A History of Speech Production Research

Maureen Stone

Postal:
Department of Neural and Pain Sciences and Department of Orthodontics
University of Maryland School of Dentistry
650 West Baltimore Street
Baltimore, Maryland 21201
USA

Email:
stone@umaryland.edu

Christine H. Shadle
Postal
Haskins Laboratories
300 George Street
New Haven, Connecticut 06511
USA

Email:
shadle@haskins.yale.edu

Many of the initial assumptions about speech and how we produce it turned out to be wrong; these myths reveal the complexity of the process.

Introduction

Researchers have studied speech production, everything that goes into talking from the thought to the articulator movements to the sound emerging from the mouth, to understand it, transmit it over long distances, mimic it, and treat speech pathologies. Progress has not been steady. Sometimes researchers had to trip over their own wrong assumptions before they could take the next step.

We think of "speech" as an acoustic signal that is heard when people talk to each other. But that sound is the end result of a complex process, starting in the brain where thoughts are translated into language and muscle commands are sent to our vocal tract to shape it and get air moving through it. As the air moves through the slowly changing vocal tract, the right sounds will be produced in the right sequence so that they can be interpreted by the listener as speech. Once the muscles are activated, production is in progress and the physical properties of aerodynamics take over to produce a particular acoustic waveform. Given the variety of vocal tract shapes possible, predicting the sounds that will be produced is itself complex.

Speech production includes all the processes involved in producing the acoustic waveform described as speech. Speech perception includes all the processes involved in receiving, processing, and interpreting that waveform as speech. In this article, we describe some of the pivotal points that have changed our understanding of speech production over the years.

Milestones in science often come from radical changes in our understanding of how things work. In speech production, a number of such changes were crucial to the development of the field. Therefore, we have organized this article as a series of "myths," that were well believed until advances in the sophistication of our instrumentation, the richness of our data, and our knowledge of other fields inspired advancements in our thinking.

Myth 1: Speech Is Made of Concatenated Sounds with Silence Between Words

In 1877, Henry Sweet, a source for *Pygmalion's* Henry Higgins, developed the science of phonetics and used it to describe Received Pronunciation, a dialect of British English traditionally associated with the nobility. At the same time, Alexander Melville Bell (father of Alexander Graham Bell) developed visible speech (Bell, 1867), which is a set of graphic diagrams of articulatory positions used to teach the deaf to learn speech. These were early efforts on the part of linguists and speech pathologists to use an orthographic alphabet to capture features of concatenated oral sounds.

In the 1930s, interest in speech disorders grew. In 1937, Charles Van Riper, a stutterer himself, developed a scientific basis for the research and remediation of stuttering. The 2010 film *The King's Speech* presents the application of such tech-

niques, at around the same time, to the stuttering malady of King George VI of England.

Semi-intelligible speech was synthesized for the first time at the 1939 New York World's Fair using the Voder, a customized filter bank controlled by highly trained human operators. Up to this time, it had been assumed that individual sounds were produced individually and strung together like beads on a chain. Then, during World War II, the speech spectrograph was developed for use in clandestine analyses of speech and speakers (cf. Solzhenitsyn's 1968 classic novel *The First Circle* for its development and use in speaker recognition in the USSR).

Observations of the acoustic spectrum for continuous speech, seen in Potter et al.'s classic 1947 book *Visible Speech*, turned this notion on its head. The spectrograms showed three remarkable features of speech, which are demonstrated in **Figure 1, bottom**. (1) There are no pauses between words (e.g., "we're shop" and "for sun"). (2) Instead, apparent pauses are due to moments of silence inherent in stop consonants, such as the "p" in "shopping." (3) Speech sounds, or phonemes, are not independent but overlap in time with the production of their neighbors, producing coarticulation and facilitating the production of rapid speech. A good example is the first word, "we're," in which two energy bands converge (**Figure 1, arrows**). Without pauses and with overlapping sounds, it may seem a wonder that the listener can hear word boundaries. In fact, word and phrase breaks are signaled by modification of the speech sounds themselves, such as the phrase final lengthening of what would otherwise be a short unstressed vowel in "dresses."

