S11.6

A PHONETIC VOCODER

Joseph Picone and George R. Doddington
Speech and Image Understanding Laboratory
Texas Instruments, Inc.
Dallas, TX

ABSTRACT

To achieve good quality synthetic speech at very low bit rates (on the order of 100 bits/sec), perceptually relevant bit allocations are essential. Since the information transfer in speech communications is an asynchronous process, classification of the signal into meaningful segments is a difficult problem. Linguistic and information theoretic approaches to speech communication suggest the most plausible domain in which the information transfer can be described is the phonemic domain. In this paper, we study the problem of coding spectral information in speech at bit rates in the range of 100-400 bits/sec using speaker independent phone-based recognition. Spectral information is coded as a sequence of phonetic events and a sequence of transitions through the corresponding HMM-based phone models.

INTRODUCTION

Segment vocoding approaches [1,2] to low rate speech coding typically must operate in a range from 400 to 800 bits/sec to maintain intelligible speech quality. The bit rates of these systems are restricted to these ranges because these systems allocate bits to reproduce the intricacies of the speech signal. To code speech at lower data rates, data rates on the order of 100 bits/sec, implies approaching the true information rate of the communication process [3]. Speech coding at these bit rates relies either on an extensive model of speech production [4], or detailed linguistic knowledge of the information embedded in the speech signal.

In this paper, we explore a simple coding technique based on phonetic speech recognition that is capable of producing intelligible synthetic speech at bit rates in the 100 to 200 bits/sec range. We will primarily focus on the problem of text independent speech coding in a speaker independent manner, though it will become obvious that speaker dependent and/or text dependent coding systems are simple extensions of the phonetic vocoding approach. We explore the type of speech quality achievable using this approach, and the various dimensions that influence bit rate in this type of system.

Speech communication is inherently an asynchronous process, mainly because the information flow embedded in the speech signal is a variable rate process. While synchronous coding techniques have proven to be viable at higher data rates [1], variable bit rate schemes that are synchronous with the information flow are mandated at extremely low bit rates. (Of course, sophisticated coding schemes can be added to these systems to create a fixed rate system, usually at the expense of increasing system delay.) Knowledge of the underlying phonetics of the signal allows more efficient and concise bit allocations.

Our motivation for pursuing this approach stems from our observations that high performance Hidden Markov Model (HMM) based speech recognition systems tend to do an excellent job of modeling the acoustic data. We have demonstrated a significant improvement in speaker independent recognition performance over our previous Dynamic Time Warping (DTW) systems using HMMs [7]. Major reasons for this improved performance include more accurate segmentation of the acoustic data, and better acoustic matching.

Using the LPC trace as an HMM development tool [5], we have repeatedly observed the power of our HMM systems to closely model the acoustic data, even in the presence of recognition errors. In fact, recognition errors are most often the result of the wrong symbolic assignment, rather than a poor acoustic match. The LPC synthesized speech generated from these LPC recognition traces generally sounds quite intelligible, even when using whole word models on a task such as digit recognition. We were motivated by these observations to examine the broader problem of speech coding using a phone-based speech recognizer.

A basic premise in this work is that the quality of the acoustic match will be good even if the phone recognition accuracy is poor. For instance, a recognition error for which a vowel is confused as another vowel due to an unusually high second formant frequency in the input utterance will not necessarily produce bad synthetic speech. The most obtrusive errors are those for which the broad phonetic class is incorrect. Choice of a short recognition unit, such as a phone, is critical to introducing enough degrees of freedom to insure a good acoustic match. Since the acoustic match is robust to within-phone-class recognition errors, a

CH2673-2/89/0000-0580 $1.00 © 1989 IEEE

small inventory of phones should suffice.

In this paper, we focus primarily on issues related to coding of spectral information. In previous work, we have developed efficient schemes for coding excitation information using Contour Quantization [1]. Here, we examine two basic issues, the design of the phonetic codebook, and the sensitivity of speech quality to recognition accuracy. The latter is an important issue in controlling the complexity of the system. Complex grammars can be used to increase recognition accuracy, but these also increase the complexity of the speech recognition subsystem.

