Phonetics Research at Bell Laboratories *

Chilin Shih and Joseph P. Olive
Bell Laboratories, Lucent Technologies
600 Mountain Avenue, Murray Hill, NJ, USA, 07974
{cls,jpo}@research.bell-labs.com

The history of phonetic and acoustic research at Bell Laboratories reflects a strong tradition of interdisciplinary inquiries. Phonetics research is inseparable from many related fields such as hearing, acoustics, speech coding, speech transmission, and speech synthesis. Generations of physicists, engineers, linguists, psychologists, mathematicians, and computer scientists joined forces to solve fundamental problems in the study of human speech. In the process, they often needed to design tools and to develop methodologies, which had profound impacts in the field, and will continue to influence speech research in the future.

Alexander Graham Bell, the founder of the Bell Telephone Company, grew up with deep awareness of the needs of the deaf community (his mother was hard of hearing) and a solid training in phonetics. His father Alexander Melville Bell designed a set of feature-based phonetic alphabet named *visible speech symbols*, [1] published first by the father (Bell 1870, Bell 1881), and later by the son (Bell 1916) with detailed notes on articulatory phonetics. The decomposition of phones into features and their consistent usage made the system easy to learn, especially for the deaf. The combinatorial possibilities of the features and an additional set of symbols for non-speech sounds gave the system the capability to transcribe different languages, accents, and non-speech sounds.

While Alexander Graham Bell was exploring methods to capture the visual image of sound waves as a teaching aid for the deaf, he realized that if he could convert sound waves into electric currents, then he could transmit speech over a long distance. Bell's waveform display in 1874 was weak and lacked discernible details. Meaningful phonetic/acoustic study from waveforms was achieved 50 years later with an improved oscillograph, made possible with the advancement of sound amplification by vacuum tubes (H. D. Arnold), distortion control (Irving B. Crandall), and their incorporation in a condenser-type microphone (E. C. Wente). Crandall and Sacia (1924) and Crandall (1925) reported the waveforms, energy, and spectrum characteristics of English vowels, semivowels, and consonants from 4 male and 4 female speakers. The study was the first to define spectral properties that differentiate one

*We would like to thank Mohan Sondhi, Joe Hall, Lloyd Nakatani, Bernd Möbius, and Richard Sproat for helpful discussions in the process of writing this paper.

[1] For example, the consonant signs were open circles with the direction of the opening reminiscent of the place of articulation. Manner of articulations were indicated by embellishments on the open circle, such as using a bar for voicing and a wavy line for fricative.

sound from the other. Around the same time, Fletcher (1922) established the perceptually salient frequency band of each English speech sound using filtered speech as stimuli. He also studied extensively the interaction of intensity level and listener's recognition rate of each English sound.

Bell's dream of visualizing sound waves was finally realized in the sound spectrograph, which soon became an indispensable tool for phoneticians. Today, 50 years later, while the generation of spectrograms is done digitally rather than in analog circuitry, the principle and even the display of spectrograms remain the same. The development of the sound spectrograph at Bell Laboratories intensified due to potential military applications during the Second World War, including speech encryption and speaker identification—the voices of radio operators reveal information on troop movement. Because of that, all publications were held back until the war ended (Dudley and Gruenz 1946, Koenig *et al.* 1946, Kopp and Green 1946, Riesz and Schott 1946, Steinberg and French 1946). A year later, the *Visible Speech* (Potter *et al.* 1947) was published—a namesake of Melville Bell's phonetic alphabet. This book is simultaneously a historical account of the spectrograph project and a technical manual for spectrogram reading. Visual cues to the interpretation of spectrograms were exemplified with ample explanation on their correlations to acoustic properties. The idea of *hubs*, target formant values of sounds, was developed to explain coarticulation effects and the resulting variations in spectrograms. The authors even designed a set of phonetic symbols derived from formant structures to facilitate speed reading of spectrograms.

