SIMULATION OF HUMAN INCREMENTAL SPEECH GATING PERFORMANCE USING TIME FREQUENCY ANALYSIS AND A SIMPLE CLASSIFIER.

Daniel Woo, Phillip Dermody
School of Electrical Engineering
University of NSW.

ABSTRACT: The incremental speech gating task is described as a task showing that human listeners can process short duration portions of speech signals to achieve speech sound identification. The results of a group of human listeners on the task is presented and an artificial system using a time frequency spectral analysis and a simple classifier is used to determine its identification performance on the same task. The use of 1 msec analysis frames and an inefficient probability summation method produces a reasonable match to the human performance-duration function in the speech gating task.

INTRODUCTION

Humans can process very brief portions of the onset of speech sounds and use these to begin processing for speech sound identification. For instance, in the incremental speech gating task human listeners are presented with a brief portion of the speech sounds that incrementally increase on each trial. After each presentation the listener is asked to identify the speech sound that was presented. For stop consonants human performance reaches above chance performance at least before 10 msec after onset and reaches a plateau around 30 msec after onset (e.g. Dermody, Mackie and Katsch, 1986). This performance is achieved without any higher linguistic knowledge about the possible identity of the speech sound and must be based on auditory information in the brief duration that is used to compare with a prototype or exemplar pattern set of the speech sound. It is possible that global features of the spectrum of the speech sound may provide sufficient cues for identification (e.g. Lee and Dermody, 1992).

An attempt to model human performance in the speech gating task was reported by Katsch, Woo and Dermody, (199). In that study brief gate durations of stop consonants were processed through an auditory model and then were presented to a neural net classifier which was trained on the identity of a set of stop consonants using only the first 30 msec of the sound as the training pattern. In subsequent testing gate durations which were shorter than the duration used in training were presented to the classifier for identification. The results suggested that the performance-duration function for 10 and 20 msec gates resembled human performance. However the study was limited by a small training database that probably produced the low overall identification performance.

In the present study we attempt to provide a fine-grain simulation of human performance on the incremental speech gating task. In order to do this we have used a time-frequency analysis to provide an estimate of the spectrum of very short durations of speech sounds. The acoustic analysis is then presented to a simple classification method that provides an estimate of recognition of the sounds. The results of the simulation are then modeled as a speech gating task and compared with a human identification function on the same set of stimuli.

Stimuli:

The stimuli were edited from a speech corpus of spoken sentences (the GLASS database) and were recorded under the same conditions reported in Millar, Dermody, Harrington and Vonwiller, 199). Four speakers were selected (two male and two females) and initial stop consonants spoken in an /a/ context were selected and segmented from the spoken sentence. This produced six stop consonants /k, p, t, b, d, g/ spoken by each of 4 speakers.
The first 30 msec of each stimulus was excised and used in the study.

Speech gates were excised from stimulus set every 1 msec for presentation to listeners and for subsequent spectral analysis.

Human performance for an incremental speech gating task:

The set of stimuli were presented to 12 untrained listeners in an incremental speech gating task in which stimuli were presented randomly on any trial. Listeners were asked to identify each stimulus on each trial by selecting one of six possible alternatives from the set /k,p,t,b,d,g/ which were presented on a display. The data was collected automatically and input to a database for subsequent analysis.

Listeners were seated in a sound treated room and listened to the speech sound via high quality earphones with the speech presented at 70dB SPL. A warning was provided before each trial and a response period identified after presentation. Listeners were given a few minutes of familiarity with the stimuli before forma testing was begun.

The listeners were aged between 20 and 30 years and passed a hearing test screen before participating in the experiment.

The results of the testing were analysed for percent correct at each gate duration. Five repetitions of each sound were randomly presented and the results showed good consistency for individuals across each gate duration.

Time-frequency analysis:

Conventional spectral estimation techniques based on Fourier transforms have well known limitations in that improved resolution in either time or frequency produces decreased resolution in the other domain. Time frequency analysis methods attempt to improve resolution jointly in the time and frequency domain.

In the present study a time frequency distribution, P (t,f) for each stimulus was generated using a geometric average of two spectrograms of window sizes 64 and 128 points using a technique described by McLoughlin, Pitton and Atlas (1994). The resulting distribution was then scaled in the frequency domain to a Bark scale. Spectral estimates for each one millisecond of the 30 msec duration speech sounds were obtained with this method.

Speech sound classification:

For each speech sound in the stimulus set 30 spectral patterns were generated using the time-frequency analysis technique. These patterns therefore constituted a reference spectral database of 720 sounds (4 speakers by 6 stops by 30 one millisecond spectral estimates). In order to attempt to classify the sounds a distance measure was calculated for each test sound with the rest of the speech database. All comparisons were then sorted and the top 100 entries (those with the smallest distance) were selected as the potential match set. The responses to each of the 6 speech categories was then assigned from the top 100 entries. A winner was then selected . These operations were performed for each analysis frame producing a histogram for each 1 millisecond time frame as a function of its score on the top 100 entries. In addition a cumulative histogram based on adding the score of the winner across time was also prepared for each speech sound.

The results of this analysis indicated that a winner consistently accumulated over time frames. That is if a /d/ was presented over each successive frame then the /d/ references would win the top 100 entries and could be accumulated to show a gradual rise in identification. The other feature that this analysis indicated is that on average winners received about 20% of the top 100 vote. That is, taking the most active 100 responses of each analysis frame the

"correct" stimulus was assigned about 20 of those responses on average across the time frames (Woo, Dermody and Philips, 1996).

