Chapter 2. Speech Processing Research Program

Academic and Research Staff

Professor Jae S. Lim, Giovanni Aliberti, Giampiero Sciutto

Graduate Students

Michael S. Brandstein, Shiufun Cheung, Warren Chou, John C. Hardwick, Rosalind W. Picard, Paul Shen, Katherine S. Wang

Technical and Support Staff

Hassan Gharavy, Cynthia LeBlanc, Julia Sharp

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 (1) real-time implementation of a multiband excitation vocoder operating at bit rates ranging between 2.4 and 8.0 kbits/sec and (2) development of algorithms for enhancing speech degraded by background noise and for modifying 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

Sponsors

National Science Foundation
Grant MIP 87-14969
National Science Foundation Fellowship
U.S. Air Force - Electronic Systems Division
Contract F19628-89-K-0041

Project Staff

Michael S. Brandstein, Professor Jae S. Lim

The recently developed Multi-Band Excitation Speech Model accurately reproduces 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. In Griffin and Lim,[2] a 9.6 Kbps speech coder based on this model was first described. Later work resulted in a 4.8 Kbps speech coding system.[3] We have shown that both of these systems are capable of high quality speech reproduction in both low and high SNR conditions.

The purpose of this research is to explore methods of using the new speech model at the 1.5 Kbps rate. Results indicate that a substantial amount of redundancy exists between the model parameters. Current research is focused on exploiting these redundancies to quantize these parameters more efficiently. Attempts are also under way to simplify the existing model without significantly reducing speech quality.

[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.

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

Since each of the spectrograms has its own advantages and weaknesses, efforts are being made to incorporate the desirable aspects of the two spectrograms into one display. One approach is to mimic the human auditory system and use critical band filters to generate neural spectrograms. Another method is to construct an optimum window, which is the temporal equivalent of the filter.

In this research, we propose an entirely different view of the problem. Instead of viewing the spectrogram as a transformed speech signal, we consider it an image or digital picture. Given this type of approach, the whole gamut of image processing techniques is available for use, and the problem becomes theoretically much simpler. In fact, the most interesting aspect of this research is that we transform a problem in speech processing into an image-processing problem.

2.4 A Dual Excitation Speech Model

Sponsors

National Science Foundation
Grant MIP 87-14969
U.S. Air Force - Electronic Systems DIvision
Contract F19628-89-K-0041

Project Staff

John C. Hardwick, Professor Jae S. Lim

One class of speech analysis/synthesis systems (vocoders) which have been extensively studied and used in practice are based on an underlying model of speech. Even though traditional vocoders have been quite successful in synthesizing intelligible speech, they have not successfully synthesized 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, results in a loss of quality as compared to 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 increases.

We have proposed a new speech model in response to the problems mentioned above. This model is referred to as the Dual Excitation (DE) speech model, because of its dual excitation and filter structure. The DE speech model is a generalization of most previous speech models which, with the proper selection of the model parameters, reduces to either the MBE speech model or to a variety of more traditional speech models.

Currently, our 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 Image Texture Modeling

Sponsors

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

Project Staff

Rosalind W. Picard, Professor Jae S. Lim

Textured regions are the nemesis of many image processing algorithms. For example, algorithms for image segmentation or image compression usually assume stationarity or high correlation of the image gray-level data, failing in textured regions. In this research, we are developing a model with only a small number of parameters which synthesizes the textures. This model would have applications for image enhancement, segmentation, scene synthesis, and low bit rate image coding.

The ability to synthesize syntactically regular "structural" textures, e.g., a tiled wall, as well as the more random "stochastic" textures, e.g., shrubbery, is desirable in a texture model. Most models assume only one of these cases and perform poorly when the data is not perfectly periodic or stochastic. Our goal is to develop a model which is capable of exhibiting a continuum of structural and stochastic behavior.

We have confirmed the fact that the Markov Random Field model, which is equivalent to the Gibbs Distribution of statistical mechanics, synthesizes a wide variety of stochastic textures. By using concepts from statistical mechanics, we have modified the Gibbs distribution so that it synthesizes a greater variety of textures. One modification, imposing an "external field," allows discontinuities to appear in the texture. We have explored other modifications which simulate "physical" processes. Recently, by using dominant spectral coefficients, we have induced some structure in the random field. We are currently investigating the limitations of this mixed structural and stochastic model by estimating the structural and stochastic parameters of natural textures and trying to synthesize visually similar textures.

2.6 Speech Enhancement Techniques for the Dual Excitation Vocoder Model

Sponsors

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

Project Staff

Katherine S. Wang, Professor Jae S. Lim

We are exploring 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] allows 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 noncausal Wiener filtering, which we could approximate by an adaptive Wiener filtering technique. Speech enhancement can also be model based, such as using classical estimation theory applied to an all pole model of speech. We will draw from some of these conventional techniques to create a speech

[4] J.S. Lim and A.V. Oppenheim, "Enhancement and Bandwidth Compression of Noisy Speech," *Proc. IEEE* 67 (12): (1979); ed. J.S. Lim, *Speech Enhancement.* (Englewood Cliffs, New Jersey: 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.

enhancement system customized to the traits of the Multi-Band Excitation (MBE) Vocoder and, subsequently, to the Dual Excitation Model.[6] The noise-like characteristics of unvoiced speech and the harmonic structure of voiced speech suggest that noise can most effectively be reduced in speech that has been separated into the two components, rather than attempting to categorize the frequency band as purely voiced and unvoiced.

We have shown that the recently developed Multi-Band Excitation (MBE) Speech Model accurately reproduces a wide range of speech signals without many of the limitations inherent in existing speech model based systems.

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