COMPARISON OF METHODS FOR SPEECH ANALYSIS

George Raicevich and Phillip Dermody.

Speech Communication Research Group,
National Acoustic Laboratories, Sydney Australia

ABSTRACT - Comparisons of performance are made between an auditory model and LPC analysis. Further more two types of auditory model outputs, mean rate and synchrony response are tested for the lowest distance metric error. A time sensitive Euclidean distance measure (Integrated Time Squared Error : ITSE) is used and compared to a Euclidean distance metric. A local speech data base of CV combinations mixed with office environment noise is used for the testing.

INTRODUCTION

Speech recognition front end processing with an auditory model (AM) may provide improved performance in maintaining speech information in noisy environments. The performance of the Seneff AM [1] mean rate response and synchrony response were compared. The mean rate response is a spectral representation of the cochlea output useful in locating acoustic events and assigning segments to broad phonetic categories. A question raised is whether the spectral sharpening provided by the synchrony response gives any performance advantage over the mean rate for speech signals in high levels of noise. As a comparative reference for the AM, LPC analysis distance metric results are presented.

Only a brief description of the Seneff model is presented while a thorough explanation can be found in [2]. The model is shown in figure 1. It consists of stage 1, a set of 40 critical band spaced bandpass filters each feeding stage 2, a half wave rectifier followed by a short term adaptation circuit, lowpass filtering and rapid automatic gain control (AGC). Stage 2 models the auditory model synapse neuro transmitter release, nerve fibre synchrony reduction and nerve fibre refractory effect. Stage 3 provides the mean rate and synchrony response functions [2,p58].

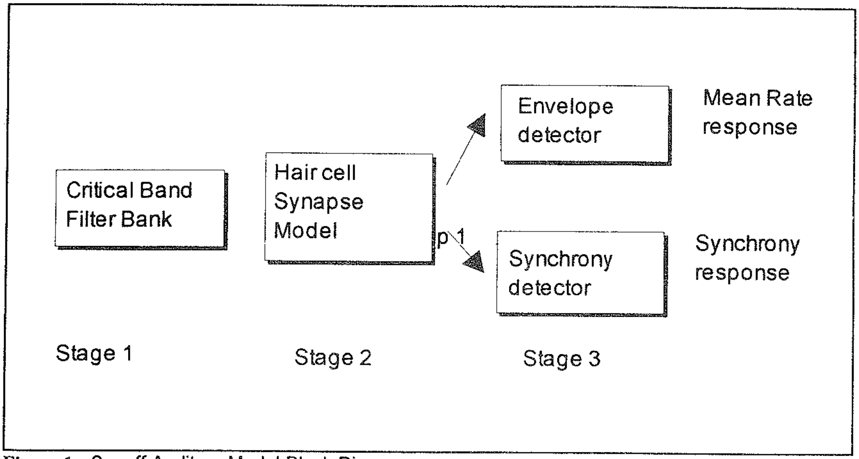

Figure 1. Seneff Auditory Model Block Diagram

The Seneff AM synchrony response yields a spectral representation with enhanced spectral contrast relative to the AM mean rate response. Signals with a periodic nature are enhanced by the synchrony stage creating distinct formant peaks during sonorant regions of speech. Both synchrony and mean rate response were compared to a 14th order LPC analysis with +6dB/oct high frequency pre emphasis.

The signal and signal plus noise differences where measured with a modified euclidean distance measure sensitive to time contiguous errors called the Integrated Time and Squared Error (ITSE) measure. The ITSE resets the integrating function whenever the error ceases to be time contiguous. The ITSE is an extension of the Euclidean distance metric [4,5]. The rationale for using the ITSE is that more errors are made in human speech recognition tasks when a critical acoustic cue is masked by time contiguous noise than for a momentarily masked acoustic cue. The ITSE may emulate this function and so provide a diagnostic tool that more closely resembles errors made in human speech recognition. The ITSE has been previously reported [3].

