer Berlin Heidelberg New York

Flanagan J, Rabiner L (eds) (1973) Speech synthesis. Dowden, Hutchinson and Ross, Stroudsburg

Flanagan JR, Ostry DJ, Feldman AG (1990) Control of human jaw and multi-joint arm movements. In: Hammond GR (ed) Cerebral control of speech and limb movements. North-Holland, Amsterdam, pp 29–58

Fletcher S, McCutcheon M, Wolf M (1975) Dynamic palatometry. J Speech Hear Res 18: 812–819

Folkins JW, Abbs JH (1975) Lip and jaw motor control during speech: responses to resistive loading of the jaw. J Speech Hear Res 18: 207–220

Fourcin AJ (1974) Laryngographic examination of vocal fold vibration. In: Wyke B (ed) Ventilatory and phonatory control systems. Oxford University Press, New York, pp 315–333

Fourcin AJ (1981) Laryngographic assessment of phonatory function. In: Ludlow CL, Hart MO (eds) Proc Conf on the Assessment of vocal pathology, ASHA Rep 11, pp 116–127

Fowler CA (1977) Timing control in speech production. Indiana University Linguistics Club, Bloomingon

Fowler CA (1984) Current perspectives on language and speech production: a critical overview. In: Daniloff R (ed) Recent advances in speech, hearing and language, vol. 4. College-Hill Press, Boston, pp 195–278

Fowler CA (1995) Speech production. In: Miller J, Eimas P (eds) Speech, language and communication. Academic Press, New York, pp 29–61

Fowler CA, Saltzman E (1993) Coordination and coarticulation in speech production. Lang Speech 36: 171–195

Fowler CA, Rubin P, Remez RE, Turvey MT (1980) Implications for speech production of a general theory of action. In: Butterworth B (ed) Language production. Academic Press, New York, pp 373–420

Fry DB (1979) The physics of speech. Cambridge University Press, Cambridge, England

Fujimura O (1977) Stereo-fiberscope. In: Sawashima M, Cooper, FS (eds) Dynamic aspects of speech production. University of Tokyo Press, Tokyo, pp 133–137

Fujimura O (1988) Vocal fold physiology vol 2. Vocal physiology: voice production, mechanisms and functions. Raven Press, New York

Fujimura O (1990) Methods and goals of speech production research. Lang Speech 33: 195–258

Fujimura O, Baer T, Niimi S (1979) A stereo-fiberscope with a magnetic interlens bridge for laryngeal observation. J Acoust Soc Am 65: 478–480

Galaburda AM (1984) Anatomical asymmetries. In: Geschwind N, Galaburda AM (eds) Cerebral dominance: the biological foundations. Harvard University Press, Cambridge, MA, pp 11–25

Gay TJ (1981) Mechanisms in the control of speech rate. Phonetica 38: 148-158

Gelfer CE, Harris KS, Baer T (1987) Controlled variables in sentence intonation. In: Baer T, Sasaki C, Harris, K (eds) Laryngeal function in phonation and respiration. College-Hill Press, Boston, pp 422–435

Gracco VL, Abbs JH (1985) Dynamic control of the perioral system during speech: kinematic analyses of autogenic and nonautogenic sensorimotor processes. J Neurophysiol (Bethesda)54: 418–432

Gracco VL, Abbs JH (1988) Central patterning of speech movements. Exp Brain Res 71: 515–526

Grossberg S (1986) The adaptive self-organization of serial order in behavior: speech, language, and motor control. In: Schwab EC, Nusbaum HC (eds) Pattern recognition by humans and machines, vol 1. Academic Press, Boston, p 187

Guiard-Marigny T, Adjoudani A, Benoît C (1996) A 3D model of the lips and of the jaw for visual speech synthesis. In: Progress in speech synthesis, Springer Berlin Heidelberg New York

Hardcastle WJ (1972) The use of electropalatography in phonetic research. Phonetica 25: 197–215

Hardcastle WJ, Marchal A (eds) (1990) Speech production and speech modelling. Kluwer, Dordrecht

Hardcastle WJ, Gibbon F, Nicolaidis K (1991) EPG data reduction methods and their implications for studies of lingual coarticulation. J Phonetics 19: 251–266