A powerful tool for studying phone based speech recognition is the TIMIT Acoustic Phonetic database [6]. Here, we will consider a subset of the TIMIT database consisting of a total of 2792 sentences from 245 male speakers and 104 female speakers. The TIMIT database represents a comprehensive coverage of spoken American English (including dialect). This database has been phonetically transcribed and labelled using a set of 60 phones [6]. This transcription data forms the basis for our work. Let us begin by examining some fundamental coding limits in the phonetic vocoding approach.

FUNDAMENTAL CODING LIMITS

The average speaking rate for the TIMIT database is 2.7 words/sec. Note that this includes a short stretch of silence at the beginning and end of file. The average phone rate, computed from the transcription data, is 12.3 phones per second. A histogram of the phones is shown in Figure 1. The entropy of this distribution is 5.5 bits. Thus, assuming perfect speech recognition performance, approximately 68 bits per second are required to transmit phone sequences.

We also need, however, to transmit duration information for each phone model. We have considered two HMM models for phones, as shown in Figure 2. The first model, the simple progressive model, implies that each incoming frame of speech data requires 1 bit/frame to code whether a stay or progress transition occurred. At a nominal frame duration of 20 msec, this information requires approximately 50 bits/sec to transmit. The second model requires, in the limit, requires 1.6 bits/frame, or a total of 80 bits/sec.

The net bit rate required, then, is on the order of 120 bits/sec to code the spectral information. Of course, the above analysis essentially assumes one HMM model per phone, and does not account for the required synchronization bits. Each factor of two increase in the size of the phone codebook, in the worst case, adds 1 bit per phone, or 12 bits/sec, to the net bit rate.

For additional redundancy removal, we can examine the histograms of phone pairs and phone triples. Of the 3600 possible phone pairs, slightly over 75% occur. The entropy of this distribution is 9.6 bits. Similarly, out of the 216,000 possible phone triples, only approximately 20,000 occur in the training database. The phone triple distribution has an entropy of 12.8 bits. Thus, incorporating some higher level constraints on sequences of phones can reduce the bit rate required to transmit phone indices by approximately 13% in the case of phone pairs, and 22% in the case of phone triples. Below, we will examine the virtue of using a phone pair grammar to aid in speech recognition subsystem.

THE PHONETIC VOCODER

The speech recognition subsystem, described in more detail in [7,8], is a layered grammar approach to speech recognition, and is shown in Figure 3. The HMM technology is a standard continuous density system. The acoustic front-end uses an LPC based feature extraction system in which 10th order LPC coefficients, pitch, and energy, computed every 20 msec. The acoustic front-end uses several auxiliary measures of energy and spectral change in addition to the core LPC derived principal components based features. A statistical clustering technique described in [8] is used to generate HMM seed models.

Phone seed models were constructed from the TIMIT database, using the phonetic transcriptions and labels provided. A contextually rich sample of excised exemplars of each of the 60 phones were clustered into codebooks of 2, 4, and 8 models per phone. These clustered HMM models were then re-estimated using supervised training at the phone level on the entire training database. In this mode of supervision, the recognizer is forced to recognize the sequence of phones contained in the phonetic transcription. Timing information from the manual transcriptions is discarded.

The basic phone recognition system with a null grammar (any phone can follow any phone) requires on the average about 60 MIPS/sec (executed on a VAX 8650). It is interesting to note that this is only two to three time greater than the processing time required by our previous VQ based 400 bits/sec systems, though the nature of the computations are significantly different. The former system relies largely on symbolic processing and data structure manipulation (mips), while the latter system requires mainly vector operations (megaflops).

CODING PATHS THROUGH HMM MODELS

The major way in which the phonetic vocoder distinguishes itself from a vector quantization system is the manner in which the spectral information is transmitted. Rather than transmit indices in a VQ codebook representing each spectral vector,

we transmit a phone index, and auxiliary information describing the path through the model. We have used simple left to right progressive models because the transitional behavior can be described efficiently by 1 bit per state. Other alternatives that are suggested include vector quantizing the paths through each model such that only a subset of all possible paths are allowed.

Naturally, the simpler the model the more likely it is that every path in the model will be exercised. We find that forcing the system to use simpler models, and coding the transitions using 1 bit/frame to be an acceptable tradeoff between complexity and performance. The use of more complicated models, such as the skip state model in Figure 1, does not significantly improve quality, and requires more effort to efficiently encode path information.