Bell Labs continued its efforts to study speech sounds, production and perception cues, acoustic and articulatory models (French and Steinberg 1947, Potter and Peterson 1948, Potter and Steinberg 1950), and pattern playback synthesis (Schott 1948). Peterson and Barney (1952) reported results from a database of 76 speakers, including men, women, and children, speaking English words with 10 different vowels in the h_d context, such as "heed", "hid", and "head". Auditory identifications of all stimuli from 70 listeners were also obtained. The paper included informative figures of the formant space of all sounds, the average F0, F1, F2, and F3 values of all vowels by men, women, and children, and explored factors that lead to vowel mis-identification by listeners. Dunn (1950) calculated vowel resonances from vocal tract dimensions, compared them to measurements from natural speech, and tested them with an electronic vocal tract model.

In 1940, Farnsworth succeeded in taking high speed motion pictures of vocal cords during speech (Farnsworth 1940), providing a valuable source of information for the study of vocal cord movements. Miller (1959) devised an electronic inverse filter that removes the vocal tract resonances, thereby revealing the glottal waveform.

The understanding of the principles of human speech naturally leads to a desire to simulate human speech. Speech synthesis, and the later text-to-speech system, is another research topic that has been carried out in tandem with phonetics research at Bell Laboratories. The creation and the improvement of a speech synthesizer requires all aspects of phonetic and acoustic knowledge. At the same time, a working synthesizer is a good testing ground for phonetic and acoustic hypotheses.

The first electronic speech synthesizer, the *Voder* (Dudley *et al.* 1939), was developed by Homer Dudley at Bell Laboratories, and was demonstrated at the World's Fair in 1939 in

New York. The Voder was a keyboard-like instrument with keys, switches, and pedals controlling variations in vowel resonance, sound source, and pitch. The picture of the Voder can be found on the Bell Labs Archive web site http://www.lucent.com/museum/1936scs.html, and a speech sample (as well as many other speech synthesis samples up to 1987) was collected in the synthesizer archive of Dennis Klatt (1987), and can be downloaded from http://www.icsi.berkeley.edu/eecs225d/klatt.html. Soon after the Voder, Dudley developed the *Vocoder* which can play back speech from phonetic specifications (Dudley 1939). Dudley also contributed to an interesting article reviewing the early history of phonetic alphabets, speech acoustics, and mechanical speech synthesis (Dudley and Tarnoczy 1950).

Digital synthesizers were developed soon after the introduction of the digital computer. Kelly and Gerstman (1961) implemented a digital formant synthesizer, which takes phonetic input and computes three formant resonators and pitch. Kelly and Lochbaum (1962) described the first articulatory synthesizer, replacing the earlier formant models with vocal-tract models. Later articulatory synthesis include Coker (1968) and Flanagan *et al.* (1975).

Fujimura employed fiberoptic and x-ray microbeam (Fujimura 1977, Fujimura *et al.* 1979, Fujimura 1980) to take measurements and to construct articulatory models of vocal cords, nasal and stop consonants, larynx, and the tongue. The microbeam research was extended to allophonic variations of English /l/ (Sproat and Fujimura 1993).

The first text-to-speech system at Bell Labs was developed by Coker *et al.* (1973). English text was converted to phonetic input by the use of a dictionary, the output was then sent to a formant synthesizer. The idea of concatenative speech synthesis—connecting segments of stored real speech to create new sentences—was proposed in the fifties by Perterson *et al.* (1958). The idea was tested with magnetic tapes. A digital concatenative synthesizer was implemented by Olive (1977) when computers had gained enough memory and processing power. This system was the predecessor of the current Bell Labs text-to-speech system. Duration models (Umeda 1975, Umeda 1977), F0 models (Olive 1975), pronunciation variations (Umeda and Coker 1974), and glottal flow models (Rosenberg 1971) were developed and were used in the text-to-speech systems. At the same time, studies by Nakatani *et al.* (Olive and Nakatani 1974, Nakatani and Dukes 1977, Nakatani and Schaffer 1978, Nakatani 1981) established the role of prosody in production, comprehension, and the naturalness of synthetic speech. They showed speech concatenation may improve the intelligibility of synthesized speech, but the naturalness of the system lies in better understanding and modeling of prosody.