$$ITSE = \left[\frac{1}{f_{\max} \cdot t_{\max}} \cdot \sum_{f=1}^{f_{\max}} \sum_{t=1}^{t_{\max}} \left(etl \cdot \left| \mathrm{F}_s(f,t) - \mathrm{F}_m(f,t) \right|^2 \right) \right]^{0.5} \qquad (1)\ \text{eq.}$$

Key

t	=	*Time*
f	=	*Frequency*
etl	=	*Err. time length*
$\mathrm{F}_s(f, t)$	=	*Signal*
$\mathrm{F}_m(f, t)$	=	*Masked signal*

METHOD

Database

The database comprised 12 Consonant Vowel (CV) combinations for six phonological categories. These included both voiced and unvoiced stops, fricatives, nasal, lateral and semivowel. The CV combinations were used as clean reference. Masked speech samples were created by mixing the reference speech with a corpus of speech like noise recorded from bank office environments. The CV combinations were mixed with the office noise samples at signal to noise ratios of 0, 5, 10, 20dB.

Normalisation.

Comparison between different AM's is complex because each model produces a different time, frequency and amplitude scale. It is necessary to normalise the three dimensions to allow performance comparison between models and different analysis methods.

The output of different analysis methods usually differ in aspects of the number of spectral channels and frequency range. A common frequency response range and channel number was achieved using spline interpolation on the spectral output of all the models. Each were normalised to a common range of 210 to 6500hz with 40 channels. The frequency limits were imposed by the range common to all the models.

Time normalisation of the clean and noisy signal was observed by running models with as similar as possible time step lengths.

Each model output was normalised for amplitude dynamic range. In each case the log output, referenced to the peak, was used. The method of normalisation for these results was to reference 0dB as the peak amplitude of the acoustic sample tested and to take all amplitudes exceeding 50% of the dynamic range.

Tests were conducted to measure the robustness of the distance metric results for a wide range of dynamic range fraction.

RESULTS

The results for the normalisation of output and the noise comparison are presented.

Normalisation :

Amplitude.
To normalise for amplitude, different portions of the dynamic range were taken and plotted against DM for the model outputs to test for robustness of method. Figure 2 illustrates that the results remain in the same relationship while measuring the DM over different portions of the dynamic range. All measurements here after were made with energy greater than 0.5 of the dynamic range.

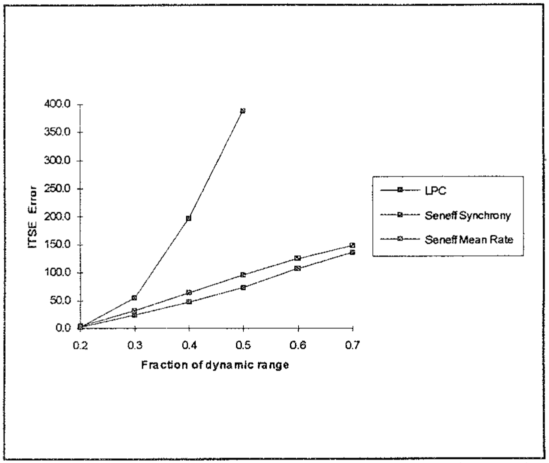

Figure 2 . ITSE Results with amplitude normalisation by the use of different dynamic range portions showing that the results are consistent.

For the case when 0.5 of the dynamic is used, figure 3 shows the auditory model performance leads at 0dB signal to noise ratio. It can be seen that for the DM there is little resolution between the mean rate and synchrony response. Also seen is the same relationship between the results with a greater differentiation as the signal to noise ratio approaches 0dB.

Time step length.
The time window length was set to 2.5mS +/- 0.05mS for each analysis and model output with a 50% window overlap. The overall signal sample length was the same for the clean and noisy sample and the same samples were used for all models. It has been found in previous experiments [6] that for a fixed sample time length, changing the resolution of the model output time window size does not significantly change the DM error. For example, increasing the number of time records from changing from 50 to 100 for the AM output produced a 1% change in the euclidean DM result. The DM and ITSE results are quite robust with respect to large changes in time step length.

