EVALUATION OF A SPEECH ENHANCEMENT SYSTEM

Yvonne M. Perlmutter*, Louis D. Braids, Ronald H. Frazier**, Alan V. Oppenheim

Research Laboratory of Electronics
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139

ABSTRACT

Frazier (1975) proposed an adaptive comb-filtering technique for enhancing the intelligibility of speech signals degraded by the addition of competing speech signals. This paper reports on a series of preliminary tests of speech intelligibility for materials processed by the proposed system. Sentences spoken by males and females were used as both targets and jammers. Tests were conducted for various combinations of system parameters, speakers and target-to-jammer amplitude ratios. Baseline tests were conducted for processed materials in the absence of interference and for unprocessed materials at various target-to-jammer ratios.

INTRODUCTION

There are many situations in which the ability to understand a given talker is severely limited by the presence of interfering signals. Since a significant portion of speech signal is quasi-periodic (the voiced segments of speech), the effects of the interference may be reduced through the use of comb-filtering techniques. Such a filter would pass only a small band of frequencies about each of the harmonics of the speech signal, rejecting those portions of the competing signal which lie outside the pass-bands of the filter. In 1970 Shields[1] proposed using a time variant comb-filtering technique based on estimates of the fundamental frequency of the speech signal. In 1975 Frazier[2] improved and elaborated upon Shields' system, in developing the adaptive comb-filtering scheme.

This paper reports the results of a series of speech intelligibility tests which were conducted on Frazier's adaptive comb-filtering system.[3]

BACKGROUND

The adaptive comb-filtering system has three internal parameters which control the quality of the processed speech. They are:

1) type of window functions;
2) size of the window function (1, 3, 5, 7, 9, 11, 13 or 15) coefficients, corresponding to a comb-filter impulse response which lasts over 1, 3, 5, 7, 9, 11, 13 or 15 pitch periods of the waveform being processed; 1 coefficient is equivalent of no processing);
3) processing technique for the nonperiodic segments of the target speech signal (the unvoiced segments and the silent intervals). Two techniques are available. In the attenuation method the input signal is attenuated by a constant between 0 and 1. In the inertial method filtering is continued, using the last valid pitch period to determine the filter characteristics.

This paper investigates the intelligibility of speech signals when processed by the adaptive comb-filter with the Blackman window function; using from 1 to 15 window coefficients; and using both methods of treatment of unvoiced/silent segments, with the attenuation constant = 0.3 .

Before the characteristics of the adaptive comb-filter could be established, the above parameters had to be known, along with the pitch contours of the target speech signal. The latter was obtained by applying a threshold test to the glottal waveform[4] (waveform at the speaker's larynx), recorded simultaneously with the speech material recordings. This pitch data was hand corrected before being applied to the filtering systems.

* GTE Sylvania, Needham, Massachusetts

** U.S. Coast Guard, Wildwood, New Jersey

All the speech processing for this research was implemented using the same PDP-11/45 computer facilities as Frazier.

TEST PROCEDURE

The adaptive comb-filtering system was tested for a signal consisting of added speech waveforms of two different speakers (the "target" speaker and the "jammer" speaker). The target speech material consisted of syntactically normal nonsense sentences of the format: "The adjective noun verb, past tense the noun"[5]; for example: "The round work came the well." The jamming material consisted of sentences drawn from the 1965 Revised list of Phonetically Balanced Sentences (Harvard Sentences); for example: "Find the twin who stole the pearl necklace." These sentences had a more varied rhythmic pattern than those used for the target signals. This eased the problem of target-jammer alignment considerably, and simulated the typical situation more closely than would the use of identical target-jammer rhythmic patterns.

All the speech materials were recorded in an anechoic environment. Four speakers were used, two female and two male, all young adults of General Americal dialect. Figure 1 shows the pitch contours of the four speakers, speaking four different target sentences.

Each speaker recorded six 10-sentence target lists, and six 10-sentence jammer lists, for totals of 240 different target sentences and 240 different jammer sentences.