Harrington J, Fletcher J, Roberts C (1995) Coarticulation and the accented/unaccented distinction: evidence from jaw movement data. J Phonetics 23: 305–322

Hauser MD (1991) Sources of acoustic variation in rhesus macaque vocalizations. Ethology 89: 29–46

Hauser MD (1992) Articulatory and social factors influence the acoustic structure of rhesus monkey vocalizations: a learned mode of production? J Acoust Soc Am 91: 2175–2179

Hauser MD (1996) Nonhuman primate vocal communication. In: Cochran M (ed) Handbook of acoustics. John Wiley, New York

Hauser MD, Fowler C (1992) Declination in fundamental frequency is not unique to human speech: evidence from nonhuman primates. J Acoust Soc Am 91: 363–369.

Hauser MD, Evans CS, Marler P (1993) The role of articulation in the production of rhesus monkey (*Macaca mulatta*) vocalizations. Anim Behav 45: 423–433

Hauser MD, Schön Ybarra M (1994) The role of lip configuration in monkey vocalizations: experiments using xylocaine as a nerve block. Brain and Language 46: 423–433

Hirayama ME, Vatikiotis-Bateson E, Kawato M, Jordan MI (1992) Forward dynamics modeling of speech motor control using physiological data. In: Moody JE, Hanson SJ, Lippmann RP (eds) Advances in neural information processing systems 4. Morgan Kaufman, San Mateo, p 191

Hirayama M, Vatikiotis-Bateson E, Kawato M (1993) Physiologically based speech synthesis using neutral networks. IEICE Trans E76-A: 1898–1910

Hirayama M, Vatikiotis-Bateson E, Kawato M (1994) Inverse dynamics of speech motor control. Adv Neural Inf Proc Syst 6: 1043–1050

Hixon TJ (1971a) Magnetometer recording of jaw movements during speech. J Acoust Soc Am 49: 104

Hixon TJ (1971b) An electromagnetic method for transducing jaw movements during speech. J Acoust Soc Am 49: 603–606

Hixon TJ (1972) Some new techniques for measuring the biomechanical events of speech production: one laboratory's experiences. ASHA Rep 7: 68–103

Hogden J, Löfquist A, Gracco V, Oshima K, Rubin P, Saltzman E (1993) Inferring articulator positions from acoustics: an electromagnetic midsagittal articulometer experiment. J Acoust Soc Am 94: 1764

Hogden J, Rubin P, Saltzman E (1996) An unsupervised method for learning to track tongue position from an acoustic signal. Bull Commun Parl 3:101–116

Hopp SL, Sinnott JM, Owren MJ, Petersen MR (1992) Differential sensitivity of Japanese macaques (*Macaca fuscata)* and humans (*Homo sapiens*) to peak position along a synthetic coo call continuum. J Comp Psychol 106: 128–136

Jakobson R, Fant G, Halle M (1963) Preliminaries to speech analysis. MIT Press, Boston

Jordan MI (1988) Supervised learning and systems with excess degrees of freedom. COINS Tech Rep . University of Massachusetts, Computer and Information Sciences, Boston, pp 99–127

Jordan MI (1989) Serial order: a parallel, distributed processing approach. In: Elman JL, Rumelhart DE (eds) Advances in connectionist theory: speech. Erlbaum, Boston, pp 44–93

Jordan MI (1990) Motor learning and the degrees of freedom problem. In: Jeannerod M (ed) Attention and performance, XIII. Erlbaum, Boston

Kaburagi T, Honda M (1994) A trajectory formation model of articulatory movements based on the motor tasks of phoneme-specific vocal tract shapes. Proc Int Conf on spoken language processing II (ICLSP 94), Acoustical Society of Japan, Yokohama, Japan, pp 579–588

Kawato M (1989) Motor theory of speech perception revisited from minimum torque-change neural network model. In: Proc 8th Symp on Future electron devices, 30–31 October, 1989, Tokyo

Kawato M (1991) Optimization and learning in neural networks for formation and control of coordinated movement. In: Meyer D (ed) Attention and performance, XIV. Erlbaum, Hillsdale, New Jersey

