INTELLIGIBILITY OF SPEECH COMPRESSED USING AN AUDITORY MODEL

Dale Carnegie*†, Geoff Holmes*, and Lloyd Smith*

*Department of Computer Science
University of Waikato

†Department of Physics
University of Waikato

ABSTRACT

We extract dominant frequencies from speech waveforms by an In-Synchrony-Bands spectrum analyzer based upon an auditory model. Experiments indicate that intelligible reconstructed speech requires only 3 such frequencies per frame. This paper presents the results of our investigation into speech compression employing this technique.

INTRODUCTION

Ghitza (1987) describes an auditory model that attempts to model speech at the auditory nerve level by capturing the phenomenon of auditory synchrony. This model exhibits good noise immunity, and provides highly intelligible results using a substantially reduced number of frequencies. Carnegie et al. (1990) experimented with this model, and examined the effect of varying the bandpass filter slopes, the setting of threshold levels to determine what would be considered as dominant frequencies, and also where the model appeared to perform poorly .

Investigations into speech reconstructed with a reduced number of frequency components indicate that highly intelligible results require only three frequencies. We verified this using 4 groups each of 10 untrained listeners, under a Diagnostic Rhyme Test format. Ideally there should be no correspondence between the form of speech processed and the ability of the model to adequately reproduce it. Results indicate that this is not the case, with the model performing especially poorly in the areas of voicing and nasality. Consequently we have modified the original model to compensate for this, and are endeavouring to produce a low-bit coding system comparable to those detailed in the literature. This paper describes our experiments in compression and intelligibility, and details the modifications necessary to produce intelligible speech at low bit rates.

IMPLEMENTATION

Speech input to the model is sampled at 8kHz and bandpass filtered at 300-3300 Hz. The speech is segmented into 128 point frames, Hamming windowed and converted to the frequency domain every 8 milliseconds by an FFT algorithm. To model the auditory nerve fibres we use 100 highly overlapping bandpass filters, equally spaced on a log scale with a 3% frequency step. The filter's slopes correspond to the tuning curves of the auditory nerve fibres in cats (Ghitza 87). Filters below 1 kHz have an incline and a decline of 18dB/octave, whilst filters whose centre frequency is greater than 1kHz have a declination of -120dB/octave (refer Figure 1).

SBS processing involves keeping track of the number of adjacent filters that have the same dominant frequency, and the SBS spectrum is the number of such adjacent filters (refer Fig 1). For example, an SBS amplitude of 5, indicates that that frequency dominates 5 adjacent filters. From past experiments (Carnegie et al. 1990), we knew that a threshold SBS amplitude of 10 would provide us with highly intelligible speech with an average of three contributing frequencies. Reconstruction with two frequencies is almost incomprehensible, and reconstructing with 4 frequencies does not significantly improve upon the performance we obtain when using 3. Hence further experiments would only use those frequencies whose SBS amplitude was greater than or equal to 10 for the speech reconstruction. In these initial investigations, we did not alter the original amplitudes and phases, however, in some circumstances, this resulted in an unacceptably low level of reconstructed intelligibility.

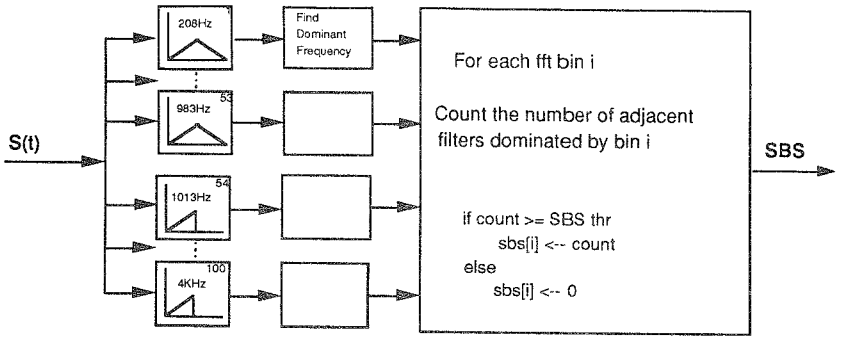

Figure 1. The Auditory Model and SBS Determination

To test the original model, initial experiments concentrated upon examining the model's performance in terms of noise immunity, formant tracking, and the introduction of tonal artefact for certain sounds (Carnegie et al. 1990). Our results show an immediate possible application in the area of speech compression. After establishing a definite number of frequency components for use in the speech reconstruction, we next varied the amplitudes and phases of the original speech waveform. Our experiments indicate that amplitude is more important than phase for intelligibility purposes, so all phase information was eliminated from the reconstructed speech. We were then able to examine the effects of varying the quantisation of the amplitude.