Myth 2: When Synthesizing Speech, Female Speakers Can Be Treated as Small Male Speakers

In 1952, Peterson and Barney published a study of the vowels of 76 speakers, showing that 10 American English vowels formed separate clusters on a graph of the second versus the first formant frequencies (F2 vs. F1). Researchers then used formant synthesizers to test how closely they could mimic their own speech while tweaking the structure and parameter set. As a result, early examples of synthetic speech sounded very much like their creators (for example, http://tcscasa.org/klatts-history-of-speech-synthesis/ examples 4 and 6: Gunnar Fant; examples 7 and 8: John Holmes; example 9: Dennis Klatt, as described in Klatt, 1987). To match women's and children's speech, the parameters were simply scaled. However, those synthesized voices were not nearly as natural sounding as the male voices (e.g., example 9, "DECTalk

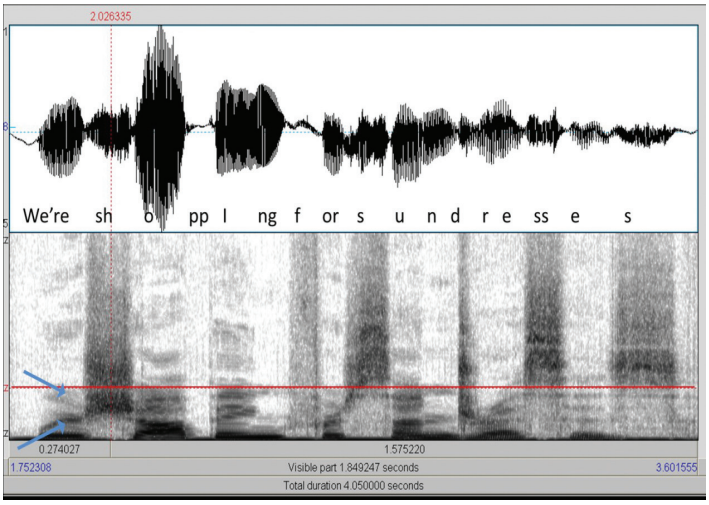

Figure 1. *Speech spectrogram of the speech signal (top) and the wideband spectrogram (bottom) for the sentence "We're shopping for sundresses" spoken by an adult female. The spectrogram shows the energy from 0 to 20 kHz. Red line: more typically used frequency range, 0-5 kHz; blue arrows: regions where the formant transitions indicate coarticulation. See the text in Myths 1 and 4.*

scaled,"(http://tcscasa.org/klatts-history-of-speech-synthesis/). What had gone wrong?

The assumption, based on Peterson and Barney's vowel clusters, had been that since men had the lowest F1-F2 frequencies, women next lowest, and children the highest, scalar changes were sufficient to convert synthesizers to sound like women's and children's voices. As source-filter models of speech production were developed in the 1950s and 1960s (Fant, 1970; Flanagan, 1972), this experimental generalization made theoretical sense. After all, the formants were the resonances of the vocal tract that filtered the source produced by vibration of the vocal folds. Women's and children's vocal tracts were shorter than men's, so their formants were higher. Their larynges and thus vocal folds were smaller, and so their ranges of phonation frequencies were higher as well.

The missing piece was that not only the phonation frequency but also the details of the voice source differed, both the spectral properties and the cycle-to-cycle variations. Methods were developed to inverse filter speech and derive the glottal waveform, which when used as the source made synthesized speech sound more natural (Holmes, 1973). Defining source parameters typical of men's or of women's speech was also helpful (Fant et al., 1985; other studies are summarized in Klatt, 1987). Women were found to be more breathy and their glottal waveforms had a longer open quotient. However, as Klatt wrote, "rules for dynamic control of these [voice-source] variables are quite primitive. The limited naturalness of synthetic speech from this and all other similar devices suggests that either something is still missing

from the voicing source models or that we do not yet know how to control them properly" (Klatt, 1987, pp. 745-746). Since 1987, several studies have resulted in improvements to the synthesis of female speech. Klatt and Klatt (1990) improved naturalness using voice quality variation. Karlsson (1992) enhanced naturalness beyond the parameters controlled in traditional synthesis of male speech by adding extra formants, a noise-modulated fundamental frequency, a vocal fry component, and an improved voice source model. Subsequent work by Karlsson and Neovius (1994) addressing both source and filter problems has led to even more natural-sounding synthesis of female voices (see http://tcscasa.org/klatts-history-of-speech-synthesis/).

Myth 3: Two-Dimensional Images and Models Sufficiently Represent Three-Dimensional Structures Like the Tongue

The first techniques used to visualize the vocal tract in motion produced 2-dimensional (2-D) data. They projected 3-dimensional (3-D) motion onto a single plane using X-rays or midsagittal point tracking (e.g., X-ray Microbeam; Kiritani et al., 1975). As a result, measurements of vocal tract motion did not capture three dimensions. Although people found ways to get 3-D information, the methods were laborious (e.g., static palatography [Ladefoged, 1957]) or inexact (e.g., cadavers [Fant, 1970]) or gave partial information (see **Figure 2**).

