Knowledge Based Speech Analysis and Enhancement*

Cory Myers, Alan Oppenheim, Randall Davis, and Webster Dove

Massachusetts Institute of Technology
Research Laboratory of Electronics
Cambridge, MA 02139

ABSTRACT

This paper describes a system for speech analysis and enhancement which combines signal processing and symbolic processing in a closely coupled manner. The system takes as input both a noisy speech signal and a symbolic description of the speech signal. The system attempts to reconstruct the original speech waveform using symbolic processing to help model the signal and to guide reconstruction. The system uses various signal processing algorithms for parameter estimation and reconstruction.

Introduction

As part of our current research program in knowledge-based signal processing we are developing a system for knowledge-based speech analysis and enhancement [1,2]. The system accepts as input both a noisy speech signal and a symbolic description of the speech utterance. The symbolic description of the signal includes information about the speaker, such as age and gender, and information about the recording environment, such as noise characterization, sampling rate and signal bandwidth. The system is also given a symbolic description of the content of the signal in the form a time-aligned phonetic transcript. The output from the system is both a reconstructed version of the original speech signal and a model of the spectral envelope of the original speech signal. Such a system would provide a useful tool for the restoration of noisy recordings of historical or archival value. For valuable recordings one desires a reconstruction of the original waveform and it is not unreasonable to expend the effort in extracting a symbolic description of the utterance.

While there are potentially many applications where a system such as ours would be useful, the problem of designing such a system was chosen principally because it contains a balance between signal processing and symbolic processing and thus can serve as a vehicle for development of techniques for integrating signal and symbolic processing. There is a rich history of study into the problems of speech analysis, synthesis and reconstruction as well as into problems of acoustic-phonetics and linguistics. Many signal processing tools are available and much is known about their behavior [3,4]. Researchers have extensively examined acoustic-phonetic properties of speech and symbolic descriptions of some acoustic-phonetic features have been collected. Numerical measurements of some acoustic-phonetic features have also been performed. Symbolic reasoning problems related to speech synthesis, such as synthesis-by-rule, have been studied [5].

The basic approach used by the system to reconstruct the original speech waveform from the noisy version is to extract a parametric model of the original speech signal, using both the available noisy signal and the phonetic transcript, and then to resynthesize from this parametric model. The parametric model is chosen to represent the spectral envelope of the underlying speech signal contained in the noisy utterance. Explicit modeling of the associated underlying spectral envelope was chosen because of the large amount of knowledge relating to the problem. Researchers have collected much knowledge, both symbolic knowledge and signal processing knowledge, relating to the problem of spectral modeling so the problem presents a good opportunity to explore interaction between signal processing and symbolic processing. The current version of the system does not extract excitation parameters. Information about the excitation is supplied either from a hand edited pitch track or from an excitation modeling algorithm run on the original speech waveform.

Knowledge about Speech Analysis and Reconstruction

Researchers have collected large amounts of knowledge which can be applied to the problem of parameterizing the spectral envelope of a speech signal in a noisy environment. Many different signal processing algorithms exist for the modeling of speech. The performance and use of these algorithms has been studied and consequently there is available a large knowledge base to guide the modeling. Information relating acoustic-phonetic properties of speech sounds to the spectral envelope has been collected and the importance of different spectral features identified. Speech synthesis systems also often build explicit models of the spectral envelope. Many methods of selecting and interpolating speech parameters have been used in synthesis systems and researchers have learned much about the important parameters for speech synthesis.

Much knowledge about speech reconstruction also exists. The importance of different speech sounds in intelligibility is well known and the ability to recognize different sounds in the presence of noise has been studied. Researchers have developed many different algorithms for the reconstruction of a speech waveform from a noisy version of the waveform.

*This work has been supported in part by Sanders Associates Incorporated, in part by the Advanced Research Projects Agency monitored by ONR under Contract N00014-81-K-0742 NR-049-506 and in part by the National Science Foundation under Grant ECS80-07102.

CH1945-5/84/0000-0423 $1.00

Similarly, speech synthesis from a phonetic transcript has also been extensively examined. Synthesis-by-rule systems also contain much information about how to make speech waveforms given symbolic descriptions of the speech signal.