Kawato M, Maeda Y, Uno Y, Suzuki R (1990) Trajectory formation of arm movement by cascade neural network model based on minimum torque-change criterion. Biol Cybern 62: 275–288

Keller E, Ostry DJ (1983) Computerized measurement of tongue dorsum movement with pulsed-echo ultrasound. J Acoust Soc Am 73: 1309–1315

Kelly JL Jr, Lochbaum C (1962) Speech synthesis. In: Proc Stockholm Speech Commun Seminar, RIT, Stockholm

Kelman AW (1981) Vibratory pattern of the vocal folds. Folia Phoniatr 33: 73-99

Kelso JAS, Tuller B, Vatikiotis-Bateson E, Fowler CA (1984) Functionally specific articulatory cooperation following jaw perturbations during speech: evidence for coordinative structures. J Exp Psychol: Hum Percept Perform 10: 812–832

Kelso JAS, Vatikiotis-Bateson E, Saltzman EL, Kay B (1985) A qualitative dynamic analysis of reiterant speech production: phase portraits, kinematics, and dynamic modeling. J Acoust Soc Am 77: 266–280

Kelso JAS, Saltzman EL, Tuller B (1986a) The dynamical perspective in speech production: data and theory. J Phonetics 14: 29–59

Kelso JAS, Saltzman EL, Tuller B (1986b) Intentional contents, communicative context, and task dynamics: a reply to the commentators. J Phonetics 14: 171–196

Kent RD, Read C (1992) The acoustic analysis of speech. Singular Publishing Group, San Diego, California

Kent RD, Atal BS, Miller JL (eds) (1991) Papers in speech communciation: speech production. Acoustical Society of America, Woodbury, New York

Kirchner JA (1988) Functional evolution of the human larynx: variations among the vertebrates. In: Fujimura O (ed) Vocal physiology: voice production mechanisms, and functions. Raven Press, New York

Kiritani S, Itoh K, Fujimura O (1975) Tongue-pellet tracking by a computer-controlled x-ray microbeam. J Acoust Soc Am 57: 1516–1520

Klatt DH (1980) Software for a cascade/parallel formant synthesizer. J Acoust Soc Am 67: 971–995

Klatt DH (1987) Review of text-to-speech conversion for English. J Acoust Soc Am 82: 737–793

Koenig W, Dunn HK, Lacey LY (1946) The sound spectrograph. J Acoust Soc Am 18: 19–49

Kröger BJ, Schröder G, Opgen-Rhein C (1995) A gesture-based dynamic model describing articulatory movement data. J Acoust Soc Am 98: 1878–1889

Kuc R, Tuteur F, Vaisnys JR (1985) Determining vocal tract shape by applying dynamic constraints. ICASSP 95. Proc of the Int Conf on Acoustics, speech and signal processing, Tampa IEEE, New York, pp 1101–1104

Kuehn DP, Moll K (1976) A cineradiographic study of VC and CV articulatory velocities. J Phonetics 4: 303–320

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

Laitman JT, Crelin ES, Conlogue GJ (1977) The function of the epiglottis in monkey and man. Yale J Biol Med 50: 43–48

Larar JN, Schroeter J, Sondhi MM (1988) Vector quantization of the articulatory space. IEEE Trans Acoust, Speech Signal Process 36: 1812–1818

Larson CR (1988) Brain mechanisms involved in the control of vocalization. J Voice 2: 301–311

Levelt WJM (1989) Speaking: from intention to articulation. MIT Press, Cambridge

Levinson SE, Schmidt CE (1983) Adaptive computation of articulatory parameters from the speech signal. J Acoust Soc Am 74: 1145–1154

Liberman AM, Studdert-Kennedy M (1978) Phonetic perception. In: Held R, Leibowitz H, Teuber HL (eds) Handbook of sensory physiology, vol. 8 Perception. Springer, Berlin Heidelberg New York, p 143

Liberman AM, Mattingly IG (1985) The motor theory of speech perception revised. Cognition 21: 1–36

Liberman AM, Ingemann F, Lisker L, Delattre P, Cooper F (1959) Minimal rules for synthesizing speech. J Acoust Soc Am 31: 1490-1499

Liberman AM, Cooper FS, Shankweiler DP, Studdert-Kennedy M (1967) Perception of the speech code. Psychol Rev 74: 431–461