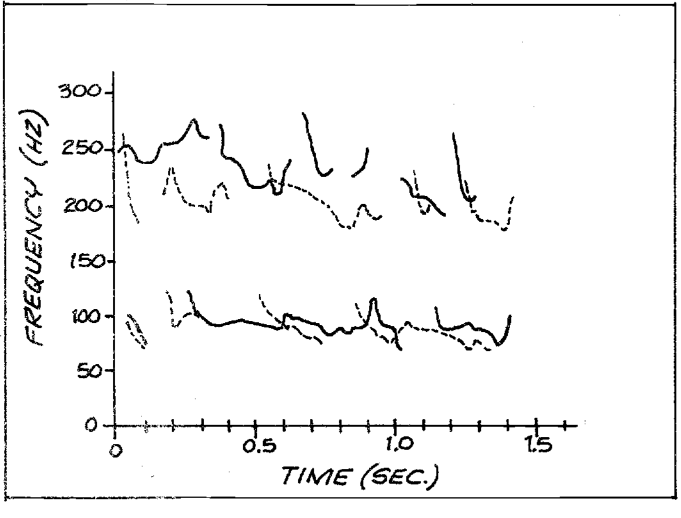

Fig. 1: Pitch contours of the four speakers speaking four different target sentences.

The level of the speech materials was monitored throughout the recording sessions, always keeping the maximum speech peak levels at 95dB. From here on all the sentences were considered of the same level, and reference signals were used to calibrate the equipment throughout the experiements.

The listening sessions were conducted in a sound-proof room, with the test sentences presented diotically over headphones. Ten listeners participated in each test. The listeners were untrained, except for a short exposure to several sample target sentences in the beginning of each listening session.

Responses obtained from the listeners were graded for the percentage of words recorded correctly. A word had to be recorded perfectly and at the correct position within the sentence in order to be considered correct. No substitutions, deletions or additions of vowels or consonants were allowed.

RESULTS

Two preliminary speech tests were conducted to help evaluate the adaptive comb-filtering process. The first test measured the intelligibility of the filtered target sentences with no jammer sentences present (T/J=∞) for all possible numbers of window coefficients (including 1 - or unprocessed signal), for both the attenuation and the inertial methods of treatment of unvoiced/silent segments. The results indicated a linearly decreasing intelligibility score with an increasing number of window coefficients for both the attenuation and inertial methods. The unprocessed target sentences had the highest intelligibility score of 97%; score for 15 window coefficients was 66%. The attenuation method of treatment of unvoiced/silent segments produced slightly higher scores than the inertial method.

The second preliminary test examined the intelligibility of the unprocessed target sentences, with the unprocessed jammer sentences present (i.e.: 1 window coefficient); for six different T/J ratios: -18, -12, -6, 0, +4, +6, and +12dB. The results of this test are shown in Figure 2 together with results obtained by Miller[6] for intelligibility of conversational speech, as masked by one or two voices. The results of our test correspond more closely to Miller's results for speech masked by two voices than by one voice. This is most likely the result of using harder speech materials (conversational speech vs. nonsense sentences), and may also be due to the effects of low-pass filtering (5KHz) and quantizing (12 bits) the waveforms for computer processing.

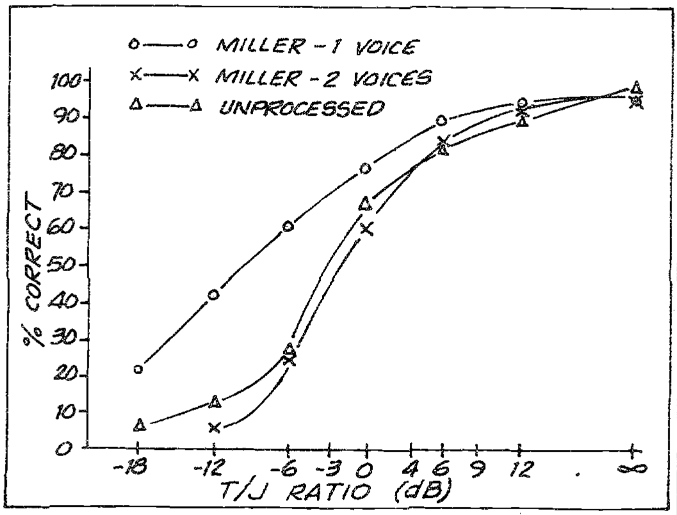

Figure 2: Intelligibility scores for unprocessed speech at different T/J ratios; and Miller's results.