Our system attempts to utilize many forms of knowledge relating to speech analysis and synthesis and to make explicit the representation of the knowledge in the system. In developing our system we have attempted to collect much knowledge relating to the problems of speech analysis, synthesis and reconstruction. The collection of knowledge is by no means complete and only those those pieces of knowledge which seemed amenable to our approach were incorporated into the system. Some examples of the types of information that we use in our system are the following:

- Linear Predictive Coding is known to provide a good model for non-nasalized voiced sounds [6].
- A reasonable window size for LPC analysis in a vowel is between 15 and 20 msec [6].
- The average first formant location for the vowel /i/ when spoken by a male speaker is 270Hz [7].
- A primary difficulting in understanding speech in noisy environments is the difficulty in differentiating among stops and fricatives [8].

System Strategy

The current system emphasizes LPC spectral analysis and standard speech synthesis-by-rule techniques. Figure 1 shows the flow of information within the system. In processing an utterance the system first attempts to decide on how to perform LPC analysis on the utterance. Rather than using a fixed LPC analysis method the system varies the LPC analysis according to the context. The system first divides the utterance up into finer regions than specified by the phonemes. A region is meant to be a portion of the utterance over which it is reasonable to do LPC analysis with a fixed set of parameters. Initially regions are generated from the phonetic transcript by breaking segments down into parts according to rules which estimate the size of these parts. The number, type and sizes of the regions generated from a phoneme depend on the characteristics of that phoneme. For example, a vowel is typically broken down into an initial transitory region, a middle stationary region and a final transitory region. Since the sizes of these regions can not be predicted exactly the system makes reasonable, potentially overlapping, guesses as to their locations. Following the initial generation of regions, the regions are refined according to the context in which they lie. For example, one rule states that the aspiration following a voiced stop is reduced before a vowel.

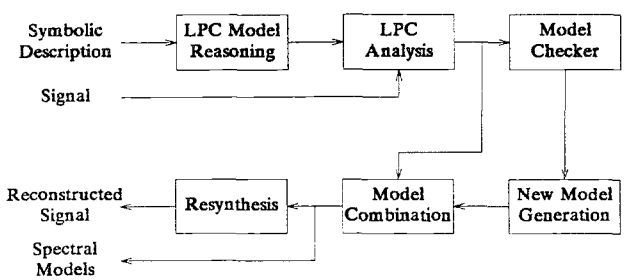

Figure 1. Information flow within the system.

Once the waveform has been broken down into regions the appropriate parameter values for LPC analysis are chosen according to properties of the regions and according to the symbolic description of the speaker and the recording environment. The system chooses the model order, the window duration and the frame rate. It also makes statements as to the expected performance of LPC analysis within each region. A typical rule states that if a region is derived from a vowel-like phoneme and was spoken by an adult male then the number of poles to use for LPC analysis is twice the bandwidth divided by 1000 plus two [6]. Other rules state that for a region which may contain a very fast spectral transition, as during a burst, the appropriate window duration is 12 msec and the appropriate analysis rate is 167 times per second.

Following reasoning about the LPC modeling parameters the system runs the LPC analysis. The reflection coefficients are generated and residual energy recorded. The system then checks the results of the LPC analysis using various model checkers. The residual energy is checked, based on both the phonetic label and the known noise level, and poorly analyzed regions are marked. A fixed threshold is not used, but rather one which varies according to context is used. For example, good LPC analysis is expected for vowels in high to moderate SNR environments but poorer LPC analysis is expected for nasals. Other model checkers examine the formant position and tracks generated from the LPC analysis. Formant locations and smoothness are checked against expected values and discrepancies are marked. Once again, the thresholds that are used are context dependent. For example, formants are expected to be moving much faster at the onset of a vowel after a stop than after an aspirate.

After the results of LPC analysis have been examined new models are proposed for those regions which do not pass one of the previously mentioned model checkers. Currently the system builds a synthetic model using a synthesis-by-rule system. The synthesis model is an adaptation of a standard formant-based synthesizer [9]. Rules in the synthesizer specify formant tracks according to the phonetic transcript. The synthetic models and the LPC models are then combined. The synthetic models are not fixed as in classical synthesizers but are built to represent the constraints that classical synthesizers know about and then are "fitted" to the properties of the underlying speech waveform. This involves measuring those properties in the speech waveform which are not obscured by the noise and selecting the other parameters according to phonetic constraints. For example, if a synthetic vowel is to be created then the first and second formants of the vowel can often be measured even in the presence of noise. The third and higher formants may then be generated using a standard constraint propagation technique. Finally, the LPC models which are considered valid are combined with the synthetic models and with an excitation sequence to resynthesize an output sequence.

