Chapter 2. Speech Processing Research Program

Academic and Research Staff

Professor Jae S. Lim, Giampiero Sciutto

Graduate Students

Michael S. Brandstein, Shiufun Cheung, John C. Hardwick, Katherine S. Wang, Chang Dong Yoo

Technical and Support Staff

Debra L. Harring, Cynthia LeBlanc

2.1 Introduction

The objective of this research program is to develop methods for solving important speech communication problems. Current research topics in progress include development of a new speech model and algorithms to (1) enhance speech degraded by background noise and (2) modify the time scale of speech. We are also investigating methods for displaying spectrograms more efficiently.

2.2 Development of a 1.5 Kbps Speech Vocoder

Sponsor

National Science Foundation Fellowship

Project Staff

Michael S. Brandstein, Professor Jae S. Lim

The recently developed Multi-Band Excitation Speech Model has been shown to accurately reproduce a wide range of speech signals without many of the limitations inherent in existing speech model based systems.[1] The robustness of this model makes it particularly applicable to low bit rate, high quality speech vocoders. Griffin and Lim first described a 9.6 Kbps speech coder based on this model.[2] Later work resulted in a 4.8 Kbps speech coding system.[3] Both of these systems are capable of high quality speech reproduction in both low and high SNR conditions.

The purpose of this research was to explore methods for using the new speech model at the 1.5 Kbps rate. Results indicated that a substantial amount of redundancy exists between the model parameters. Research focused on exploiting redundancies to quantize these parameters more efficiently. Attempts were also made to simplify the existing model without significantly reducing speech quality.

This research was completed in June 1990.

2.3 A New Method for Representing Speech Spectrograms

Sponsors

National Science Foundation
Grant MIP 87-14969
U.S. Navy - Office of Naval Research
Contract N00014-89-J-1489

Project Staff

Shiufun Cheung, Professor Jae S. Lim

The spectrogram, which is a two-dimensional time-frequency display of a one-dimensional signal, is used extensively in speech research. Existing spectrograms are generally divided into

---

[1] D.W. Griffin and J.S. Lim, "A New Model-Based Speech Analysis/Synthesis System," *IEEE International Conference on Acoustic, Speech and Signal Processing,* Tampa, Florida, March 26-29, 1985, pp. 513-516.

[2] D.W. Griffin and J.S. Lim, "A High Quality 9.6 Kbps Speech Coding System," *IEEE International Conference on Acoustic, Speech and Signal Processing,* Tokyo, Japan, April 8-11, 1986.

[3] J.C. Hardwick, "A 4.8 Kbps Multi-Band Excitation Speech Coder," S.M. thesis, Dept. of Electr. Eng. and Comput. Sci., MIT, 1988.

two types, wideband spectrograms and narrowband spectrograms, according to the bandwidth of the analysis filters used to generate them. Due to the different characteristics of the two types of spectrograms, they are employed for different purposes. The wideband spectrogram is valued for its quick temporal response and is used for word boundary location and formant tracking. On the other hand, the narrowband spectrogram, with its high frequency resolution, is primarily used for measuring the pitch frequency.

Various attempts have been made to improve the spectrographic display. Past efforts include the development of neural spectrograms which use critical bandwidth analysis filters in the imitation of the human auditory system and the development of better time-frequency distributions such as the Wigner distribution.

In this research, we propose a different approach. The spectrogram is viewed as a two-dimensional digital image instead of a transformed one-dimensional speech signal. Image processing techniques are used to create an improved spectrogram which preserves the desirable visual features of the wideband and narrowband spectrograms. This transforms a speech processing problem into an image processing problem.

At this point, we have developed a simple but effective method for combining the two types of spectrograms. In this method, each pixel of the combined spectrogram is the geometric mean of corresponding pixel values of the two original spectrograms. Apart from the geometric-mean merge, we are also experimenting with the use of color and other merge algorithms. Initial results are promising.