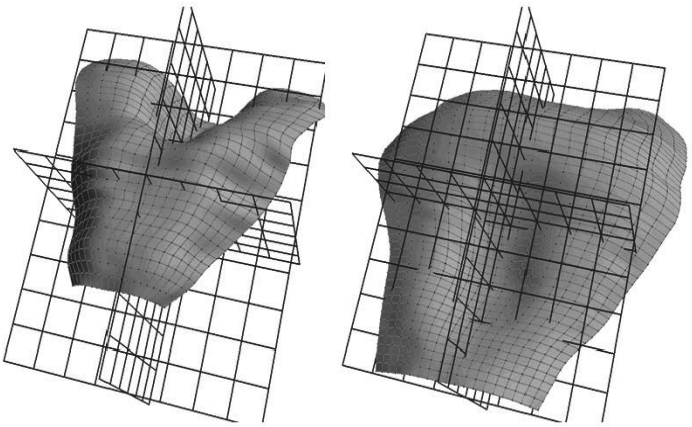

Figure 2. *The 3-D tongue shapes for /i/ (left) and /l/ (right) show how deceptive a lateral X-ray of the highest edge would be.*

Despite the 3-D efforts, many articulatory models were based only on the midsagittal plane, and this shaped people's thinking accordingly. Thus, physiological models mostly used cylindrical cross-sectional areas formulated mathematically from 2-D cross-sectional distances (cf. Fant, 1970). Similarly, theories of speech production were based on 2-D representations of articulatory motion (Hardcastle, 1976; Browman and Goldstein, 1989; Saltzmann and Munhall, 1989).

These conceptual blinders were removed in the late 1980s when ultrasound, and later MRI, captured tissue slices of the tongue and vocal tract. Suddenly, tongue motions that appeared to occur exclusively in the anterior-posterior and superior-inferior directions on 2-D projection X-rays were found to contain dramatic nonuniform shape changes in the cross-sectional dimension as well (Stone et al., 1988; Baer et al., 1991; Badin et al., 2002).

At the same time, a new theory emerged that redefined the field's understanding of muscular structures such as tongues and tentacles. These structures have 3-D orthogonal muscle architecture and volume preservation, which makes them highly deformable in 3-D space (Kier and Smith, 1985). In addition, neuromuscular and anatomical explorations (Slaughter et al., 2005; Stone et al., 2016) have shown that innervation and fiber architecture of tongue muscles support complex patterns of muscle coactivation to stabilize, stiffen, and deform local tongue regions during speech motions.

These developments, combined with increased computer power, drastically changed our understanding of the complexity of the vocal tract tube and the structures that shape it. Three-dimensional finite-element models now predict tongue and airway deformation, capturing their complexity and allowing reevaluation of previous ideas about speech motor control (cf. Stavness et al., 2012, Bijar et al., 2015).

Myth 4: All the Information in Speech Is Contained Within a Set Bandwidth. Therefore, We Don't Need to Consider Frequencies Above 5 kHz in Describing Speech Sounds.

Although young humans can hear from 20 Hz to 20 kHz, many speech sounds extend up to only 7 kHz or so, and many studies analyze speech in a bandwidth up to 4 or 5 kHz. There are interlocking reasons for this, but one consequence is that it is easy to forget that speech sounds do extend to higher frequencies.

In the early days of Bell Telephone Laboratories, research was done to find the optimal bandwidth for the telephone (Flanagan, 2009). The bandwidth of 300 Hz to 3,500 Hz was agreed on, which corresponds to the most sensitive range of

human hearing. Some of the research conducted revealed a high degree of redundancy in speech. For instance, filtering out all sound above 1,800 Hz reduced intelligibility to 67%, but filtering out all sound below 1,800 Hz also reduced intelligibility to 67% (Moore, 1997). Miller and Nicely's (1955) study of consonant confusions that result from limiting the bandwidth and adding noise to the transmitted signal described in detail which consonants are misperceived in the various conditions.

A further contributing factor is the way in which sound propagation in the vocal tract is modeled. If only acoustic propagation of plane waves is considered, the wave equation simplifies to be analogous to the equations governing voltage and current in electrical circuits, as shown by Fant (1970, pp. 27-36). This allows sound propagation in the vocal tract to be modeled as a transmission line and for circuit theory and linear system theory to be used. One consequence is that nonplane wave modes cannot be predicted by such models. These cross modes begin to propagate (rather than dying out) above about 4-5 kHz for typical vocal tract dimensions. Above this limit, the plane wave modes still exist and are predicted correctly, but because the cross-modes are not predicted, the estimated sound spectrum is progressively less accurate.

As a result of these two bandwidth restrictions, one using speech perception to address a practical limitation on telephone bandwidths and the other using a simplifying assumption to allow circuit analogs for sound propagation in the vocal tract, it is easy to forget that speech sounds are produced and can be heard above 5 kHz. In particular, noise sources, which are the essence of many consonants, use higher frequencies. The difficulty of distinguishing "esss" from "efff" over a telephone is an obvious example, where the noise-excited broad peak at about 6 kHz that occurs in /s/ and not /f/ is not present in the transmitted signal. As can be seen in **Figure 1**, significant noise can extend all the way up to 20 kHz for fricatives such as "f," "s," and "sh," and stop releases that occur at the end of "p, t, and k." For this (female) speaker, vowel formants appear distinct well above 5 kHz in some syllables, such as in "shopping," "sun," and "…sses." Characteristics that help us to identify particular speakers and the emotional state of a speaker also appear to extend up to 8 kHz (O'Shaughnessy, 2000, p. 452).