System Architecture

The overall system architecture is not the same as the flow of information show in Figure 1. The system is organized as a collection of knowledge sources which communicate using a common data base and which share knowledge in a common knowledge base. This allows for a more explicit representation of the knowledge of the system than is typically found in most signal processing systems. All symbolic

descriptions, including the original description, and all signals are contained in the common data base. These are accessible to all the knowledge sources. (In fact, Figure 1 should show the symbolic description as input to every module.) The knowledge base contains pieces of information which are needed by the various knowledge sources. For example, a hierarchy of speech descriptions (vowel, consonant, front, /i/, etc.) is stored in the knowledge base. This hierarchy records the relationships between various speech sounds. In this hierarchy the phoneme /i/ is described as both a front vowel and a high vowel and a knowledge source which contained a rule specifying how to choose the number of poles for LPC analysis for a vowel would use the speech hierarchy to find out that such a rule also applies to an instances of the phoneme /i/. The knowledge sources are invoked as they are needed. Most of the knowledge sources are straight-forward collection of rules. Special purpose structures are used freely to improve efficiency.

Examples

In this section we show some examples of the processing done by the system. The system was written in LISP and runs on a Symbolics 3600 Lisp Machine. Figure 2 shows the utterance "He has the bluest eyes" recorded by a male adult speaker, bandlimited to 5kHz and sampled at 10kHz. Figure 3 shows the time-aligned transcript and the resulting regions chosen for LPC analysis. In this case the original utterance was 2.3 seconds in duration and contained 16 phonemes and 3 silence segments. The LPC reasoning system identified 36 different regions ranging in duration from less than 1 msec to 350 msec. Figure 4 shows the model order and window durations recommended by the system as a function of position. The system recommended LPC model orders ranging from 4 poles in the silence regions up to 12 poles in the vowel regions. Window durations from 12 msec, for the glides, up to 30 msec, for the silence regions, were recommended. The average number of poles was slightly less than 10 and the average window duration was about 21 msec.

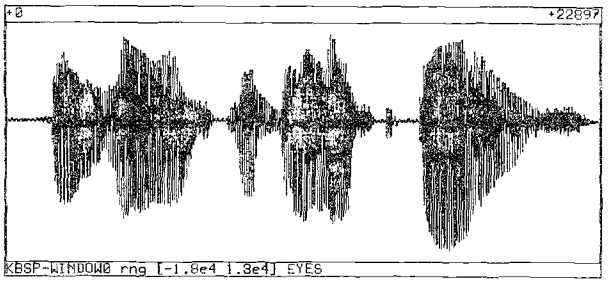

Figure 2. Original waveform.

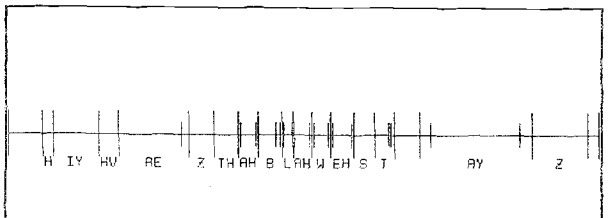

Figure 3. Time-aligned transcript and regions for LPC analysis generated from reasoning about LPC modeling. Phonemes are separated by large vertical lines. Regions are separated by the boundaries of the phonemes and by the shorter vertical lines within the phonemes.