2.4 A Dual Excitation Speech Model

Sponsors

U.S. Air Force - Electronic Systems Division
Contract F19628-89-K-0041
U.S. Navy - Office of Naval Research
Contract N00014-89-J-1489

Project Staff

John C. Hardwick, Professor Jae S. Lim

One class of speech analysis/synthesis system (vocoder) which has been extensively studied and used in practice is based on an underlying model of speech. Even though traditional vocoders have been quite successful in synthesizing intelligible speech, they have not been successful in synthesizing high quality speech. The Multi-Band Excitation (MBE) speech model, introduced by Griffin, improves the quality of vocoder speech through the use of a series of frequency dependent voiced/unvoiced decisions. The MBE speech model, however, still results in a loss of quality when compared with the original speech. This degradation is caused in part by the voiced/unvoiced decision process. A large number of frequency regions contain a substantial amount of both voiced and unvoiced energy. If a region of this type is declared voiced, then a tonal or hollow quality is added to the synthesized speech. Similarly, if the region is declared unvoiced, then additional noise occurs in the synthesized speech. As the signal-to-noise ratio decreases, the classification of speech as either voiced or unvoiced becomes more difficult, and, consequently, the degradation is increased.

A new speech model has been proposed in response to the aforementioned problems. This model is referred to as the Dual Excitation (DE) speech model, due to its dual excitation and filter structure. The DE speech model is a generalization of most previous speech models, and, with the proper selection of the model parameters, it reduces to either the MBE speech model or to a variety of more traditional speech models.

Current research is examining the use of this speech model for speech enhancement, time scale modification and bandwidth compression. Additional areas of study include further refinements to the model and improvements in the estimation algorithms.

2.5 Speech Enhancement Techniques for the Dual Excitation Vocoder Model

Sponsors

National Science Foundation Fellowship
U.S. Air Force - Electronic Systems Division
Contract F19628-89-K-0041
U.S. Navy - Office of Naval Research
Contract N00014-89-J-1489

Project Staff

Katherine S. Wang, Professor Jae S. Lim

We explored some conventional methods for speech enhancement in the presence of additive

white noise[4] in a new framework where the voiced estimation provided by the MBE model[5] allowed us to perform noise reduction separately on voiced and unvoiced components.

Conventional methods which take advantage of the periodic structure of voiced speech include comb filtering and adaptive noise cancellation. A technique based on short-time spectral amplitude estimation obtains the minimum mean-square-error, linear estimator of the speech signal by non-causal Wiener filtering, which we could approximate by an adaptive Wiener filtering technique. Speech enhancement can also be model based, for example, by using classical estimation theory applied to an all pole model of speech. We drew from some of these conventional techniques to create a speech enhancement system customized to the traits of the Multi-Band Excitation (MBE) Vocoder and, subsequently, to the Dual Excitation Model.[6] The noiselike characteristics of unvoiced speech and the harmonic structure of voiced speech suggest that noise reduction can most effectively be done on speech that has been separated into the two components, rather than attempting to categorize the frequency band as purely voiced and unvoiced.

The recently developed Multi-Band Excitation (MBE) Speech Model has been shown to accurately reproduce a wide range of speech signals without many of the limitations inherent in existing speech model based systems.

This research was completed in August 1990.

2.6 Nonlinear and Statistical Approach to Speech Synthesis

Sponsor

U.S. Navy - Office of Naval Research
Contract N00014-89-J-1489

Project Staff

Chang Dong Yoo, Professor Jae S. Lim

Numerous speech models have been proposed, and most of these models have been incorporated with some variation into the basic speech production model where a linear time invariant system is excited by periodic or random pulses depending on voiced/unvoiced decision. But studies have indicated that speech production is a nonlinear process and that for some sounds such as the voiced fricatives, it is difficult to make a hard voiced/unvoiced decision. Discrepancy between actual speech production and production from the linear model system might stem from the Teager paradigm, considering the vocal tract operation from the air flow point of view.

For higher quality speech synthesis, we need to derive either an enhanced speech production model incorporating more of the nonlinear effects of speech production or a speech synthesis method totally divorced from the mechanism of speech production such as the hidden Mardov modeling of speech production.