EVALUATION OF PERFORMANCE

Initial experiments focused on studying the number of phones required in the codebook. Three codebooks were generated: 2 clusters per phone, 4 clusters per phone, and 8 clusters per phone. These codebooks require 7, 8, and 9 bits, respectively. A null grammar was used.

Not surprisingly, overall speech quality was found to be the best with an 8 cluster codebook. The change is speech quality is most noticeable when the codebool size is increased from 2 to 4 clusters. Speech quality with 2 clusters is significantly buzzy. As is typical with most low rate coding schemes operating at their lower bounds, the regeneration of high frequency information is poor, giving the overall spectrum a low pass quality. This results in synthetic speech that sounds thin, and very nasalized. Increasing the size of the codebook makes the speech sound richer, and improves high frequency reproduction.

The phone recognition accuracy for these three codebooks is approximately the same. Because of the absence of any real grammar constraints, phone recognition performance is generally poor, typically about 35% of the phones are correct. Approximately 50% of the errors are substitution errors, while the other 50% are insertions and deletions. The insertion error rate drops about 20% going from the two cluster codebook to the eight cluster codebook.

In the absence of any real grammar constraints, HMM systems that model every frame of data have a tendency to insert hypotheses. This has the tendency to artificially raise the bit rate. In fact, the codebook of 8 clusters per phone hypothesizes, on the average, 12% more phones per second than the lower bound.

The phone pair grammar described above was then used to impose some simple phonotactic constraints, in hope of reducing the phone recognition error rate. The average number of phones hypothesized in the 8 cluster case remained approximately the same, and the recognition accuracy did not change significantly. No improvements in speech quality were observed. The phone pair grammar also required 25 times more CPU time to process than the simple null grammar. Thus, this approach was not pursued.

The codebooks generated for these experiments were actually structured by sex. That is, in the 8 cluster codebook, 4 clusters were allocated to males, and four clusters to females. By strictly imposing the constraint that all recognition must occur with either all male or all female models, we can reduce the number of bits required for the codebook by one.

Thus, the basic phonetic coding system can operate with an 8 cluster codebook, and results in an average bit rate of 170 bits/sec. A spectrogram of the synthetic speech is compared to the original in Figure 4. The speech quality compares very favorably with previous VQ-based coders that have used 300 bits/sec to code spectral information. This 170 bits/sec bit rate is also very close to the lower bound.

CONCLUSIONS

A simple phonetic speech coding system has been shown to be a promising approach to low rate speech coding. A simple inventory of phonemes was shown to be sufficient for capturing the bulk of the acoustic information. There are three major issues that need to be explored at this point.

First, the bit rate can be reduced by improving recognition performance, and encoding larger units of speech. No attempts have been made to encode silence efficiently, or to delete some phones that are insignificant perceptually. We are currently developing more comprehensive grammar structures and recognition systems that efficiently process natural language grammar structures. Maintaining a small grammar complexity, however, is a desirable way to control complexity.

Second, and most importantly, segmental encoding of the excitation using similar acoustic-phonetic approaches is worth investigating. We have learned from previous work [1] that it is important to encode excitation on a linguistically meaningful segmental basis to achieve low bit rates.

Finally, it is obvious that speech quality can be improved by adding several larger units to the the recognition system. Function words and other similar morphs are good candidates to be added to the system with a minimal increase in complexity.

REFERENCES

[1] J. Picone and G.R. Doddington, "Low Rate Speech Coding Using Contour Quantization," Proc. 1987 IEEE Int. Conf. Acoust., Speech, and Signal Proc., pp.1653-1656, April 1987.

[2] R.M. Schwartz and S.E. Roucos, "A Comparison Of Methods For 300-400 B/S Vocoders," Proc. 1983 IEEE Int. Conf. Acoust., Speech, and Signal Proc., pp.65-68, April 1983.

[3] L.R. Rabiner and R.W. Schafer, Digital Processing of Speech Signals, Prentice Hall, New Jersey, 1978.