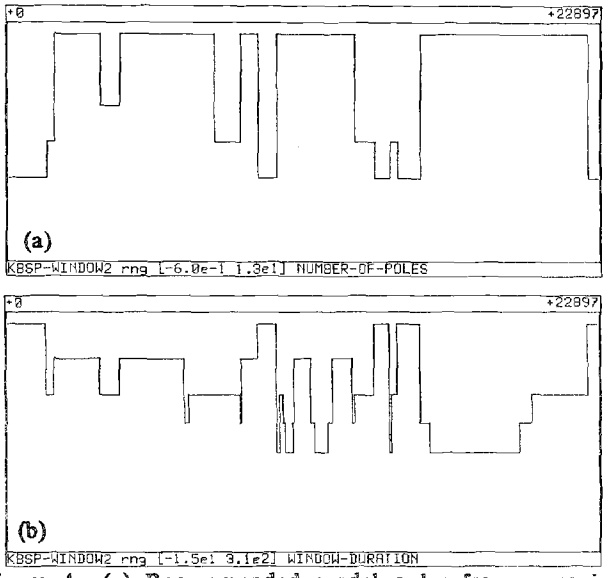

Figure 4. (a) Recommended model order from reasoning about LPC modeling. (b) Recommended window duration from reasoning about LPC modeling.

In this example the system ran 19 rules to hypothesize regions from phonetic labels and ran four rules to refine the regions. Three of the four rules hypothesized the elimination of transitional regions in vowels adjacent to aspirates. The other rule hypothesized the removal of aspiration in the stop /b/ when it preceded the voiced sound /l/. This was accomplished by instantiating the more general rule that the aspiration in a voiced stop may be removed before a voiced sound. After generation of the regions the system ran 144 rules to set the LPC modeling parameters.

Figure 5 shows the normalized residual when the recommended LPC analysis is run on a noisy version (10dB SNR) of the utterance. This curve is not substantially different from one obtained by a fixed high order LPC analysis. In this case the system has parameterized the utterance using approximately 1100 parameters per second of speech. When resynthesis from these parameters was compared with a fixed LPC analysis and resynthesis procedure, which utilized 1500 parameters per second, no difference was apparent. Similarly, little degradation in LPC resynthesis was found in using the modeling method specified by the reasoning system as compared to a fixed LPC analysis when run on noiseless speech.

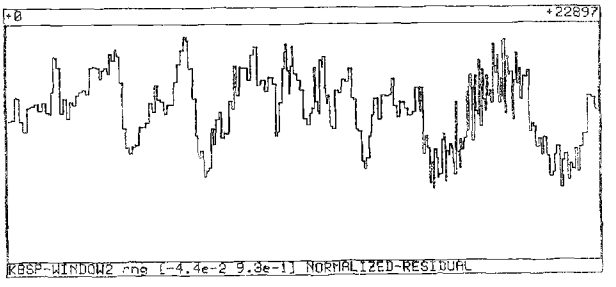

Figure 5. Normalized residual on noisy speech (10dB SNR) as computed by the LPC modeling recommended by the reasoning system.

The system also has a model checking knowledge source that understands two different forms of errors. It understands LPC modeling errors which generate too large a residual for the particular speech sound. Figure 6 shows the regions marked as failing the residual tests. These regions were derived from the normalized residual shown in Figure 5. The system does not use a fixed threshold to test the residual. Instead it employs a variable threshold which takes into account the symbolic description. For example, a lower threshold is used for stops in noisy speech because they are known to be difficult to recognize when obscured by noise. Thus, unless they are well modeled they should probably be synthesized. The model checker also understands formant positions and formant tracks. The system generates formant tracks from the results of LPC analysis. Figure 7 shows the formant tracks generated from the LPC analysis of the noisy speech. Figure 8 shows those regions which are marked as failing one of a set of tests on the formant tracks. Failures are noted due to formant locations which make no sense in terms of the phonetic label or due to formant tracks whose movement is impossible in the given phonetic environment.

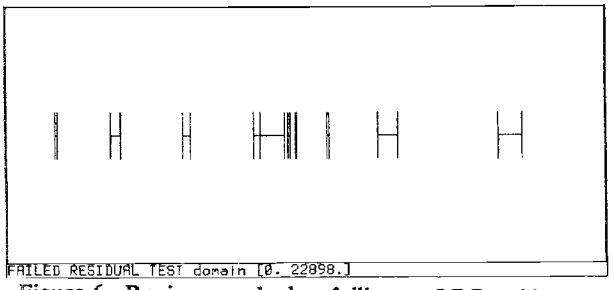

Figure 6. Regions marked as failing an LPC residual test.