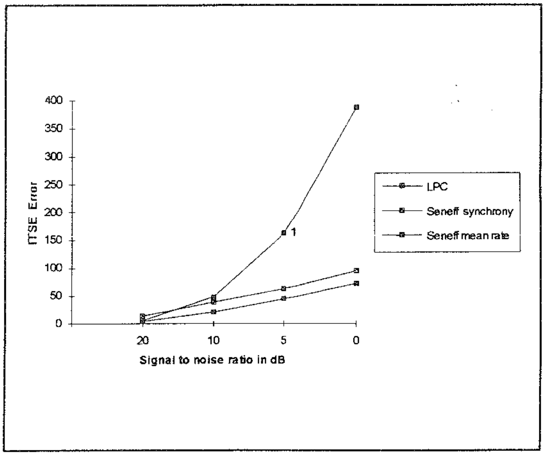

Figure 3. ITSE results for an amplitude normalisation set to 0.5 of the dynamic range.

Table 1 Illustrates the results of the Euclidean distance and the ITSE. The ITSE gives greater difference in error at higher noise levels for the model and analysis. The results are consistent to the Euclidean metric. The results for ITSE criteria show an improved performance in noise over the DM measure.

Table I. Results for the comparison of DM and ITSE using two AM type output and LPC.

Euclidean DM.	Signal to noise ratio in dB			
	20	10	5	0
LPC analysis	0.8	3.1	5.5	8.8
Seneff Mean Rate response	0.5	1.7	2.7	3.8
Seneff Synchrony response	1.5	2.9	3.7	4.5

ITSE DM.	Signal to noise ratio in dB			
	20	10	5	0
LPC analysis	5.8	48.5	163.0	387.6
Seneff Mean Rate response	4.1	21.8	44.8	73.6
Seneff Synchrony response	14.0	39.0	63.1	96

EVALUATION OF SIGNAL ANALYSIS IN NOISE

Performance of the AM and LPC compared quite favourably at 20dB s/n but at 0dB there was a performance advantage for the auditory model as can be seen in figure 3. Both the mean rate and synchrony response of the AM displayed a lower error score for both distance metrics when compared to the LPC analysis. This advantage is the greatest at signal to noise levels around 0dB S/N in speech like noise. It can also be seen that the performance of LPC deteriorated markedly in comparison to the auditory model results. The mean rate response gave a lower DM and ITSE error compared to the synchrony response.

DISCUSSION

Normalisation of model outputs.

Normalisation of frequency response and time step length were carried out successfully. Each model has a unique amplitude dynamic range dependent on implementation and this must also be normalised when comparing different models. The rationale for using the top half of the signal dynamic range as a form of amplitude normalisation was that in difficult acoustic environments it is perceived that human listeners do likewise. Cues are sought in speech signal peaks. The present method of amplitude normalisation requires improvement since it is prone to error from "statistical outliers" i.e., referencing the 0dB to a "spike" in the noise or signal (the data was examined for such an occurrence).

Comparison of DM and ITSE.

Both the Euclidean and ITSE gave comparable results, with the ITSE giving greater DM error differences in high noise conditions. The theoretical advantage of the ITSE is that it provides greater sensitivity to time contiguous noise type errors.

Comparison of alternate auditory model outputs.
It was interesting to note that the mean rate response performed slightly better than the synchrony detector at high (0dB) levels of noise. This may be attributed a data base primarily composed of unvoiced components. Since the synchrony response enhances periodic sounds it will have a minor influence on these unvoiced sounds.

On inspection of the noise sample used it was seen that a portion was a periodic-like background voice. It was reasoned the synchrony spectral sharpening tended to reduce the volume under the curve of non periodic like sounds (noise) and hence the DM and ITSE. The synchrony detector in effect enhanced the speech type noise and increased the error scores. Note also that the computational cost of a mean rate is less than for the synchrony response. The particular speech like noise used in the test represents a worst case condition. These results may not have been recorded if white noise had been used for the tests.