We chose the Diagnostic Rhyme Test (DRT) as the format for the experiments (Voiers 1977, 1983). Our implementation of this test provides an intelligibility score for words tested in isolation, with each word representative of voicing, nasality, sustention, sibilation, graveness or compactness. Results provide an indication of the model's performance in each area, for each level of quantisation. Four groups each comprising 10 untrained listeners undertook two tests. There were a total of four different tests. The first three tests differed only in the quantisation of the amplitude, 3 bits, 5 bits, and 32 bits (effectively no quantisation). The final test was a control and involved unprocessed words being played to the listeners. Each DRT comprised 96 word pairs, and the tests required the listener to choose one word from a pair that most closely resembled the SBS processed word.

The processed word pairs were recorded by a single male speaker and generated 1.4 seconds apart. The order of the words appearing in the test exhibit no correspondence to which speech grouping it belongs (sibilation etc.). Our aim is to produce a highly intelligible, low bit rate system whose use extends beyond the laboratory environment. With this in mind, we avoided the use of headsets during the DRT, and instead employed a commercially available loudspeaker, playing to listeners sitting at varying distances in a large room.

RESULTS

Somewhat surprisingly, the unprocessed speech used as a control in the experiment only results in a 90% intelligibility score. This deviation from an ideal score is most likely due to the nature of our test facilities, and also possibly due to speaker bias. Despite this minor problem, a statistical analysis of the DRT does provide us with information concerning the model's performance given varying levels of amplitude quantisation on the 6 listed forms of speech.

Literature studies (Pinto et. al. 1989, Rothauser et. al. 1969) tend to indicate that a minimum requirement for speech to be intelligible is for at least 75% of the listeners to correctly identify the word. With the exception of a few words indicative of sibilation, none of the words processed with 3 bit amplitude quantisation reached this level. At five bit amplitude quantisation, the intelligibility scores improve markedly. Although the scores for no amplitude quantisation are higher, the improvement is not dramatic. The test does however, indicate a markedly poorer result for those words designed to test voicing and nasality.

The time consuming nature of such an extensive DRT test makes it impractical as an evaluation method for each parameter change. Instead, we asked a trained listener to attempt to detect any difference between different recordings of the same word. As an example, the listener was asked to compare the intelligibility of words produced with 5 bit amplitude quantisation with words reproduced with 6 bits. We knew from the DRT that there is a slight degradation in intelligibility going from full amplitude to 5 bit quantisation. It is reasonable to assume that the loss going from 6 bits to 5 bits would be more significant than the loss in going from 7 to 6, or 8 to 7 etc. We also knew that the overall loss is not great and could perhaps be considered as not particularly significant in terms of intelligibility. Our listener had trouble reliably and consistently detecting a difference between words generated with 5 bit quantisation and words generated with 6 bits. This, combined with the results from the DRT indicate that the loss in intelligibility is comparatively small, and hence future experiments could be based on 3 frequencies with an "upper limit" of 5 bits phase and amplitude quantisation.

The model performs poorly over most of the 6 DRT categories, and needs to be improved considerably before it is of any real use. To determine a starting point for improvements we noted the degradation in intelligibility of each of the word types with increasing quantisation. Voicing is of particular interest. Not only did the voicing performance degrade rapidly from the unprocessed to the amplitude-only test, but the degradation took it well below the 75% acceptance level. This we assumed was a result of endeavouring to reproduce unvoiced sounds that normally exhibit "white noise" spectra, with only three frequencies. To test this, we recorded a selection of 17 phones, chosen from a list of fricatives, stops and affricatives that would enable us to explicitly test the model's reproduction of unvoiced speech. The results indicate that our assumption was incorrect.