Lieberman P (1969) On the acoustic analysis of primate vocalizations. Behav Res Methods Instrum 5: 169–174

Lieberman P (1975) On the origins of language. MacMillan, New York

Lieberman P (1984) The biology and evolution of language. Harvard University Press, Cambridge

Lieberman P, Blumstein SE (1988) Speech physiology, speech perception, and acoustic phonetics. Cambridge University Press, New York

Lieberman PH, Klatt DH, Wilson WH (1969) Vocal tract limitations on the vowel repertoires of rhesus monkey and other nonhuman primates. Science 164: 1185–1187

Lindblom B, Lubker J, Gay T (1979) Formant frequencies of some fixed mandible vowels and a model of speech motor programming by predictive simulation. J Phonetics 7: 147–161

Lisker L (1957) Closure duration and the intervocalic voiced-voiceless distinction in English. Language 33: 42–49

Lisker L, Abramson AS (1967) Some effects of context on voice onset time in English stops. Lang Speech 10: 1–28

Löfqvist A (1990) Speech as audible gestures. In: Hardcastle W, Marchal A (eds) Speech production and speech modelling. Kluwer, Dordrecht, pp 289–322

Löfqvist A, Gracco VL (1994) Tongue body kinematics in velar stop production: influences of consonant voicing and vowel context. Phonetica 51: 52–67

Löfqvist A, Gracco VL, Nye PW (1993) Recording speech movements using magnetometry: one laboratory's experience. Proc ACCOR Worksh on Electromagnetic articulography in phonetic research. Forschungsberichte des Instituts fr Phonetik und Sprachliche Kommunikation der Universität München 31, pp 143–162

Luria AR (1975) The man with a shattered world. Penguin, London

Maeda S (1979) An articulatory model of the tongue based on a statistical analysis. J Acoust Soc Am 65: S22

Manuel S, Vatikiotis-Bateson E (1988) Oral and glottal gestures and acoustics of underlying /t/ in English. J Acoust Soc Am 84: S84

Markel JD, Gray AH (1976) Linear prediction of speech. Springer, Berlin Heidelberg New York

Massaro DW, Cohen MH, Gesi A, Heredia R, Tsuzaki M (1993) Bimodal speech perception: an examination across languages. J Phonet 21: 445–478

Mattingly IG, Liberman AM (1969) The speech code and the physiology of language. In: Leibovic KN (ed) Information processing in the nervous system. Springer, Berlin Heidelberg New York, p 97

Mattingly IG, Liberman AM (1988) Specialized perceiving systems for speech and other biologically significant sounds. In: Edelman GM, Gall WE, Cowan ME (eds) Auditory function: the neurobiological bases of hearing. Wiley, New York, pp 775–793

May B, Moody D, Stebbins W (1988) The significant features of Japanese macaque coo sounds: a psychophysical study. Anim Behav 36: 1432–1444

May B, Moody D, Stebbins W (1989) Categorical perception of conspecific communication sounds by Japanese macaques. J Acoust Soc Am 85: 837–847

McGowan R (1987) Articulatory synthesis: numerical solution of a hyperbolic differential equation. Haskins Laboratories Status Report on Speech Research SR-89/90, New Haven pp 69–79

McGowan R (1988) An aeroacoustic approach to phonation. J Acoust Soc Am 83: 696–704

McGowan R (1994) Recovering articulatory movement from formant frequency trajectories using task dynamics and a genetic algorithm: preliminary model test. Speech Commun 14: 19–48

McGowan R (1995) Recovering task dynamics from formant frequency trajectories: results using computer "babbling" to form an indexed data base. In: Bell-Berti F, Raphael LJ (eds) Producing speech: contemporary issues. For Katherine Safford Harris. AIP Press, New York, pp 489–504

McGowan R, Saltzman E (1995) Incorporating aerodynamic and laryngeal components into task dynamics. J Phonet 23: 255–269

Mead J, Peterson N, Grimby C, Mead J (1967) Pulmonary ventilation measured from body surface movements. Science 156: 1383–1384

Mermelstein P (1973) Articulatory model for the study of speech production. J Acoust Soc Am 53: 1070–1082