ACKNOWLEDGMENT

This work was carried as part of the GLASS project. We gratefully acknowledge the financial support of the Department of Industry, Technology and Commerce. We also thank Dr S. Seneff for assistance with her auditory model.

REFERENCES

[1] S. Seneff (1985) Pitch and Spectral Estimation of Speech Based on an Auditory Synchrony Model. RLE Technical Report, No 504, Massachusetts Institute of Technology.

[2] S. Seneff (1988) A Joint Synchrony/Mean-rate Model of Auditory Speech Processing. Journal of Phonetics, vol. 16 (Academic Press) 55-76

[3] Dermody P. , Raicevich G. , Katsch R., (1992) Comparative Evaluations of Auditory Representations of Speech. ESCA Book chapter 1 . (John Wiley & Son)

[4] Shinners S.M. (1979) Modern Control System Theory and Application 2and Edition (Addison Wesley Publishing.) 180

[5] Shultz W.C. Rideout V.C. (1961) Control System Performance Measures: Past, Present and future. IRE transactions on automatic control. Feb. 22-35

[6] Raicevich G. (1991) Distance Metric Evaluations (unpublished report) N.A.L.

AUDITORY MODELS AS FRONT-ENDS FOR SPEECH RECOGNITION IN HIGH NOISE ENVIRONMENTS

M.D. Chau and C. D. Summerfield

Syrinx Speech Systems Pty Ltd

ABSTRACT -- This paper describes a series of experiments conducted by Syrinx to determine performance improvements offered by Auditory Model based speech signal processing front-ends for HMM recognisers. The experiments tested an implementation of the Ghitza Model connected to a HMM recogniser through a number of interface algorithms that reduces the Auditory Model's representation dimensionality to a manageable size. The results show that in high noise environments recognisers incorporating front-ends based on the Ghitza Auditory Model outperform those implemented using traditional Delta Cepstrum speech processing algorithms.

INTRODUCTION

Over the past 5 years there has been continued interest in the applications of Auditory models to increase robustness of speech recognition in high noise environments. Experiments by Seneff (1985) and Ghitza (1987, 1987) have produced some evidence that Auditory models do improve recogniser performance in high noise conditions.

Syrinx was keen to determine the performance improvement offered by using auditory based front-end signal processing algorithms when compared to conventional speech processing algorithms. In a series of experiments, Syrinx compared the performance of a conventional Delta Cepstrum front-end signal processing algorithm with its implementation of the Ghitza Ensemble Interval Histogram (EIH) model.

The Ghitza model consists of three processing elements. Input speech is applied to a filter bank consisting of 40 band pass filters distributed linearly on a frequency scale from 200 Hz to 6707 Hz. Filter outputs are then passed to a threshold crossing detection processor from which 40 individual histograms are constructed for the whole interval of speech, where each histogram is the integration of the period between threshold crossing detected during the processing interval. The final Ensemble Interval Histogram (EIH) output is the composite of the individual interval histogram outputs.

The EIH is a measure of the periods detected between threshold crossing for each bandpass filter in the front-end filter bank. As a consequence, the EIH is effectively a "periodgram" of the input speech signal, where the x-axis corresponds to period (1/f seconds) and the y-axis corresponds to a sum of threshold crossing periods. As the EIH represents a periodgram, its dimensionality needs to be large to both adequately resolve high frequency components and to obtain the necessary frequency coverage. In Syrinx's implementation the EIH has dimensionality of 200. This is well above the dimensionality of conventional speech recogniser front-ends, which are typically 24.

Although the performance gains offered by the EIH could be informally observed through spectrogram representation and by comparisons of LPC fittings, it was difficult to assess the improvements in recogniser performance in noise, if any, offered by the EIH representation. Informal observations have also established that any reductions in algorithm complexity also lead to a concomitant reduction in performance benefits offered by Ghitza Auditory model. The problem