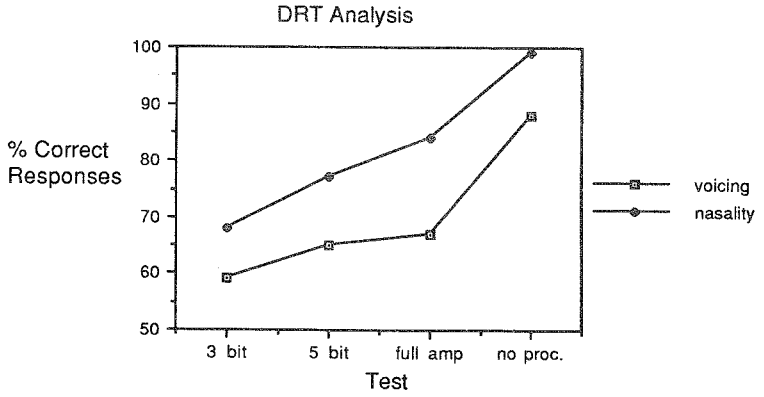

Figure 2. Quantisation of Voicing and Nasality

Rather than the problem being with the unvoiced speech segments, it became clear that the amplitude of the frame segment in the original data is far more important. For example, the word "SHOE" reproduces very well with very little tonal artefact. The original amplitude of the /i/ is quite high, and when reproduced the contributing frequencies contain sufficient energy to render that frame easily audible. "Valve" on the other hand, reproduces poorly although the /v/ sounds are voiced. Upon inspection of the original waveform, it is clear that the initial amplitudes are quite low. Hence upon reconstruction, so much energy is lost from an already quiet frame that that frame becomes quite inaudible. As such quiet frames are common in words indicative of nasality and voicing, this problem could account for their poor DRT performance.

To further examine this loss of energy, we amplified those frames we knew to be poorly reproduced by applying an arbitrary gain of 10 to those frames in the original waveform. This gain factor is excessively high, and results in some distortion as well as producing a very unnatural sound. However, there is a marked increase in intelligibility in most cases. A method of manual application of gain is clearly not an

acceptable solution for implementing in a computational model. One possibility is to set the energy of the initial and final frames to be equal, regardless of the number of contributing frequency bins. One implementation of this is to force the rms energy for each frame to remain constant.

To reproduce the same rms energy, given fewer contributing elements, the following expression was derived :

$$N \times rms^2 = \left(n_1^2 + n_2^2 + \ldots + n_{k-1}^2 + 1 \right) x_k^2$$

where N is the number of frequency bins in the original speech frame, rms is the original root mean square energy of the speech frame, $x_1 = n_1 x_k, \ldots, x_{k-1} = n_{k-1} x_k$, x_k is the last contributing amplitude and k is the number of frequency bins used in the reconstruction. For our experiments k has an average value of 3.

Applying this energy gain term does result in an increase in intelligibility from the unmodified model, but in many respects is inferior to the results obtained from the manual application of the huge gain. Continuing experiments are attempting to derive some optimum automatic gain value for these frames.

CONCLUSIONS

Speech compressed via SBS processing, and then subjected to quantisation of the amplitude components remains highly intelligible. Speech transmission, assuming no quantisation of pitch and a 50% overlap of succeeding frames, is at 125*((3*5) + (3*6)) = 4kbits/second. Comparable compression schemes have a transmission rate of 1.2kbits/sec (O'Shaughnessy 1987). Preliminary experiments indicate the frame overlap is not necessary for intelligibility, and that by non-linear coding of the frequency bins we may be able to halve the 6 bits that represent the unquantised frequency. Our transmission rate then approaches 1500 bits/second.

One serious problem is with the gain factor that amplifies those frames that exhibit considerable energy loss upon reconstruction. If we were to transmit the rms energy of the original frame with the frequency and amplitude data, the number of bits we would need to transmit would increase considerably. This is not an acceptable option for compression. It should not however, be necessary to resort to this. Instead we require a software test to recognise a frame with low energy and amplify it accordingly. So, even though our experiments are concentrating on comparing initial and reconstructed energies, what is envisioned for application purposes is to apply some gain factor that need not necessarily form part of the data transmission. We estimate that the most appropriate way to do this will be to actually track the energies of the reproduced frames. Work is continuing in this area.