Moore CA (1992) The correspondence of vocal tract resonance with volumes obtained from magnetic resonance images. J Speech Hearing Res 35: 1009–1023

Morrish K, Stone M, Shawker T, Sonies B (1985) Distinguishability of tongue shape during vowel production. J Phonetics 13: 189–203

Müller J (1848) The physiology of the senses, voice and muscular motion with the mental faculties. (Translation) W. Baly, Walton and Maberly, London

Munhall K, Löfqvist A (1992) Gestural aggregation in speech: laryngeal gestures. J Phonet 20: 111–126

Munhall KG, Löfqvist A, Kelso JAS (1986) Laryngeal compensation following sudden oral perturbation. J Acoust Soc Am Suppl 1 80: S109

Nadler R, Abbs J (1988) Use of the x-ray microbeam system for the study of articulatory dynamics. J Acoust Soc Am Suppl 1 84: S124

Neff WD (1964) Temporal pattern discrimination in lower animals and its relation to language perception in man. In: de Reuck, O'Connor (eds) Ciba foundation symposium on disorders of language. Churchill, London, p 183

Neff WD, Diamond IT, Casseday JH (1975) Behavioral studies of auditory discrimination: central nervous system. In: Keidel Neff (eds) Handbook of sensory physiology, vol V/2. Springer, Berlin Heidelberg New York, p 307

Nittrouer S, Munhall K, Kelso JAS, Tuller B, Harris KS (1988) Patterns of interarticulator phasing and their relation to linguistic structure. J Acoust Soc Am 85: 1653–1661

Norris KS, Møhl B (1983) Can odontocetes debilitate prey with sound? Am Nat 122: 85–114

O'Shaughnessy D (1987) Speech communication: human and machine. Addison-Wesley, New York

O'Shaughnessy D (1995) Speech technology. In: Syrdal A, Bennett R, Greenspan S (eds) Applied speech technology. CRC Press, Boca Raton, pp 47–98

Ostry DJ, Munhall KG (1994) Control of jaw orientation and position in mastication and speech. J Neurophysiol 71: 1528–1545

Ostry DJ, Keller E, Parush A (1983) Similarities in the control of speech articulators and the limbs: kinematics of tongue dorsum movement in speech. J Exp Psychol Hum Percept Perform 9: 622–636

Owren MJ (1990) Acoustic classification of alarm calls by vervet monkeys (*Cercopithecus aethiops*) and humans: II. Synthetic calls. J Comp PsychOL 104: 29–40

Owren MJ, Bernacki RH (1988) The acoustic features of vervet monkey alarm calls. J Acoust Soc Am 83: 1927–1935

Owren MJ, Linker CD (1995) Some analysis techniques that may be useful to acoustic primatologists. In: Zimmermann E, Newman J, Jurgens (eds) Current topics in primate vocal communication, Plenum Press, New York, pp 1–27

Owren MJ, Linker CD, Rowe MP (1993) Acoustic features of tonal "grunt" calls in baboons. J Acoust Soc Am 94: 1823

Owren MJ, Seyfarth RM, Cheney DL (1995) Acoustic indices of production mechanisms underlying tonal "grunt" calls in baboons. J Acoust Soc Am 98: 2965

Papçun G, Hochberg J, Thomas TR, Laroche F, Zacks J, Levy S (1992) Inferring articulation and recognizing gestures from acoustics with a neural network trained on X-ray microbeam data. J Acoust Soc Am 92: 688–700

Parush A, Ostry D, Munhall K (1983) A kinematic study of lingual coarticulation in VCV sequences. J Acoust Soc Am 74: 1115–1125

Perkell JS (1969) Physiology of speech production: results and implications of a quantitative cineradiographic study. MIT Press, Cambridge

Perkell J, Cohen M, Garabieta I (1988) Techniques for transducing movements of points on articulatory structures. J Acoust Soc Am (Suppl) 1 84: S145

Perkell JS, Cohen MH, Svirsky MA, Matthies ML, Garabieta I, Jackson MTT (1992) Electromagnetic midsagittal articulometer systems for transducing speech articulatory movements. J Acoust Soc Am 92: 3078–3096

Perrier P, Laboissière R, Eck L (1991) Modelling of speech motor control and articulatory trajectories. Proc 12th Int Congr Phonet Sci 2 pp 62–65

