Signal Processing for a Visual Hearing Aid

E. McDonald, H. Kunov and W. Wong
Sensory Communication Laboratory, Institute of Biomaterials and Biomedical Engineering
University of Toronto, Toronto, Ontario, Canada

Introduction

Speech is one of the most important forms of human communication. Unfortunately, many people who suffer from hearing loss have trouble perceiving and understanding speech - particularly in noisy environments. The long-term goal of this research is to develop a visual aid for people with high frequency hearing loss, the most prevalent form of auditory impairment. This aid would present important speech information through peripheral vision using an LED bar graph mounted in the frame of a pair of glasses.

Previous research suggests that acoustical enhancement of plosives and fricatives can improve the intelligibility of fluent speech[1,2]. Due to the manner in which plosives, fricatives, and affricates are produced, these phonemes should contain significant high frequency energy content. Two strategies were devised to try and detect plosives and fricatives based on the high frequency energy content.

Methods

For an application of a visual hearing aid, it is important that the visual output is perceived as being synchronous with the audio stimulus the user receives. The total delay through the system was required to be less than 15 ms. As each strategy was being simulated in LabVIEW, it was easier to design algorithms which operated on short non-overlapping sequences of data, rather than those which updated the output with every input sample. For the simulation of each strategy, the input signal was broken up into non-overlapping segments of 220 samples (approximately 5 ms at a sampling rate of 44.1 kHz).

The first strategy, High Frequency Energy (HFE), filtered each segment with a fourth order butterworth high-pass filter (f_0 = 3.5kHz) and then calculated the total energy in each segment. This energy was then converted into arbitrary decibel units. The LED value for each segment was then calculated by quantizing the energy output (in dB) into 1 of 9 levels. An 8 (the highest level) corresponded to the global peak value. A 0 (the lowest level) corresponded to an energy level 25 dB or more below the global peak value. In a real time system, the thresholds used for quantization would be based on a long-term average energy.

The second strategy was the High Frequency Energy Ratio (HFER). The first step in this strategy was to estimate the power spectrum using an FFT. The energy in the 3.6-6 kHz band was estimated by summing the energy in the bins corresponding to this region. The energy in the 600-1000 Hz band was estimated in the same way. Finally, the LED value was calculated by quantizing the ratio of the energies in the higher region versus the lower region into 1 of 9 levels. A ratio of 20 or more would correspond to an LED value of 8 (the highest level). A ratio of 1 or less would correspond to an LED value of 0 (the lowest level).

Testing of each detection strategy was carried out using the sentence "Jeff's toy go cart never worked" as spoken by four male and two female speakers. The speech waveforms were downloaded from the TIMIT speech database[3]. Each waveform was inspected by hand to determine the regions where plosives and fricatives were present and those where they were not. Pink noise was then added to each waveform to create four test conditions: clean speech, 12 dB SNR, 6 dB SNR, and 0 dB SNR. A LabVIEW program was used to compare the output of the detector with the hand marked regions for each waveform and calculate the true positive and false positive rates based on a threshold.

By varying the threshold used, Receiver Operator Characteristic (ROC) curves were generated from the resultant pairs of false positive and true positive rates.

Results and Discussion

The areas under the ROC curve for each condition is given in Table 1.

Test Condition	HFE	HFER
Clean Speech	0.670	0.678
12 dB SNR	0.644	0.710
6 dB SNR	0.633	0.720
0 dB SNR	0.591	0.650

Table 1. Areas under ROC curves for Plosive/Fricative/Affricate detector

It is clear that both methods performed poorly at detecting plosives, fricatives and affricates. The area under each curve is not significantly greater than 0.5

(which corresponds to random guessing). A closer examination of the errors suggested that both methods were not detecting the plosive phonemes or voiced fricatives. However, each detector appeared to be reasonably good at detecting unvoiced fricative and affricate phonemes.

A second run was conducted using each method as an unvoiced fricative and affricate detector (detection of a plosive or voiced fricative was be considered a false detection). The areas under the ROC curve for each condition is given in Table 2. In comparison of these results with the previous results, it is clear that both methods performed significantly better as a fricative detector alone than as a plosive and fricative detector (areas of 0.98 vs. 0.67 for HFE, 0.91 vs. 0.68 for HFER).

Test Condition	HFE	HFER
Clean Speech	0.984	0.908
12 dB SNR	0.971	0.975
6 dB SNR	0.951	0.947
0 dB SNR	0.850	0.876

Table 2. Areas under ROC curves for Unvoiced Fricative/Affricate detector on "Jeff's toy go-cart never worked"

Unfortunately, the sentence "Jeff's toy go-cart never worked" does not contain many unvoiced fricatives or affricates. Thus, a second sentence was found which had more unvoiced fricatives and affricates. The sentence chosen was "She always jokes about too much garlic in his food". The TIMIT database contained recordings from seven male speakers. This sentence was processed in the same manner as the first sentence. The areas under the ROC curve for each condition is given in Table 3.

Test Condition	HFE	HFER
Clean Speech	0.936	0.859
12 dB SNR	0.909	0.925
6 dB SNR	0.857	0.880
0 dB SNR	0.727	0.789

Table 3. Areas under ROC curves for Unvoiced Fricative/Affricate detector on both test sentences

The two detection strategies did not perform as well on the second sentence as they did on the first sentence. The cause of this may be due to effects of coarticulation. It was noted when hand marking the wav files for the second sentence that several phonemes appeared to have been significantly affected by coarticulation.

It is also interesting to note that the HFE method performed better than the HFER method in the condition of clean speech (ie. no noise). However, in pink noise, the HFER method performed better than the HFE. As both methods do not require significant computation to perform, perhaps both could be offered in an aid with an option to switch between each strategy depending on the noise condition.

References

[1] D. Ebrahimi, Preliminary research on a peripheral vision lipreading aid, M.A.Sc. Thesis, University of Toronto, 1987.

[2] V. Hazan, S. Simpson, The effect of cue-enhancement on the intelligibility of nonsense word and sentence materials presented in noise, Speech Communication, 24: 1998, 211-22

[3] TIMIT Speech database, Linguistic Data Consortium, Univeristy of Pennsylvania, http://morph.ldc.upenn.edu/lol/timit.html