Pierrehumbert (1980) developed a formal grammar of English intonation. Intonational contours were represented as tonal sequences, and were classified into pitch accents, phrase accents, and boundary tones. Liberman and Pierrehumbert (1984) explored several factors affecting the phonetic implementation of pitch accents, such foreground and background reading, downstep, and final lowering. Hirschberg (1992) linked the pitch accent predictions to discourse structure. The model was generalized to Japanese (Pierrehumbert and Beckman 1988) and to Chinese (Shih 1988). An implementation of this model for the English text-to-speech system is described in Anderson *et al.* (1984), and the implementation for Chinese, Navajo and Japanese in Sproat (1998).

The current text-to-speech system of Bell Labs includes the following languages/dialects:

English, German, Chinese (Mandarin and Taiwanese), Spanish (Latin American and Castillian), French, Russian, Italian, Japanese, Romanian, and Navajo. Many of these systems are accessible on the web: http://www.bell-labs.com/project/tts/.

Behind each language there is an extensive research effort, including the collection of speech databases, typically of thousands of sentences, containing multiple instances of all possible phone-to-phone sequences and many triphone sequences in the language. The databases offer valuable information on phone inventories and coarticulation effects. Olive *et al.* (1993) provides detailed discussions of all sound transition patterns in English with illustrations of waveforms and spectrograms. Sproat (1998) reported F1/F2 space of vowel inventories of a few languages, and Shih and Sproat (1996) reported the Chinese data.

The current duration models and a set of tools for analysis were developed by Jan van Santen for English (van Santen 1992, van Santen 1994), and applied successfully to other languages (Shih and Ao 1996, Möbius and van P. H. Santen 1996). Many interacting factors affecting duration were considered, and greedy algorithms were employed to choose text materials for the duration database that provide maximum coverage of desired factor combinations. A central idea of the theory, with support from multilingual data, is that most of the durational effects are *monotonic* in nature—the relative durational scale of members of phone classes, such as voiceless fricatives, tends to be preserved under different contexts—therefore the durational variation in speech can be captured by additive or multiplicative models. In cases where factors interact, duration can be predicted by sums-of-products models (van Santen 1993).

The current intonation model, with precise alignment of accent curves with the segmental material, were developed by van Santen (van Santen and Möbius 1997, van Santen *et al.* 1998). The model has been successfully applied to Germanic and Romance languages, as well as Russian.

We have often been asked in the past why a telecommunications company is interested in phonetics research, and what the current direction is. The application of phonetics research has taken many unexpected turns over the years. In the early days, the understanding of speech and hearing is vital to the design of telephone equipments, which must convert and amplify speech with minimum distortion in the frequency range that is important to human speech production and perception. There is also a continuous quest for a cost-efficient way of speech compression for the purpose of data transmission. Dudley's vocoder was conceived as such a system: speech was converted into phonetic specifications and could be synthesized from such specifications. Then only the specifications, not the speech, need to be transmitted over the phone line, at a much narrower bandwidth. As new technologies evolved, the scope of phonetics research has widen to include high quality text-to-speech systems and speech recognition systems (Riley and Ljolje 1991, Giachin *et al.* 1991) as a service on today's communication network.

For readers who would like to delve into the subject more deeply, we highly recommend the following books: *A History of Engineering and Science in the Bell System: The Early Years (1875–1925)*, and *A History of Engineering and Science in the Bell System: Communications Sciences (1925–1980)*. These are two volumes of a set of five, which summarize Bell Labs research up to the eighties. The chapters were written by experts in the fields,

with fascinating technical and historical details that are accessible to general readership.

Fletcher (1953) presented work on phonetics, acoustics, loudness, and hearing as integral components of the human communication network, and how the advancement of these areas impacted telecommunications. The book gave detailed summaries of works by Crandall *et al.*, Potter *et al.*, and Farnsworth. The 1997 reprint by the Acoustical Society of America makes this valuable resource available again. For easier reading on the same subject, there are two popular books, also from Bell Labs researchers, *Man's World of Sound* (Pierce and David 1958), and *Speech Chain* (Denes and Pinson 1963).

