OBJECTIVE IDENTIFICATION OF SPEECH PRESENTED IN NOISE

Richard Katsch , Phillip Dermody[1]John Seymour, Loredana Cerrato[2]
Speech Communication Research, National Acoustic Laboratories

Abstract: The initial investigation of an objective measure of speech processing in noise is presented that uses models of auditory processing as the speech analysis stage and a simple distance measure classifier to provide identification scores for speech presented in noise.

Introduction:

Traditionally, speech intelligibility in noise is assessed by measuring percent correct scores from human listeners who are required to identify the speech presented in a background of noise. The ability of human listeners to process speech in noise can be manipulated by a large range of variables (Dermody, 1994). However, even in difficult listening tasks such as when listeners are asked to identify monaurally presented nonsense syllables presented in noise, their performance levels are still quite high (Dermody, 1994).

For nonsense syllables presented in noise listeners do not have high level linguistic or world knowledge to assist their judgements and so it can be assumed that listeners identify spoken syllables in noise by comparing the incoming acoustic pattern with a stored prototype or exemplar pattern or set. Of course the question of what aspects of the pattern (global or feature components) are important for perception constitutes a major focus of many speech perception studies (eg Lee and Dermody, 1992).

In the present study we investigated the performance of an objective system for measuring speech intelligibility performance. The objective system includes speech processing methods based on human auditory functions and a simple distant measure classifier that uses the signal in quiet as the reference pattern and compares it to the representation of the same sound when presented in noise. The main purpose of the study is to compare a method of auditory processing based on a psychoacoustic model with a composite model incorporating higher level processing capability. Auditory models have frequently been reported to produce good results for speech processing in noise (e.g. Dermody,1993)

Auditory models

The model referred to as the Zwicker model in this report is based on the work of Zwicker (1990) who provides extensive psychoacoustic data about processing in the peripheral auditory system. Our implementation does not model the external and middle ear systems that predominantly add a high frequency emphasis to the spectral output. The auditory model analyses the input signal using an FFT based on a Bark scale. Across channel interaction is modelled using a smearing function that is convolved with the FFT output. It should be noted that our approach only models across frequency channel interaction in a single frame and does not model across frame interaction that occurs from temporal integration of signals. The output of the modelling process is a spectrum every 5 msecs that are added to form an integrated spectrum of the full duration of the speech sound.

The second model used in this study is based on an auditory model by Ghitza (1986; 1987;1988). The model employs 24 filter channels based on the ERB scale. These filters represent the basilar membrane filtering action of the cochlea. The operation of the hair cells and nerve fibres is simulated as an array of level-crossing detectors at the output of each filter. These are logarithmically spaced providing a compressive function for increases of signal power at each filter. The output of the level crossing detectors represents the discharge activity of an ensemble of auditory nerve fibres.

The next stage modelled in the Ghitza model is assumed to be a more central stage of auditory processing which assumes a place-independent structure and operates on detailed timing information in the auditory nerve fibre responses. This structure is simulated by an Interval Histogram at the output of each channel which are summed into an Ensemble Interval Histogram (EIH). The EIH is a measure of the spatial extent of coherent neural activity across the simulated auditory nerve and is represented by the short-term probability density function of the reciprocal of the intervals between successive nerve firings.

The EIH can be converted back to a frequency scale because the tonotopic information is present implicitly in the timing representation. In the implementation of the model used in this study the pseudo-spectrum produced by the EIH is produced every 5 msec and then added to produce an integrated spectrum of the duration of the speech sample.

Speech Sound Classifier

The present method extends a technique investigated by Dermody, Raicevich and Katsch (1993) for measuring performance capability of auditory models in noise. In the present method the integrated spectrum produced as the output of both auditory processing models is used to 'identify' the speech sound presented in noise. The technique uses the integrated spectrum of each speech sound analysed in a quiet condition (no additive noise) as a prototype against which to compare the speech sound spectrum from the same sounds presented in different noise conditions. The comparison of each noisy speech spectrum with each quiet prototype provides an unbiased estimate of identification performance (within the constraints of the speech database used). The technique provides a percent correct score by using the best match from the classification process and sums these across speech sounds and speakers to provide a total percent correct score. Diagnostic information is also available from confusion matrices produced with the method but only percent correct results are presented in this report.

Speech Database and Noise Conditions

The speech database was produced by editing 18 English consonants from continuously spoken sentence material. The recording environment was the same as that reported in Millar, Dermody, Harrington and Vonwiller, (1993). All consonants were spoken in a quiet environment by 40 speakers with approximately equal numbers of different genders. The text of the database had been constructed to produce an example of each of the 18 consonant sounds in only one vowel environment, /a/. The material therefore represents realistically spoken material (as a target in a continuous sentence which was unknown by the speaker) but is constrained by use of only one vowel context.

The spoken corpus was edited to extract the 18 consonant targets of interest in such a way that only the initial consonant portion was segmented. This involved editing the speech sounds using waveform information (supported by spectrograms if necessary) so that no specific transition information into the vowel was included. The editing was carried out by a trained acoustic phonetician who also listened in good listening conditions and confirmed that vowel identity could not be heard in the edited material. This strategy was adopted to ensure that the speech sound database used in this study represented reasonably homogeneous speech sound information about consonant identity. Previous work (eg. Dermody, Mackie and Katsch, 1986) has shown that listeners are able to use just the consonantal portion (ie. before the transition) to accurately identify speech sounds. The speech sounds were checked and then included in the speech database used in this investigation.