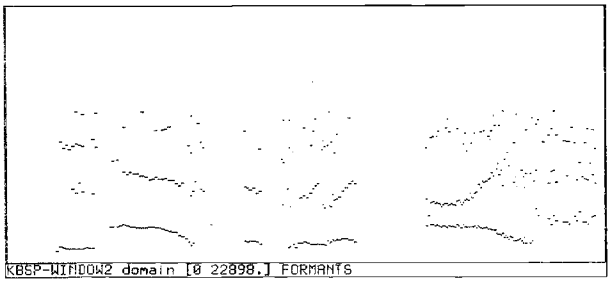

Figure 7. Formant tracks computed from the noisy speech (10dB SNR) after LPC analysis.

System Status

The model combination and reconstruction portions of the system are not yet complete but will be soon. An evaluation of the overall ability of the system will be undertaken at that time, and rule modifications and additions will be made as needed. In the long term, we would like to extend the system to understand more and different speech models. Pole-zero modeling for nasals and nasalized vowels could be incorporated. Mechanisms for choosing among more models will need to be examined. The system currently only understands noise which is stationary and white. More extensive modeling of noise sources could be used. An interactive component may be added which would allow the user to point out regions that need more work. It would also allow the user to give the system descriptions (buzzy, muffled, etc.) of the current reconstruction.

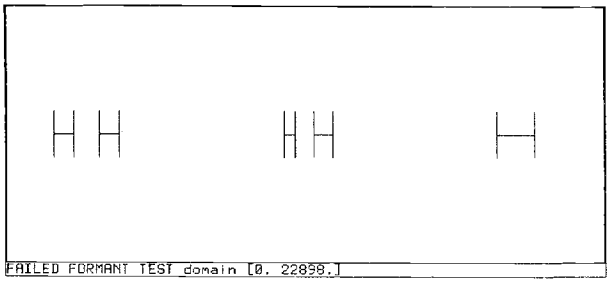

Figure 8. Regions marked as failing a formant position and track test.

Discussion

While the system is not complete, we have been encouraged in its development. The system contains an interesting mix of signal processing and symbolic processing. The combination of a rule-driven system for picking LPC analysis parameters and signal processing modules for the LPC analysis has worked well. Incorporation of new rules has been straightforward and the knowledge that is represented in these rules is explicitly available to the system. The experiments comparing the rule-based LPC analysis and resynthesis to a standard LPC analysis and resynthesis have shown the value of the symbolic information. Further development of the system, particularly the interaction between signal processing and symbolic processing in resynthesis, should substantially enhance the performance of the system and will provide an exciting area for future work.

References

[1] W. Dove, C. Myers, A. Oppenheim, R. Davis, and G. Kopec, "Knowledge Based Pitch Detection," *Proceedings International Conference on Acoustics, Speech and Signal Processing*, Boston, Massachusetts, pp. 1348-1351, April, 1983.

[2] R. Davis, G. Kopec, A. Oppenheim, "Artificial Intelligence and Signal Processing: Overview and Experiments," to be published.

[3] *Speech Analysis*, R. Schafer and J. Markel, Eds., IEEE Press, New York, 1979.

[4] *Speech Enhancement*, J. Lim, Ed., Prentice-Hall, Inc. Englewood Cliffs, New Jersey, 1983.

[5] D. H. Klatt, "Structure of a Phonological Rule Component for a Synthesis-by-Rule Program," *IEEE Transactions on Acoustic, Speech and Signal Processing*, ASSP-24, pp. 391-398, October, 1976.

[6] J. Markel and A. Gray, *Linear Prediction of Speech*, Springer-Verlag, New York, 1976.

[7] G. Peterson and H. Barney, "Control Methods Used in a Study of the Vowels," *Journal of the Acoustical Society of America*, Vol. 24, pp. 175-184, March 1952.

[8] H. Drucker, "Speech Processing in a High Ambient Noise Environment," *IEEE Transactions on Audio and Electroacoustics*, Vol. AU-16, No. 2, pp. 165-168, June 1968.

[9] D. H. Klatt, "Software for a Cascade/Parallel Formant Synthesizer," *Journal of the Acoustical Society of America*, Vol. 67, No. 3, pp. 971-995, March 1980.