Finally, when studying speech production, the entire sound spectrum provides clues to the speech production mechanism. For instance, a small spectral peak or trough attains greater significance when its integral multiple is detected at a higher frequency. Such acoustic evidence should be noted before filtering and down-sampling to the frequency range of greatest interest for a particular study.

Myth 5: Some Aspect of Produced Speech Must Be Invariant, Such as Acoustics or Articulation

One of the earliest assumptions in speech research was that the specific spectral features associated with a specific speech sound were immutable. Therefore, multiple repetitions of a sound would contain identical representations of these features, which, when extracted by the brain of a listener, would result in perception of the spoken speech sound. The sound spectrogram, which debunked Myth 1, also revealed that the spectra of speech sounds were not invariant but differed with every repetition; they reflected acoustic features of neighboring sounds. In addition, multiple repetitions of the exact same speech task could vary in their spectral and temporal features. Thus, repetitions of perceptually identical sounds were not acoustically invariant.

The search for invariance then moved to the physical articulation of sounds, with the idea that the brain refers the acoustic signal back to its knowledge of the vocal tract (Liberman et al., 1967). Alas, articulator motions were also variable due to biomechanical constraints, preference for ease of production, and linguistic rules that enhance acoustical salience and distinctiveness of speech sounds. Other candidates for invariance included the constriction size, the vocal tract area function, and the electromyography (EMG) signals of the muscles (Perkell and Klatt, 1986). None of these components was invariant. However, we learned a lot about control patterns and timing. In the end, variability at all levels of production has been accepted for the most part, and theories of speech perception now seek to explain why invariance is not a problem for the brain in the human perception of speech despite its being an enormous problem in machine recognition of speech (cf. Guenther et al., 2006).

Myth 6: There Are One-to-One Mappings in Speech Production

There are two versions of this myth. One-to-one mappings have been sought between acoustic spectra and vocal tract shape and also between tongue surface shapes and tongue muscle activity. The first variant of this myth arose because in predictive models, a single vocal tract shape is linked to a single acoustic spectrum. The inverse assumption was embraced as well because one-to-one inverse mappings seemed

reasonable and using them was very convenient. Inverse mappings allowed us to take audio recordings, which are easy to make, and estimate vocal tract features, such as the source of voice qualities like hoarseness, tongue position in different accents and languages, or the subject identifiers used in forensics. Unfortunately, inverse maps are not one-to-one. A single acoustic spectrum can be produced by more than one vocal tract shape. This is easily exemplified by ventriloquists who routinely produce acoustic features associated with lip motion using other parts of the vocal tract.

Strong evidence for a many-to-one relationship between vocal tract shapes and an acoustic spectrum came from two sources. The first source was a series of studies that measured vowels while the jaw was held rigid by a bite block inserted between the molars like the stem of a smoker's pipe. Speech acoustics and perception were the same with and without the bite block, indicating that the subjects had found a different articulatory position to produce the same vowel (Gay et al., 1981) and alveolar consonant (Flege et al., 1988) sounds. The second source was acoustic-to-articulatory inversion studies, which used vocal tract models to show that many different vocal tract shapes could produce a specific sound (Atal et al., 1978).

The second version of this myth arose from the expectation that there would be a one-to-one relationship between muscle activation and specific tongue surface shapes. In the 1970s, extensive EMG studies were conducted to establish these relationships. However, EMG studies showed that muscle activity is quite variable for the same speech task and must be averaged across repetitions to reveal activation patterns (see **Figure 3**). Moreover, tongue muscle activation is not simple. Instead, local motor units can coactivate within and across muscles to create local internal motions and supporting regions of stiffness (Cope and Sokoloff, 1999). Other research considers how muscle activation links to internal tongue motion patterns and finally to surface tongue deformations (Stone et al., 2008).