Flanagan (1972) gave an excellent account of speech synthesis works from the earliest records up to the seventies. Sproat (1998), written by the current members of the TTS team, describes the Bell Labs multilingual text-to-speech system, which is a modern product made possible by a tradition of speech research that goes back more than a hundred years.

REFERENCES

Mark Anderson, Janet Pierrehumbert, and Mark Liberman. Synthesis by Rule of English Intonation Patterns. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, Vol. 1, (San Diego, CA, USA), pp. 2.8.1–2.8.4, ICASSP, 1984.

Alexander Melville Bell. *Explanatory Lecture on Visible Speech.* Simpkin, Marshall, & Co., London, 1870.

Alexander Melville Bell. *Sounds and Their Relations, A Complete Manual of Universal Alphabetics.* J. H. Choates & Co., Salem, Mass, 1881.

Alexander Graham Bell. *The Mechanism of Speech.* Funk & Wagnalls Company, New York and London, 1916.

Cecil H. Coker, Noriko Umeda, and Catherine P. Browman. Automatic Synthesis from Ordinary English Text. *IEEE Transactions of Acoustics, Speech and Signal Processing*, Vol. 21, pp. 293–297, 1973.

Cecil H. Coker. Speech Synthesis with a Parametric Articulatory Model. In *Speech Synthesis* (J. L. Flanagan and L. R. Rabiner, editors), pp. 135–139, Stroudsburg, PA: Dowden, Hutchinson and Ross, 1968.

I. B. Crandall and C. F. Sacia. A Dynamical Study of the Vowel Sounds. *Bell System Technical Journal*, April 1924.

Irving B. Crandall. The Sounds of Speech. *Bell System Technical Journal*, October 1925.

Peter B. Denes and Elliot N. Pinson. *The Speech Chain: The Physics and Biology of Spoken Language.* Bell Telephone Laboratories, Inc., 1963.

Homer Dudley and Otto O. Gruenz. Visible Speech Translators with External Phosphors. *Journal of Acoustical Society of America*, Vol. 18, No. 1, pp. 62–73, 1946.

Homer Dudley and T. H. Tarnoczy. The Speaking Machine of Wolfgand von Kempelen. *Journal of Acoustical Society of America*, Vol. 22, pp. 151–166, 1950.

H. Dudley, R. Riesz, and S. Watkins. A Synthetic Speaker. *Journal of the Franklin Institute*, Vol. 227, pp. 739–764, 1939.

Homer Dudley. The Vocoder. *Bell Laboratories Record*, Vol. 17, pp. 122–126, 1939.

H. K. Dunn. The Calculation of Vowel Resonances, and an Electrical Vocal Tract. *Journal of Acoustical Society of America*, Vol. 22, pp. 740–753, 1950.

D. W. Farnsworth. High-speed Motion Pictures of the Human Vocal Cords. *Bell Laboratories Record*, Vol. 18, pp. 203–208, 1940.

J. L. Flanagan, K. Ishizaka, and K. L. Shipley. Synthesis of Speech From a Dynamic Model of the Vocal Cords and Vocal Tract. *The Bell System Technical Journal*, Vol. 54, pp. 485–506, March 1975.

James L. Flanagan. *Speech Analysis: Synthesis and Perception.* Springer-Verlag, New York, 1972.

Harvey Fletcher. The Nature of Speech and its Interpretation. *Journal of the Franklin Institute*, Vol. 193, No. 6, pp. 729–747, 1922.

Harvey Fletcher. *Speech and Hearing in Communication.* Van Nostrand Company, Inc., 1953.

N. R. French and J. C. Steinberg. Factors Governing the Intelligibility of Speech Sounds. *Journal of Acoustical Society of America*, Vol. 19, No. 1, pp. 90–119, 1947.

O. Fujimura, T. Baer, and S. Niimi. A Stereo-fiberscope with a Magnetic Interlens Bridge for Laryngeal Observation. *Journal of Acoustical Society of America*, Vol. 65, pp. 478–480, 1979.

Osamu Fujimura. Control of the Larynx in Speech. *Phonetica*, Vol. 34, pp. 280–288, 1977.