Petersen MR (1981) The perception of species-specific vocalizations by animals: developmental perspectives and implications. In: Aslin R, Alberts J, Petersen M (eds) Development of perception: psychobiological perspectives, vol. 1 Auditory, chemosensory and somatosensory systems. Academic Press, New York , pp 67

Petersen M, Beecher M, Zoloth S, Marler P, Moody D, Stebbins W (1984) Neural lateralization of vocalizations by Japanese macaques: communicative significance is more important than acoustic structure. Behav Neurosci 98: 779–790

Rabiner LR, Schafer RW (1978) Digital processing of speech signals. Prentice Hall, Englewood Cliffs

Rahim MG, Goodyear CC (1990) Estimation of vocal tract filter parameters using a neural net. Speech Commun 9: 49–55

Rahim MG, Goodyear CC, Kleijn WB, Schroeter J, Sondhi MM (1993) On the use of neural networks in articulatory speech synthesis. J Acoust Soc Am 93: 1109–1121

Recasens D (1984) Timing constraints and coarticulation: alveolar-palatals and sequences of alveolar and /j/ in Catalan. Phonetica 41: 125–139

Remez RE, Rubin PE, Pisoni DB, Carrell TO (1981) Speech perception without traditional speech cues. Science 212: 947–950

Repp B (1983) Trading relations among acoustic cues in speech perception are largely a result of phonetic categorization. Speech Commun 2: 341–361

Repp B (1988) Integration and segregation in speech perception. Lang Speech 31: 239–271

Rosenberg A (1971) Effect of glottal pulse shape on the quality of natural vowels. J Acoust Soc Am 49: 583–590

Rothenberg M (1977) Measurement of airflow in speech. J Speech Hear Res 20: 155–176

Rothenberg M (1981) Some relations between glottal air flow and vocal fold contact area. In: Ludlow CL, Hart MOC (eds) Proc Con on the Assessment of vocal pathology, ASHA Rep 11, Am Speech-Language-Hearing Assoc, Rockville, Maryland, pp 88–96

Rubin JA (1983). Static and dynamic information in vowels produced by the hearing impaired. Doctoral Diss, City University of New York

Rubin PE (1995) HADES: a case study of the development of a signal analysis system. In: Syrdal A, Bennett R, Greenspan S (eds) Applied speech technology. CRC Press, Boca Raton, pp 501–520

Rubin PE, Baer T, Mermelstein P (1981) An articulatory synthesizer for perceptual research. J Acoust Soc Am 70: 321–328

Rubin PE, Tiede M, Vatikiotis-Bateson E, Goldstein L, Browman C, Levy S (1995) V-TV: the Haskins vocal tract visualizer CD-ROM. Paper presented at the ACCOR Workshop on Articulatory Databases, Munich, Germany, 25–26 May, 1995

Sackner MA (1980) Monitoring of ventilation without a physical connection to the airway. In: Sackner MA (ed) Diagnostic techniques in pulmonary disease, part I. Dekker, New York, pp 503–537

Saltzman E (1986) Task dynamic coordination of the speech articulators: a preliminary model. Exp Brain Res, Series 15: 129–144

Saltzman E, Kelso JAS (1987) Skilled actions: a task dynamic approach. Psychol Rev 94: 84–106

Saltzman E, Munhall KG (1989) A dynamical approach to gestural patterning in speech production. Ecol Psychol 1: 333–382

Saltzman E, Rubin P, Goldstein L, Browman CP (1987) Task-dynamic modeling of interarticulator coordination. J Acoust Soc Am 82: S15

Saltzman E, Goldstein L, Browman CP, Rubin P (1988a) Modeling speech production using dynamic gestural structures. J Acoust Soc Am 84: S146

Saltzman E, Goldstein L, Browman CP, Rubin P (1988b) Dynamics of gestural blending during speech production. Neural Networks 1: 316

Saltzman E, Lõfqvist A, Kinsella-Shaw J, Kay B, Rubin P (1995) On the dynamics of temporal patterning in speech. In: Bell-Berti F, Raphael LJ (eds) Producing speech: contemporary issues. For Katherine Safford Harris. AIP Press, New York, pp 469–487