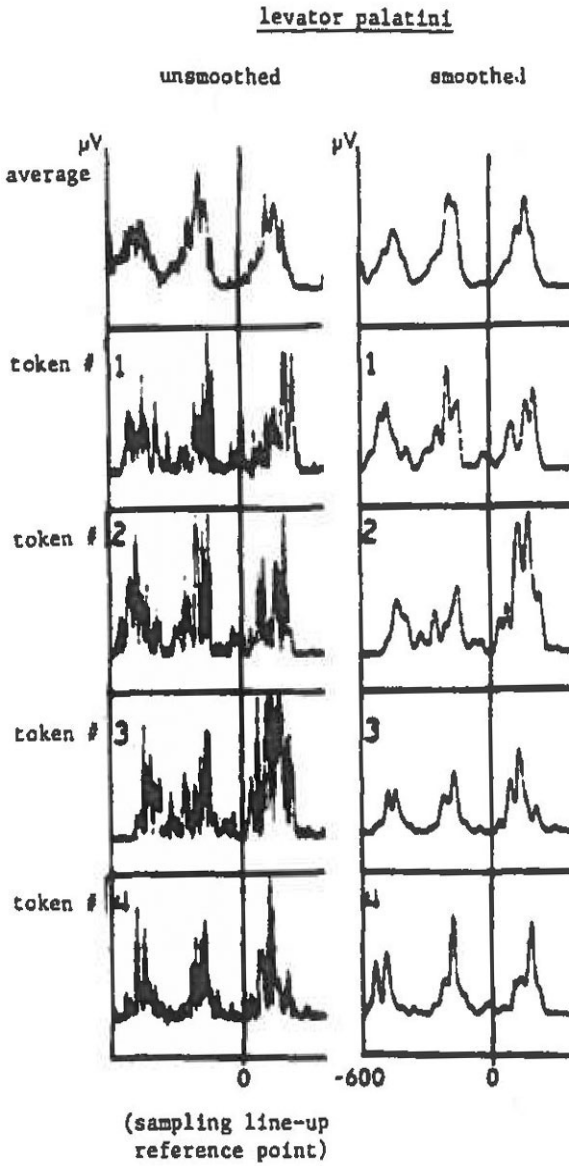

Figure 3. *Individual (rows 1-4) and 20 token average (top row) of EMG signals for the spoken utterance "fax map." From Harris (1982), with permission from ASHA.*

Myth 7: Aerodynamics and Acoustics Can Be Neatly Separated in Vocal Tract Models and Speech Synthesizers Without a Loss of Predictive or Conceptual Power

There are three basic types of speech synthesis: articulatory, formant, and concatenative. *Articulatory synthesizers* model the positions and movement of articulators. *Formant synthesizers* model the sequence of resonances and antiresonances. *Concatenative synthesizers* string together prerecorded and coded speech segments. All three types must include the effects of coarticulation, although these are included in different ways. The particular type that offers the best quality for commercial synthesis has varied over the decades. In the 1970s and 1980s, formant synthesis was the best commercial method of synthesizing speech, whereas currently it is concatenative synthesis. However, only articulatory synthesis allows for the synthesis of any sound as produced by any vocal tract and thus has more potential for synthesis in clinical applications.

In the earliest articulatory synthesizers, sound sources, for both phonation and supraglottal noise sources, were para-

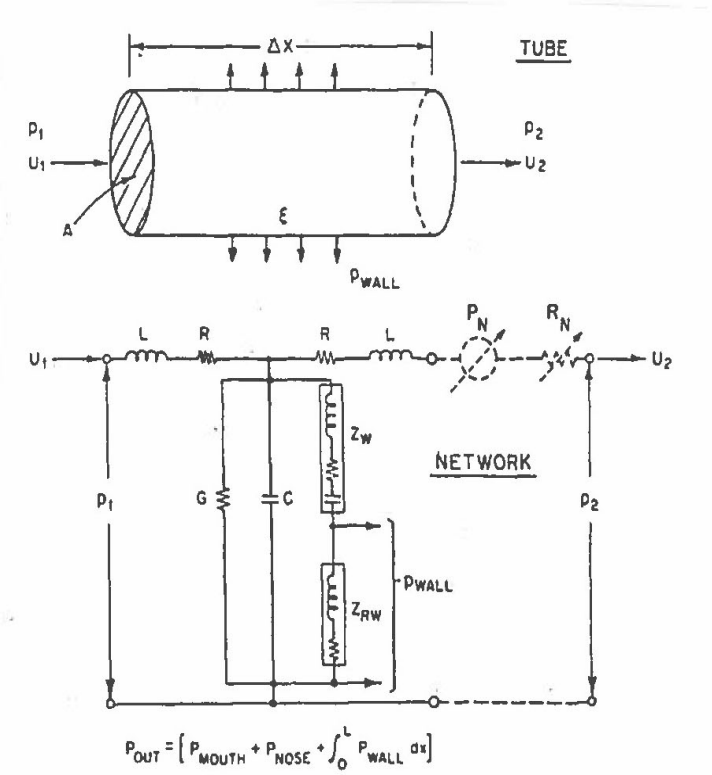

Fig. 1. Representation of one-dimensional acoustic wave propagation in a right-circular tube with yielding side wall.