[4] J. Schroeter, J.N. Larar, and M.M. Sondhi, "Multi-Frame Approach for Parameter Estimation of a Physiological Model of Speech Production," Proc. 1988 IEEE Int. Conf. Acoust., Speech, and Signal Proc., pp.83-86, April 1988.

[5] G.R. Doddington, J. Picone, and J.J. Godfrey, "The LPC trace as an HMM development tool," Journal Acoust. Soc. Am., Suppl. 1, Vol. 84, Fall 1988.

[6] W.M. Fisher, G.R. Doddington, and K.M. Goudie-Marshall, "The DARPA Speech Recognition Research Data base: Specifications and Status," in Proceedings of DARPA Speech Recognition Workshop, pp. 93-99, February 1986.

[7] J. Picone, G.R. Doddington, and J.J. Godfrey, "A Layered Grammar Approach to Speaker Independent Speech Recognition", presented at the 1988 IEEE Speech Recognition Workshop at Harriman, NY, June 1988.

[8] J. Picone, "On Modeling Duration In Context In Speech Recognition", to be published in the Proc. of the 1989 Int. Conf. on Acoust., Speech, and Signal Proc. in Glasgow, Scotland.

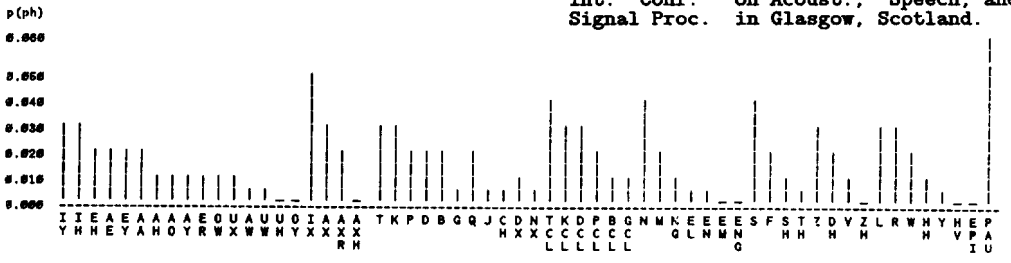

Figure 1. A Phone Histogram for a Subset of the Acoustic Phonetic Database

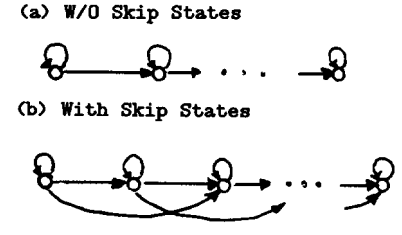

Figure 2. HMM Phone Models Used In The Phonetic Vocoder

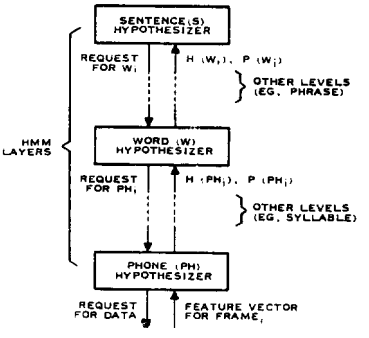

Figure 3. A Layered Grammar Recognizer

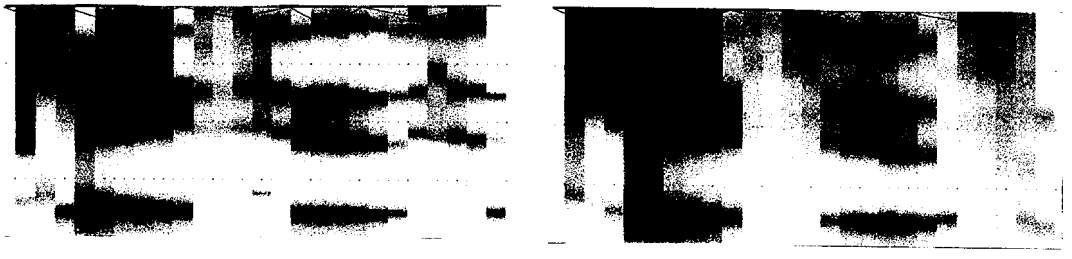

Figure 4. Spectrograms Comparing Original Speech (Left) to Synthetic (Right)