Sasaki CT, Levine PA, Laitman JT, Crelin ES Jr (1977) Postnatal descent of the epiglottis in man. Arch Otolaryngolica 103: 169–171

Sawashima M (1977) Current instrumentation and technique for observing speech organs. Technocrat 9-4 : 19–26

Sawashima M, Abramson AS, Cooper FS, Lisker L (1970) Observing laryngeal adjustments during running speech by use of a fiberoptics system. Phonetica 22: 193–201

Schönle P, Grabe K, Wenig P, Hohne J, Schrader J, Conrad B (1987) Electromagnetic articulography: use of alternating magnetic fields for tracking movements of multiple points inside and outside the vocal tract. Brain Lang 31: 26–35

Schroeder MR (1967) Determination of the geometry of the human vocal tract by acoustic measurements. J Acoust Soc Am 41: 1002–1010

Schroeter J, Sondhi MM (1992) Speech coding based on physiological models of speech production. In: Furui S, Sondhi MM (eds) Advances in speech signal processing. Dekker, New York, p 231–268

Sears CH (1902) A contribution to the psychology of rhythm. Am J Psychol 13: 28–61

Sekiyama K, Tohkura, Y (1993) Inter-language differences in the influence of visual cues in speech perception. J Phonet 21: 427–444

Seyfarth RM, Cheney DL, Harcourt AH, Stewart K (1994) The acoustic features of double-grunts by mountain gorillas and their relation to behavior. Am J Primatol 33: 31–50

Shipley C, Carterette EC, Buchwald JS (1991) The effects of articulation on the acoustical structure of feline vocalizations. J Acoust Soc Am 89: 902–909

Shirai K (1993) Estimation and generation of articulatory motion using neuronal networks. Speech Commun 13: 45–51

Shirai K, Kobayashi T (1986) Estimating articulatory motion from speech wave. Speech Commun 5: 159–170

Shirai K, Kobayashi T (1991) Estimation of articulatory motion using neural networks. J Phonetics 19: 379–385

Simmons JA, Grinnell AD (1988) The performance of echolocation: acoustic images perceived by echolocating bats. In: Nachtigall PE, Moore PWB (eds) Animal sonar. Plenum, New York, p 353

Sondhi MM, Schroeter J (1987) A hybrid time-frequency domain articulatory speech synthesizer. IEEE Trans ASSP 35: 955–967

Sonoda Y, Wanishi S (1982) New optical method for recording lip and jaw movements. J Acoust Soc Am 72: 700–704

Stebbins WC, Sommers MS (1992) Evolution, perception and the comparative method. In: Webster DB, Fay RR, Popper AN (eds) The evolutionary biology of hearing. Springer, Berlin Heidelberg New York, p 211

Stetson RH (1905) A motor theory of rhythm and discrete succession. II. Psychol Rev 12: 293–350

Stetson RH (1928) Motor phonetics: a study of speech movements in action, 2nd edn. North Holland, Amsterdam (1951). (1st edn 1928 in Arch Neerl phonetique Exp 3)

Stone M (1990) A three-dimensional model of tongue movement based on ultrasound and x-ray microbeam data. J Acoust Soc Am 87: 2207–2217

Stone M (1991) Toward a model of three-dimensional tongue movement. J Phonetics 19: 309–320

Stone M, Vatikiotis-Bateson E (1995) Coarticulatory effects on tongue, jaw, and palate beavior. J Phonet 23: 81–100

Stone M, Shawker TH, Talbot TL, Rich AH (1988) Cross-sectional tongue shape during the production of vowels. J Acoust Soc Am 83: 1586–1596

Subtelny J, Oya N, Subtelny JD (1972) Cineradiographic study of sibilants. Folia Phoniatrica 24: 30–50

Summerfield Q (1987) Some preliminaries to a comprehensive account of audio-visual speech perception. In: Dodd B, Campbell R (eds) Hearing by eye: the psychology of lip-reading. Erlbaum, Hillsdale, New Jersey, p 3

Summerfield, Q (1991) Visual perception of phonetic gestures. In: Mattingly IG, Studdert-Kennedy M (eds) Modularity and the motor theory of speech perception. Erlbaum, Hillsdale, New Jersey, pp 117–137