Osamu Fujimura. Modern Methods of Investigation in Speech Production. *Phonetica*, Vol. 37, pp. 38–54, 1980.

E. P. Giachin, C.-H. Lee, and A. E. Rosenberg. Word Juncture Modeling Using Phonological Rules for HMM-based Continuous Speech Recognition. *Computer Speech and Language*, Vol. 5, pp. 155–168, April 1991.

Julia Hirschberg. Using Discourse Context to Guide Pitch Accent Decisions in Synthetic Speech. In *Talking Machines: Theories Models and Applications* (Gérard Bailly and Christian Benoit, editors), pp. 367–376, Amsterdam: North Holland, 1992.

John Kelly and Louis Gerstman. An Artificial Talker Driven from Phonetic Input. *Journal of Acoustical Society of America*, Vol. 33, p. 835, 1961.

John Kelly and C. Lochbaum. Speech Synthesis. In *Proceedings from the Fourth International Congress on Acoustics*, pp. 1–4, 1962.

Dennis H. Klatt. Review of Text-to-Speech Conversion for English. *Journal of Acoustical Society of America*, Vol. 82, No. 3, pp. 737–793, 1987.

W. Koenig, H. K. Dunn, and L. Y. Lacy. The Sound Spectrograph. *Journal of Acoustical Society of America*, Vol. 18, No. 1, pp. 19–49, 1946.

G. A. Kopp and H. C. Green. Basic Phonetic Principles of Visible Speech. *Journal of Acoustical Society of America*, Vol. 18, No. 1, pp. 74–89, 1946.

Mark Y. Liberman and Janet B. Pierrehumbert. Intonational Invariance under Changes in Pitch Range and Length. In *Language Sound Structure* (M. Aronoff and R. Oehrle, editors), pp. 157–233, Cambridge, Massachusetts: M.I.T. Press, 1984.

R. L. Miller. Nature of the Vocal Cord Wave. *Journal of Acoustical Society of America*, Vol. 31, pp. 667–679, 1959.

Bernd Möbius and Jan van P. H. Santen. Modeling Segmental Duration in German Text-to-Speech Synthesis. In *ICSLP*, Vol. 4, pp. 2395–2398, 1996.

L. H. Nakatani and K. D. Dukes. Locus of Segmental Cues for Word Junctures. *Journal of Acoustical Society of America*, Vol. 62, pp. 714–719, 1977.

L. H. Nakatani and J. Schaffer. Hearing Words without Words: Prosodic Cues for Word Perception. *Journal of Acoustical Society of America*, Vol. 63, pp. 234–244, 1978.

Lloyd H. Nakatani. Prosodic Aspects of American English Speech Rhythm. *Phonetica*, Vol. 38, pp. 84–105, 1981.

Joseph P. Olive and Lloyd H. Nakatani. Rule-synthesis of Speech by Word Concatenation: A First Step. *Journal of Acoustical Society of America*, Vol. 55, No. 3, pp. 660–666, 1974.

J. P. Olive, A. Greenwood, and J. S. Coleman. *Acoustics of American English Speech.* Springer-Verlag, 1993.

Joseph P. Olive. Fundamental Frequency Rules for the Synthesis of Simple Declarative English Sentences. *Journal of Acoustical Society of America*, Vol. 57, pp. 476–482, 1975.

Joseph P. Olive. Rule Synthesis of Speech from Diadic Units. *ICASSP*, Vol. 77, pp. 568–570, 1977.

Gordon E. Peterson and Harold L. Barney. Control Methods Used in a Study of the Vowels. *Journal of Acoustical Society of America*, Vol. 24, No. 2, pp. 175–184, 1952.

Gordon E. Peterson, William S-Y. Wang, and Eva Sivertsen. Segmentation Techniques in Speech Synthesis. *Journal of Acoustical Society of America*, Vol. 30, pp. 739–742, 1958.

J. R. Pierce and E. E. David. *Man's World of Sound.* Double Day, 1958.

Janet Pierrehumbert and Mary Beckman. *Japanese Tone Structure.* The MIT Press, Cambridge, Massachussetts, 1988.