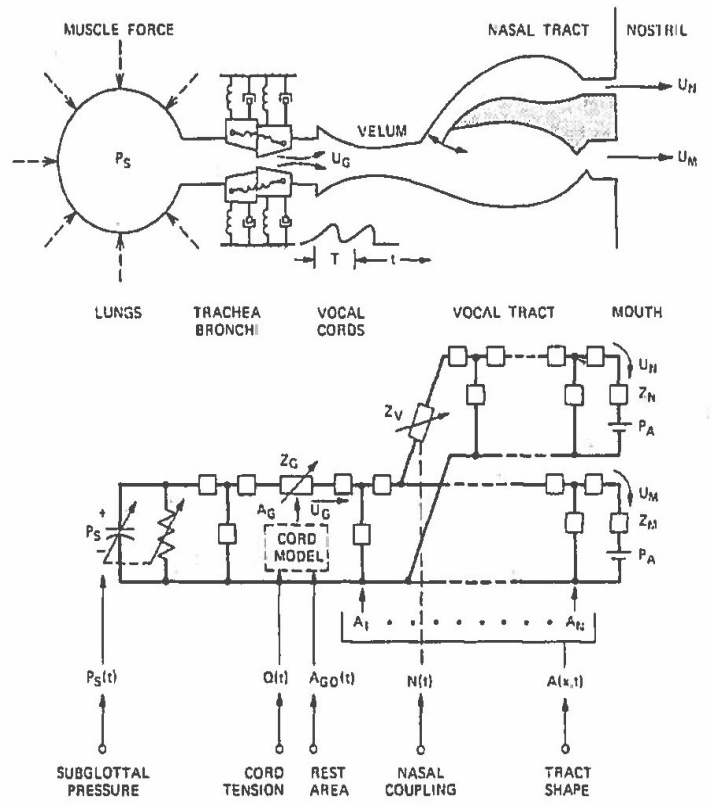

Fig. 2. Network representation of the vocal cord–vocal tract synthesizer.

Figure 4. *Two diagrams of the articulatory synthesizer incorporating aerodynamics such that sound sources were automatically activated. Top half (original* ***Figure 1****): Circuit analog for one section showing the turbulence noise source that allowed automatic synthesis of consonants (PN). Other components are circuit elements of inductance (L), resistance (R) and conductance (G), and capacitance (C) that model the mass of the air or tract wall, losses due to friction, and compliance of the air or walls, respectively. Pressure (p) and volume velocity (U) are equivalent to voltage and current, respectively. Bottom half (original* ***Figure 2****): Top diagram shows the system from lungs to oral and nasal cavities. Bottom diagram shows the transmission line representation with subglottal and atmospheric pressure sources (P_S and P_A) and the vocal cord model in addition to the three-element transmission line sections. Reprinted from Flanagan and Ishizaka (1976, Figures 1 and 2), with permission.*

metric, that is, they were placed at the appropriate location in the circuit analog of the vocal tract and their output was activated when needed. The circuit analog thus modeled only the filter properties of acoustic resonance and sound propagation, whereas the sources modeled the aeroacoustic phenomena that led to the generation of sound.

Flanagan and colleagues sought ways to incorporate aerodynamics more directly into the source and the filter. In their synthesizers, current was divided into low- and high-frequency components with the DC component modeling convection velocity and the AC component modeling the acoustic volume velocity. This allowed the DC flow to be used to calculate the Reynolds number along the tract and generate turbulent noise sources if it rose above a critical value (Flanagan and Cherry, 1969; Flanagan and Ishizaka, 1976). The controllable noise source included in every section of the vocal tract model is shown in **Figure 4, top half**. The use of low- and high-frequency components also allowed respiration to be modeled by including an external DC voltage source to model atmospheric pressure and a variable capacitor to model lung pressure (see **Figure 4, bottom half**).

This division of the current into the "acoustic volume velocity" and the "DC flow," although allowing the powerful concepts of circuit theory to be applied to speech production, is, unfortunately, a place where the circuit model misleads. Air travels from the lungs through the vocal tract at a mean

flow rate that is much slower than the speed of sound waves traveling through it. These two (or more) velocities cannot be captured by an electrical analogue. As a result, the details of various phenomena are incorrectly modeled, including the meaning of the DC flow computed by inverse filtering (see Shadle et al., 1999) and the mechanism by which the noise source is modulated by the voice source in voiced fricatives (Jackson and Shadle, 2000). More importantly, the geometric details incorporated in such a synthesizer make one think it is more physically accurate than it is.

Computational fluid dynamics (CFD) allows the fluid properties to be modeled. Although CFD has been a very useful third domain in which to study phonation (Zhang et al., 2002; Zhao et al., 2002) and, to some extent, turbulence noise in fricatives (Adachi and Honda, 2003), it is not at all straightforward to extract the sound that results from a given flow field. The addition of turbulence imposes a big computational burden. Although this method combines aerodynamics and acoustics successfully, it relies on computer power to handle the fine details of fluid motion rather than simplifying them so that only the aspects of the flow known to be crucial to the sound produced are modeled. It thus offers a more physically realistic method of predicting the flow field and the sound produced but not yet a different way to think about that process.