Finally, noting that we have effectively been endeavouring to track formants, we may be able to estimate, with a reasonable degree of accuracy, the dominant frequencies of a succeeding frame. Another DRT is required to evaluate the effects of this. We eventually expect this model to be able to intelligibly reproduce speech at below 1000 bits per second transmission rate. Techniques such as Vector Quantisation will then be explored to lower this even further.

REFERENCES

Carnegie, D., Holmes, G. and Smith, L. (1990) "Implementation of an Auditory Model", Proc. Australian International Conference on Speech Science and Technology , 214-217.

Ghitza, O. (1987) "Auditory Nerve Representation Criteria for Speech Analysis/Synthesis" IEEE Transactions on Acoustics, Speech and Signal Processing, ASSP-35, 736-740.

O'Shaughnessy, D. (1987) *Speech Communication - Human and Machine.* p. 331 (Addison-Wesley).

Pinto, N.B., Childers, D.G. and Lalwani, A.L. (1989) "Formant Speech Synthesis: Improving Production Quality" IEEE Transactions on Acoustics, Speech, and Signal Processing, 37, 1870-1886.

Rothauser, E.H. et al. (1969) "IEEE Recommended Practice for Speech Quality Measurements" IEEE Transactions on Audio and Electroacoustics, AU-17, 225-246.

Voiers, W.D. (1977) "Diagnostic Evaluation of Speech Intelligibility", In Speech Intelligibility and Speaker Recognition: Benchmark Papers in Acoustics II, edited by M.E. Hawley p. 374-387, (Dowden, Hutchinson, and Stroudsburg, PA) .

Voiers, W.D. (1983) "Evaluating Processed Speech Using the Diagnostic Rhyme Test", Speech Tech. 1, 30-39.

A METHOD TO EVALUATE THE PRE-PROCESSING STAGE OF ISOLATED-WORD RECOGNITION SYSTEMS

Hiroaki Oasa and Michael Wagner
Department of Computer Science
University College
University of New South Wales

ABSTRACT - This paper describes a cluster analysis method developed to evaluate pre-processors for isolated-word recognition. The method derives various statistical measures on inter-word and intra-word distances, whose ratios are analysed to evaluate the relative effectiveness of the pre-processors.

The study uses a speech database comprising (a) the 36-word set of alphabetic letters and digits and (b) a phonetically balanced set of 36 CVC words, and evaluates three pre-processors which extract FFT log-power spectrum coefficients, linear-prediction based cepstral coefficients, and critical band energies from the speech signal.

INTRODUCTION

Evaluating an automatic speech recognition system by stating the overall recognition rate of the system for certain vocabularies, speaker modes and syntactic support is often unsatisfactory. This stems from the fact that the overall recognition rate is a highly non-linear function of the distribution of the test and reference tokens in the chosen parameter space. As such the recognition rate often provides little enlightenment as to why a system performs as well or as badly as it does.

A better approach to evaluate automatic speech recognition systems is the separate examination of the major modules of the system, for example the parameter or feature extraction module, the pattern recognition module or the syntax, semantics or pragmatics modules. For each module to be thus examined it is necessary to derive suitable measures against which the module can be evaluated in order (1) to provide developers with more useful information on the strengths and weaknesses of the system and (2) to give potential users of a speech recognition system a robust and detailed yardstick to compare between systems.

This study presents an evaluation of the pre-processing stage of an isolated word recognition system. In any pattern recognition system it is important that a suitable set of parameters or features is extracted from the signal in order to provide a good separation of the classes of tokens which are present in the signal and which are to be distinguished by the recognition system. The study therefore proposes a number of statistical measures which relate to the clustering power of the pre-processing module, that is the ability of the system to cluster tokens of the same class closely together while separating clusters of tokens from one class as widely as possible from clusters of tokens of another class.

As the automatic speech recognition system in question is a speaker-dependent isolated-word recognition system, tokens are isolated words from a given vocabulary which are spoken by a given speaker. There are as many classes of tokens, and therefore clusters of tokens, as they are words in the vocabulary and each cluster contains as many tokens as repetitions of the word were recorded by the speaker.

This study introduces a number of statistical measures which are closely related to the well-known F-ratio, that is the ratio of inter-cluster over intra-cluster variance. These statistical measures are then used to evaluate three different pre-processors, namely, the log power spectrum, the LPC cepstrum and the critical band spectrum.