Janet Pierrehumbert. *The Phonology and Phonetics of English Intonation.* PhD thesis, MIT, 1980.

R. K. Potter and G. E. Peterson. The Representation of Vowels and their Movements. *Journal of Acoustical Society of America*, Vol. 20, p. 528, 1948.

R. K. Potter and J. C. Steinberg. Toward the Specification of Speech. *Journal of Acoustical Society of America*, Vol. 22, No. 6, pp. 807–820, 1950.

Ralph K. Potter, George A. Kopp, and Harriet C. Green. *Visible Speech.* D. Van Nostrand Company, Inc., New York, 1947.

R. R. Riesz and L. Schott. Visible Speech Cathode-Ray Translator. *Journal of Acoustical Society of America*, Vol. 18, No. 1, pp. 50–61, 1946.

Michael D. Riley and Andrej Ljolje. Lexical Access with a Statistically-Derived Phonetic Network. In *Eurospeech 91*, Vol. 2, pp. 585–588, 1991.

Aaron Rosenberg. Effect of Glottal Pulse Shape on the Quality of Natural Vowels. *Journal of Acoustical Society of America*, Vol. 49, pp. 583–590, 1971.

L. O. Schott. A Playback for Visible Speech. *Bell Laboratories Record*, Vol. 26, pp. 333–339, 1948.

Chilin Shih and Benjamin Ao. Duration Study for the Bell Laboratories Mandarin Text-to-Speech System. In *Progress in Speech Synthesis* (Jan van Santen, Richard Sproat, Joseph Olive, and Julia Hirschberg, editors), pp. 383–399, New York: Springer, 1996.

Chilin Shih and Richard W. Sproat. Issues in Text-to-Speech Conversion for Mandarin. *Computational Linguistics and Chinese Language Processing*, Vol. 1, No. 1, pp. 37–86, 1996.

Chilin Shih. Tone and Intonation in Mandarin. In *Working Papers of the Cornell Phonetics Laboratory, Number 3: Stress, Tone and Intonation*, pp. 83–109, Ithaca, NY, USA: Cornell University, 1988.

Richard W. Sproat and Osamu Fujimura. Allophonic Variation in English /l/ and its Implications for Phonetic Implementation. *Journal of Phonetics*, Vol. 21, No. 3, pp. 291–311, 1993.

Richard W. Sproat, editor. *Multilingual Text-to-Speech Synthesis: The Bell Labs Approach.* Kluwer Academic Publishers, 1998.

J. C. Steinberg and N. R. French. The Portrayal of Visible Speech. *Journal of Acoustical Society of America*, Vol. 18, No. 1, pp. 4–18, 1946.

Noriko Umeda and Cecil H. Coker. Allophonic Variation in American English. *Journal of Phonetics*, Vol. 2, pp. 1–5, 1974.

Noriko Umeda. Vowel Duration in American English. *Journal of Acoustical Society of America*, Vol. 58, pp. 434–445, 1975.

Noriko Umeda. Consonant Duration in American English. *Journal of Acoustical Society of America*, Vol. 61, pp. 846–858, 1977.

Jan P. H. van Santen and Bernd Möbius. Modeling Pitch Accent Curves. In *Intonation: Theory, Models and Applications—Proceedings of an ESCA Workshop*, (Athens), pp. 321–324, ESCA, Sept. 18–20 1997.

Jan P. H. van Santen, Chilin Shih, and Bernd Möbius. Intonation. In *Multilingual Text-to-Speech Synthesis: The Bell Labs Approach* (R. Sproat, editor), Boston: Kluwer Academic Publishers, 1998.

Jan P. H. van Santen. Contextual Effects on Vowel Duration. *Speech Communication*, Vol. 11, pp. 513–546, 1992.

Jan P. H. van Santen. Analyzing N-way Tables with Sums-of-Products Models. *Journal of Mathematical Psychology*, Vol. 37, pp. 327–371, 1993.

Jan P. H. van Santen. Assignment of Segmental Duration in Text-to-Speech Synthesis. *Computer Speech and Language*, Vol. 8, pp. 95–128, April 1994.