Noise:

The noise was digitally added to the speech files and was one sample of large noise database collected in real office environments (Raicevich and Dermody, 1991). The duration of the noise was matched to the duration of each sample speech and adjusted to equal the power of each speech sample. Five different signal to noise ratios were included in the study: +20, +10, +5, 0, -5 dB.

Method

Each speech sample was analysed in quiet and in each signal to noise ratio and processed by each auditory model. The spectral outputs of the models were used to classify the speech sounds by comparing the quiet prototype with the noisy spectra using a distance measure to determine the best match between the quiet and noisy spectra. The results are presented as percent correct for each auditory processing model based on the adding the correct matches and presenting them as a percentage of possible correct matches.

Results

The results of the processing and classification of the speech database are presented in Figure 1.

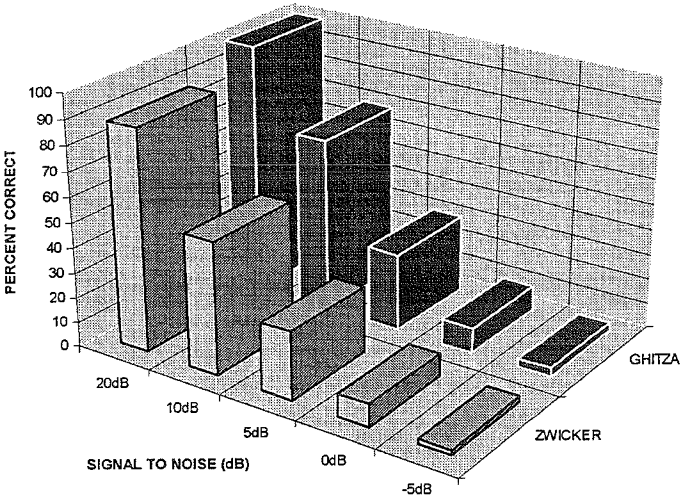

Figure 1

Discussion

The present results indicate that while performance of both auditory model processors is good at a high signal to noise ratio of +20 the performance decreases rapidly for poorer signal to noise ratios.

The present results provide a baseline of performance capability of two auditory processing models. The Zwicker model is based on a simple implementation of filtering and smearing while the implementation of the Ghitza model increases computational complexity to provide better noise cancelling output. The implementation of the Ghitza model does perform significantly better at the +10 and +20 dB signal to noise ratios which suggests that the enhancements provided in that model may be beneficial for processing speech in noise. However both auditory model processors degrade significantly at -5 dB and 0 dB signal to noise ratios although performance from both is still above chance even at the poorest signal to noise ratio.

Further investigations are now needed to determine the level of human identification for the same stimulus and whether additional tuning of the auditory model provides improved performance.

References:

Dermody, P. Mackie, K. and Katsch, R. (1988) Initial speech sound processing in spoken work recognition. Proceedings of First Australian Conference of Speech Science and Technology. Canberra: ANU Printing Service.

Dermody, P. (1992) Human capabilities for speech processing in noise. In Grenie, M. and Junqua, J-C. (Ed) Speech Processing in Adverse Conditions. ESCA: Cannes-Mandelieu.

Dermody, P. (1993) Auditory models for speech processing applications. In Proceedings of Workshop on Signal Processing and its Applications 1993. University of Queensland, Brisbane.

Dermody, P., Raicevich, G. and Katsch, R. (1993) Comparative evaluations of auditory representations of speech In M. Cooke, S. Beet, and M. Crawford (Eds) Visual Representations of Speech. Chichester: John Wiley

Ghitza, O. (1986) Auditory nerve representation as a front-end for speech recognition in a noisy environment. Computer Speech and Language, 1, 109-130.

Ghitza, O. (1987) Robustness against noise: The role of timing-synchrony measurement. Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing.

Ghitza, O. (1988) Temporal non-place information in the auditory-nerve firing patterns as a front-end for speech recognition in a noisy environment. Journal of Phonetics, 16, 109-123.

Lee, K. and Dermody, P. (1992) The relationship between perceptual and acoustic analysis of speech sounds. In Proceedings of the 4th Australian International Conference on Speech Science and Technology. Canberra: ASSTA

Millar, B., Vonwiller, J., Harrington, J. and Dermody, P. (1994) The Australian National Database of Spoken Language. Proceedings of International Conference on Acoustics, Speech and Signal Processing. ICASSP: Adelaide.

Raicevich, G. and Dermody, P. (1991) Predictive analysis of speech communication in noisy telephone oriented office environments In Lawrence, A. (Ed) Proceedings of Internoise '91. Sydney: Australian Acoustical Society
Pp 449-453

Zwicker, E. and Fastl, H. (1990) Psychoacoustics. (Berlin: Springer-Verlag)

[1] Phillip Dermody is now working in less restrictive environments

[2] Loredana was a visiting scientist from Italy.