Some smaller studies offer a way forward. For instance, by relaxing some of the assumptions in circuit analogues, it was possible to test how much of a difference each makes (Davies et al., 1993). Particular articulatory-aerodynamic interactions have been studied, such as whether intraoral pressure has an appreciable effect on tongue motion (Mooshammer et al., 1995) or whether articulatory changes such as cavity expansion have an appreciable effect on phonation (Westbury, 1983). The interaction of a sound wave with a cloud of turbulence has been studied to explain the production of [s] (Howe and McGowan, 2005). Measurements in a mechanical model of the vocal folds and tract were interpreted in terms of a combined flow and acoustic field (Barney et al., 1999). It has long been accepted that vocal fold vibration is a complex system. To predict the effect of any physical change accurately, a vocal fold model must include aerodynamic, mechanical, and acoustic elements. We have not yet arrived at comparable models of supraglottal aeroacoustic sources.

Final Thoughts

The study of speech production is an ongoing endeavor. The well-accepted wisdom of any era is subject to revision in the future with the advent of new ideas, new instrumentation, and new research. This cycle of repudiation and revision will surely happen again with the ideas we hold dear today. These seven myths were once among the most prominent theories of their day, and some aspects of them are still open to debate. We have presented them to describe the history and development of the field.

Sometimes the major theories of the past were converted to "myth" status by the development of instruments and methodologies that provided additional perspective. Such changes in perspective occurred when the sound spectrograph revealed the true nature of speech sound sequencing and when sophisticated imaging techniques revealed what the 3-D vocal tract looked like in motion. In other cases, existing instruments or slight modifications added the information needed to rethink our idea, such as when studies included frequencies higher than telephone bandwidth and when bite block studies revealed the variety of ways one can produce the same speech sound. Most importantly, openness of thought and discussion of ideas allowed us to change our theories and models, even when the models were more elegant and more appealing than the true data.

Acknowledgments

We would like to thank Ereni Sevasti, the subject for Figure 1, and Richard Lissemore, who recorded her speech for us.

Biosketches

Maureen Stone is a professor at the University of Maryland School of Dentistry, Baltimore, with a joint appointment in the Department of Neural and Pain Sciences and the Department of Orthodontics. She is also director of the Vocal Tract Visualization Laboratory. She is a speech scientist by profession and has spent her career using imaging techniques such as MRI and ultrasound to study the behavior of the human tongue during speech. She began her career as a research scientist at the National Institutes of Health and then at Johns Hopkins University before coming to the University of Maryland. She is a Fellow of the Acoustical Society of America.

Christine Shadle, an electrical engineer with a strong interest in music, began doing speech research at Bell Telephone Laboratories. She studied the aeroacoustics of fricative consonants at MIT and continued that work as a Hunt and NATO postdoctoral fellow at the Institute of Sound and Vibration Research (ISVR) in Southampton, UK, and at KTH in Stockholm, Sweden. At the University of Southampton and, since 2004, as a senior research scientist at Haskins Laboratories, she uses mechanical models, vocal tract imaging, and signal processing in her research. A Fellow of the Acoustical Society of America, she has served as an associate editor of *The Journal of the Acoustical Society of America* and on the Executive Council.

References

Adachi, S., and Honda, K. (2003). CFD approach to fricative sound sources. In Palethorpe, S., and Tabain, M. (Eds.). *Proceedings of the 6th International Seminar on Speech Production*, Sydney, Australia, December 7-10, 2003, pp. 1-6.

Atal, B. S., Chang, J. J., Mathews, M. V., and Tukey, J. W. (1978). Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer-sorting technique. *The Journal of the Acoustical Society of America* 63, 1535-1555.

Badin, P., Bailly, G., Revéret, L., Baciu, M., Segebarth, C., and Savariaux, C. (2002). Three-dimensional linear articulatory modeling of tongue, lips and face, based on MRI and video images. *Journal of Phonetics* 30, 533–553.

Baer, T., Gore, J. C., Gracco, L. C., and Nye, P. W. (1991). Analysis of vocal tract shape and dimensions using magnetic resonance imaging: Vowels. *The Journal of the Acoustical Society of America* 90, 799-828.

Barney, A., Shadle, C. H., and Davies, P. O. A. L. (1999). Fluid flow in a dynamic mechanical model of the vocal folds and tract, I: Measurements and theory. *The Journal of the Acoustical Society of America* 105, 444-455.

Bell, A. M. (1867). *Visible Speech: The Science of Universal Alphabetics.* Simkin, Marshall & Co., London.