Modeling continuous gating functions with discrete frame response estimates

If correct responses are assigned only a percentage of responses on each frame the problem can be reasonably described as a fuzzy classification problem. Alternatively the problem could be investigated with a probability summation model. The latter approach is adopted in this report.

We assume that on each frame the correct response is assigned winner status (it gets more votes than other categories). If the winner's vote is a proportion averaging 0.2 then we can calculate an approximate growth of information across frames. There are a large number of potential models of combination of information in joint events. In this report we describe one based on what we will call an inefficient probability summation model (Dermody, 1972).

This models the accumulation of information in the joint events as:

$I\,ab = SQRT\,(Ia^2 + Ib^{2)}$

where Iab is the combined information in the joint events and Ia and Ib are the independent information in the events a and b respectively.

If the information accumulates in this way then we can predict the information in a cumulative gating function in which each 1 msec frame is "correctly" identified at around 0.2 on each frame.

RESULTS

The results for the inefficient probability summation model based on parameters from our classification results are presented in Figure 1.

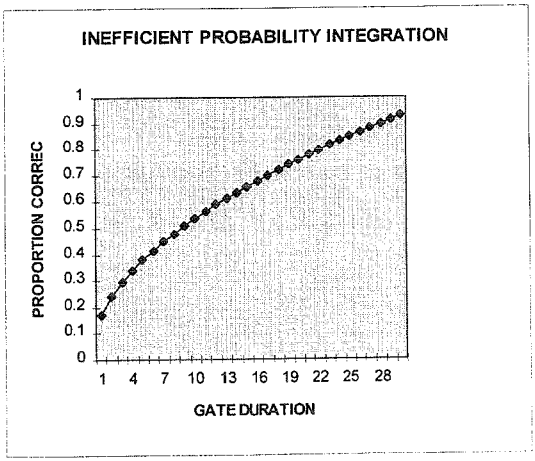

Figure 1 Results of probability summation simulation.

These results show a rising function of proportion correct with increasing gate durations which can be compared with the results of human performance with accumulated information in the incremental speech gating task.

The human listener results are presented in Figure 2:

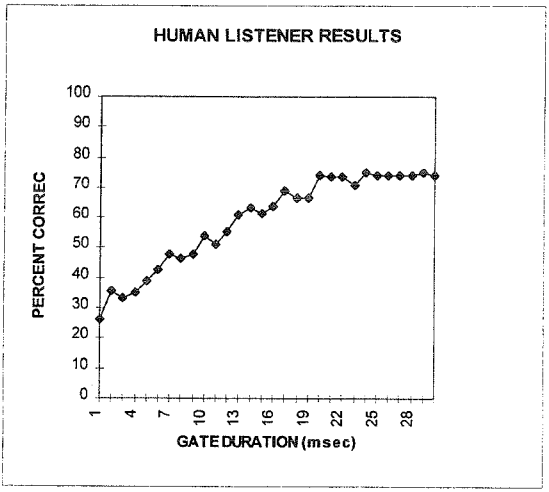

Figure 2. Performance- duration functions for the incremental speech gating task.

DISCUSSION

These preliminary results suggest that short duration portions of the signal can be analysed with sufficient resolution to provide a reasonably consistent input to a simple classifier. The classifier operates by an initial thresholding technique to select the most likely candidates for identification set inclusion. Investigation of the accumulation of the winners performance indicates that the winners do accumulate status over time (Woo,Dermody and Philips, 1996). The amount of information about the winner in any frame is represented by the number of times it is identified in the top 100 identifications. On average this is about 0.2.

In order to calculate the cumulative information in a continuous speech signal of 30 msec duration it is assumed that the information in each frame (e.i.,0.2 correct) accumulates as an inefficient probability summation that produces a predicted identification function across the duration of the speech signal. The predicted function bears some similarity to the obtained human gating function when speech is actually presented as incrementing gate durations. This suggests that a mechanism that analyses very short durations of the speech signal and summates this information across successive time frames is a potential candidate model to describe short duration incremental speech gating results obtained from human listeners. This is consistent with the model of speech gating proposed by Dermody, (1988).

Additional work is underway to determine the effects of fluctuating proportions for the winners across the analysis frames and to determine the best model for combining information across frames.

ACKNOWLEDGEMENTS: The authors would like to thank Chris Philips, Electrical Engineering, University of NSW for support and useful discussions about this study.

REFERENCES:

Dermody, P. (1972) The detection of low energy stimuli presented simultaneously to both visual and auditory modalities as measures of bisensory functions in a psychophysical task. Unpublished honours year thesis, University of Sydney.

Dermody, P. (1988) Perceptual processing of stop consonants. Unpublished PhD thesis, University of NSW.

Katsch, R. Dermody, P. and Woo, D. (1993) Human performance criteria for stop consonant identification using artificial neural nets. Proceedings of the 4th Australian Conference on Neural Nets (ACNN'93) Melbourne.

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

McLoughlin, P., Pitton, J. and Atlas, L. (1994) Construction of positive time-frequency distributions. IEEE Transactions on Signal Processing, 42, 2697-2705.

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.

Woo, D., Dermody, P. and Philips, C. (1996) Analysis of human gating performance using time frequency analysis. Proceedings of the ANZIIS-96.