Sussman HM, MacNeilage PF, Hanson RJ (1973) Labial and mandibular dynamics during the production of bilabial consonants: preliminary observations. J Speech Hear Res 16: 397–420

Tank DW, Hopfield JJ (1987) Neural computation by concentrating information in time. Proc Natl Acad Sci USA 84: 1896–1900

Tiede MK (1993) An MRI-based study of pharyngeal volume contrasts in Akan. Haskins Laboratories Status Report on Speech Research SR-113, New Haven, pp 107–130

Tuller B, Kelso JAS (1984) The timing of articulatory gestures: evidence for relational invariants. J Acoust Soc Am 76: 1030–1036

Tuller B, Kelso JAS (1995) Speech dynamics. In: Bell-Berti F, Raphael LJ (eds) Producing speech: contemporary issues. For Katherine Safford Harris. AIP Press, New York, pp 505–519

Tuller B, Kelso JAS, Harris KS (1982) Interarticulator phasing as an index of temporal regularity in speech. J Exp Psychol: Hum Percept Perform 8: 460–472

Tuller B, Kelso JAS, Harris KS (1983) Converging evidence for the role of relative timing in speech. J Exp Psychol: Hum Percept Perform 9: 829–833

Tuller B, Shao S, Kelso JAS (1990) An evaluation of an alternating magnetic field device for monitoring tongue movements. J Acoust Soc Am 88: 674–679

Turvey MT (1977) Preliminaries to a theory of action with reference to vision. In: Shaw R, Bransford J (eds) Perceiving, acting, and knowing: toward an ecological psychology. Lawrence, Hillsdale, New Jersey, pp 211–265

Vatikiotis-Bateson E (1988) Linguistic structure and articulatory dynamics. Indiana University Linguistics Club, Bloomington

Vatikiotis-Bateson E, Kelso JAS (1984) Remote and autogenic articulatory adaptation to jaw perturbation during speech: more on functional synergies. J Acoust Soc Am 75: S23–24

Vatikiotis-Bateson E, Kelso JAS (1993) Rhythm type and articulatory dynamics in English, French, and Japanese. J Phonet 21: 231–265

Vatikiotis-Bateson E, Ostry J (1995) An analysis of the dimensionality of jaw motion in speech. J Phonet 23: 101–117

Vatikiotis-Bateson E, Stone M (1989) In search of lingual stability. J Acoust Soc Am 86: S115

Vatikiotis-Bateson E, Hirayama M, Kawato M (1991) Neural network modelling of speech motor control using physiological data. Phonetic Exp Res. Inst Linguistics, U Stockholm (PERILUS) 14: 63–68

Wada Y, Kawato M (1995) A theory for cursive handwriting based on the minimization principle. Biol Cybern 73: 3–13

Wada Y, Koike Y, Vatikiotis-Bateson E, Kawato M (1995) A computational theory for movement pattern recognition based on optimal movement pattern generation. Biol Cybern 73: 15–25

Wakita H (1979) Estimation of vocal-tract shapes from acoustical analysis of the speech wave: the state of the art. IEEE Trans on Acoustics, Speech, and Signal Proc, ASSP-27: 281–285

Walsh EG (1957) An investigation of sound localization in patients with neurological abnormalities. Brain 80: 222–250

Warren DW (1986) Compensatory speech behaviors in individuals with cleft palate: a regulation/control phenomenon? Cleft Palate J 23: 251–260

Weismer G (1983) Acoustic descriptions of dysarthric speech: perceptual correlates and physiological inferences. Speech Motor Control Laboratory, Preprints. University of Wisconsin, Madison

Westbury JR (1994) On coordinate systems and the representation of articulatory movements. J Acoust Soc Am 95: 2271–73

Whalen DH, Liberman AM (1987) Speech perception takes precedence over nonspeech perception. Science 237: 169–171

Wilhelms-Tricarico R (1995) Physiological modeling of speech production: methods for modeling soft-tissue articulators. J Acoust Soc Am 97: 3085–3098

Witten IH (1982) Principles of computer speech. Academic Press, Orlando

Wood S (1979) A radiographic analysis of constriction location for vowels. J Phonet 7: 25–43