Bijar, A., Rohan, P. Y., Perrier, P., and Payan, Y. (2015). Atlas-based automatic generation of subject-specific finite element tongue meshes. *Annals of Biomedical Engineering* 44, 16-34.

Browman, C. P., and Goldstein, L. (1989). Articulatory gestures as phonological units. *Phonology* 6, 201-251.

Cope, T. C., and Sokoloff, A. J. (1999). Orderly recruitment among motoneurons supplying different muscles. *Journal of Physiology* (Paris) 93, 81-85.

Davies, P. O. A. L., McGowan, R. S., and Shadle, C. H. (1993). Practical flow duct acoustics applied to the vocal tract. In Titze, I. R. (Ed.), *Vocal Fold Physiology: Frontiers in Basic Science.* San Diego: Singular Publishing Group, Inc., pp. 93-142.

Fant, G. (1970). *Acoustic Theory of Speech Production: With Calculations Based on X-ray Studies of Russian Articulations.* Mouton and Company, The Hague.

Fant, G., Lin, Q. C., and Gobl, C. (1985). Notes on glottal flow interaction. *Speech Transmission Laboratory-Quarterly Progress and Status Report* 2-3, 21-45.

Flanagan, J. L. (1972). *Speech Analysis Synthesis and Perception.* Springer-Verlag, New York.

Flanagan, J. L. (2009). Curious science in Ma Bell's house of magic during the golden years. *IEEE Signal Processing Magazine* 26, 10-36.

Flanagan, J. L., and Cherry, L. (1969). Excitation of vocal-tract synthesizers. *The Journal of the Acoustical Society of America* 45, 764-769.

Flanagan, J. L., and Ishizaka, K. (1976). Automatic generation of voiceless excitation in a vocal cord-vocal tract speech synthesizer. *IEEE Transactions on Acoustics Speech and Signal Processing* 24, 163-170.

Flege, J. E., Fletcher, S. G., and Homiedan, A. (1988). Compensating for a bite block in /s/ and /t/ production: Palatographic, acoustic, and perceptual data. *The Journal of the Acoustical Society of America* 83, 212-228.

Gay, T., Lindblom, B., and Lubker, J. (1981). Production of bite-block vowels: Acoustic equivalence by selective compensation. *The Journal of the Acoustical Society of America* 69, 802-810.

Guenther, F. H., Ghosh, S. S., and Tourville, J. A. (2006). Neural modeling and imaging of the cortical interactions underlying syllable production. *Brain and Language* 96, 280–301.

Hardcastle, W. J. (1976). *Physiology of Speech Production.* Academic Press, London.

Harris, K. (1982). Electromyography as a technique for laryngeal investigation. *ASHA Reports* 11, 70-86.

Holmes, J. N. (1973). The influence of glottal waveform on the naturalness of speech from a parallel formant synthesizer. *IEEE Transactions on Audio and Electroacoustics* 21, 298-305.

Howe, M. S., and McGowan, R. S. (2005). Aeroacoustics of [s]. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* 461, 1005-1028.

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

Karlsson, I. (1992). *Analysis and Synthesis of Different Voices with Emphasis on Female Speech.* PhD Thesis, Department of Speech Communication and Music Acoustics, Royal Institute of Technology, KTH, Stockholm.

Karlsson, I., and Neovius, L. (1994). VCV-sequences in a preliminary text-to-speech system for female speech. *Speech Transmission Laboratory-Quarterly Progress and Status Report* 35, 47-58.

Kier, W. M., and Smith, K. K. (1985). Tongues, tentacles and trunks: The biomechanics of movement in muscular-hydrostats. *Zoological Journal of the Linnean Society* 83, 307–324.

Kiritani, S., Itoh, K., and Fujimura, O. (1975). Tongue-pellet tracking by a computer-controlled x-ray microbeam system. *The Journal of the Acoustical Society of America* 57, 1516-1520.

Klatt, D. H. (1987). Review of text-to-speech conversion for English. *The Journal of the Acoustical Society of America* 82, 737-793.

Klatt, D. H, and Klatt, L. C. (1990). Analysis, synthesis, and perception of voice quality variations among female and male talkers. *The Journal of the Acoustical Society of America* 87, 820-857.

Ladefoged, P. (1957). Use of palatography. *Journal of Speech and Hearing Disorders* 22, 764-774.

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., and Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review* 74, 431–461.

Miller, G. A., and Nicely, P. E. (1955). An analysis of perceptual confusions among some English consonants. *The Journal of the Acoustical Society of America* 27, 338-352.

Moore, B. J. (1997). *An Introduction to the Psychology of Hearing*, 3rd ed. Academic Press, Ltd., London.

Mooshammer, C., Hoole, P., and Kühnert, B. (1995). On loops. *Journal of Phonetics* 23, 3-21.

Continued on Page 61