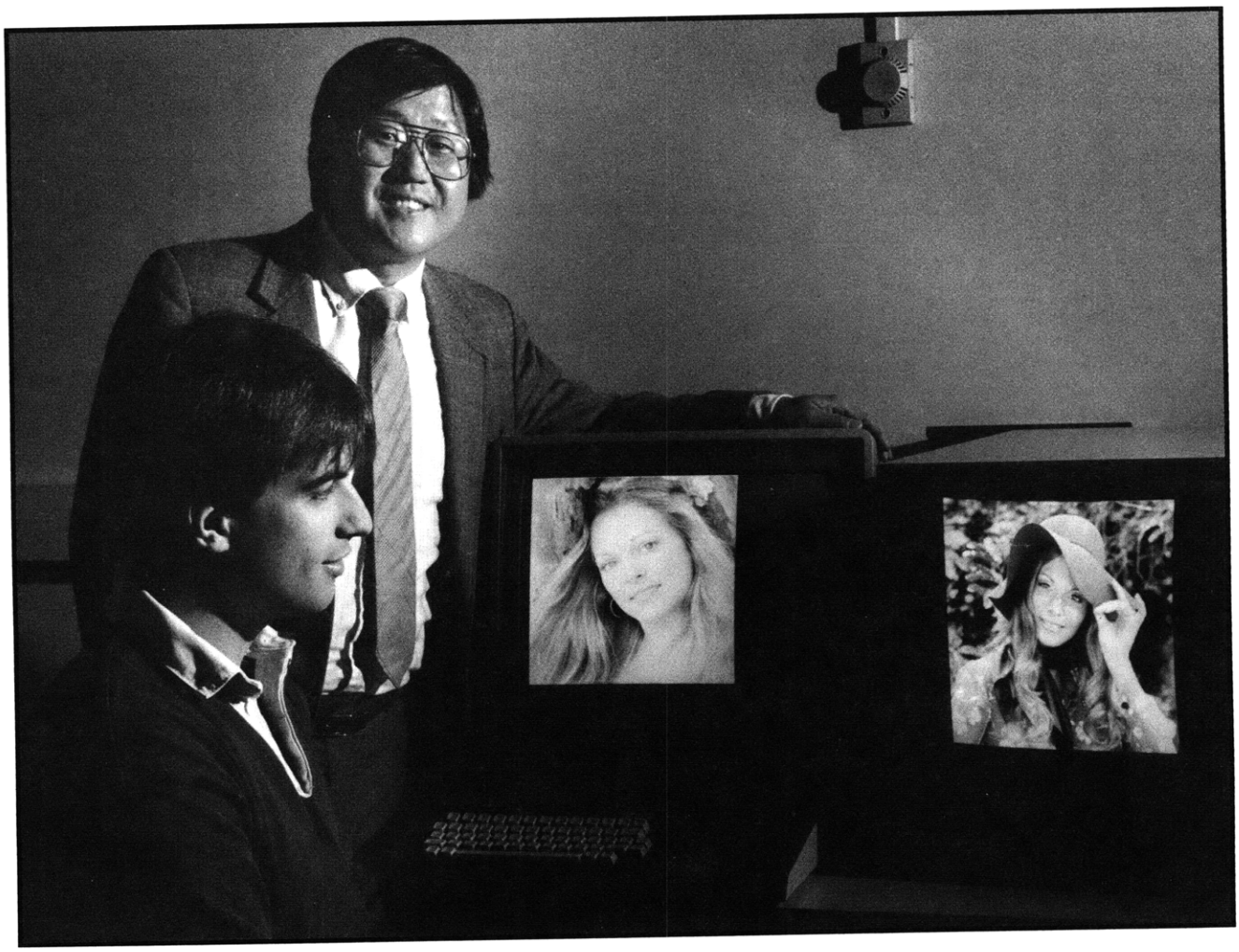
In this research, various nonlinear and statistical aspects of speech characteristics are being studied to derive a more efficient method of speech analysis/synthesis.

---

[4] J.S. Lim and A.V. Oppenheim, "Enhancement and Bandwidth Compression of Noisy Speech," *IEEE Proc.* 67 (12): 1586-1604 (1979); J.S. Lim, ed. *Speech Enhancement* (Englewood Cliffs, N.J.: Prentice Hall, 1983).

[5] D.W. Griffin and J.S. Lim, "A New Model-Based Speech Analysis/Synthesis System," *IEEE International Conference on Acoustics, Speech and Signal Processing,* Tampa, Florida, March 26-29, 1985, pp. 513-516.

[6] John Hardwick, Ph.D. research, MIT.



Professor Jae S. Lim (standing) and graduate student Matthew M. Bace demonstrate a new method that was recently developed to reduce channel degradation in the National Television Standards Committee system.