The main speech test investigated the intelligibility of processed target sentences, with processed jammer sentences present. The following input parameters were examined: three different numbers of window coefficients (3, 7, 13); two methods of treatment of unvoiced/silent segments; all the possible combinations of the target-jammer speakers (including self-jamming). The test was conducted for three different T/J ratios: -3, +4 and +9dB.

The most important result of the intelligibility test was that the adaptive comb-filter processing did not improve the intelligibility of speech for any of the system parameter combinations, nor for any of the input conditions. Intelligibility scores on the average decreased as more window coefficients were used for both attenuation and inertial methods of treatment of unvoiced/silent segments, for all T/J ratios tested (see Figures 3 and 4).

The attenuation technique (using attenuation constant of 0.3) gave generally higher scores than the inertial method.

Different target sentence speakers obtained significantly different intelligibility scores. When averaged over all other parameters, the scores for different speakers ranged from 40% to 63%. Since the female speakers had both the highest and the lowest average intelligibility scores, no significance should be assigned to the average score for the male target speakers, as opposed to the female target speakers.

In terms of the masking ability of the jammer sentence speakers, no significant differences were present between either individual, or male vs. female speakers.

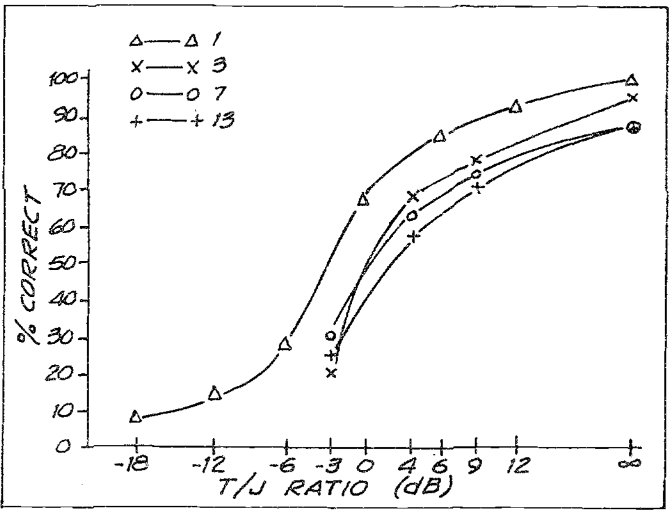

Figure 3: Intelligibility scores for different numbers of window coefficients for the attenuation method.

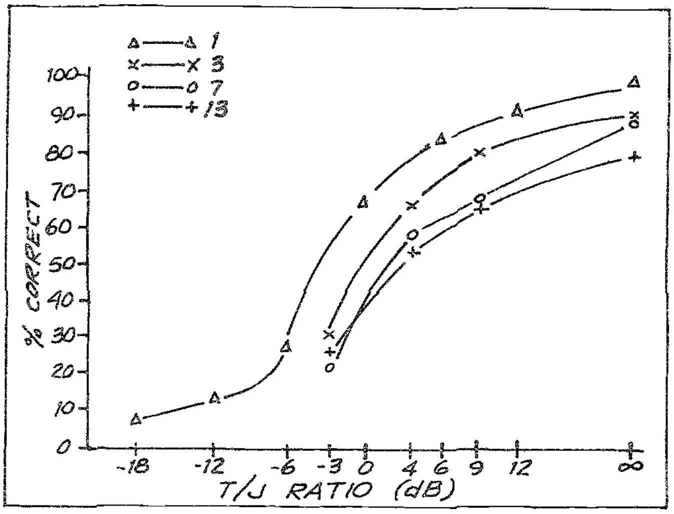

Figure 4: Intelligibility scores for different numbers of window coefficients for the inertial method.

The inter-sex speaker combinations resulted in higher intelligibility scores than intra-sex speaker combinations (see Figure 5). The male-female combination (male target sentence speaker; female jammer sentence speaker) had the highest average intelligibility score of 61%; female-male combination was in second place with a score of 57%; male-male combination obtained a score of 50%; and female-female combination was the hardest to understand with an average intelligibility score of 45%.

Interestingly enough the score for self-masking by male speakers was higher than the score for male-male speaker combinations by about 6%. In case of the female speakers, self-masking produced a score 6% lower than for female-female speaker combinations.

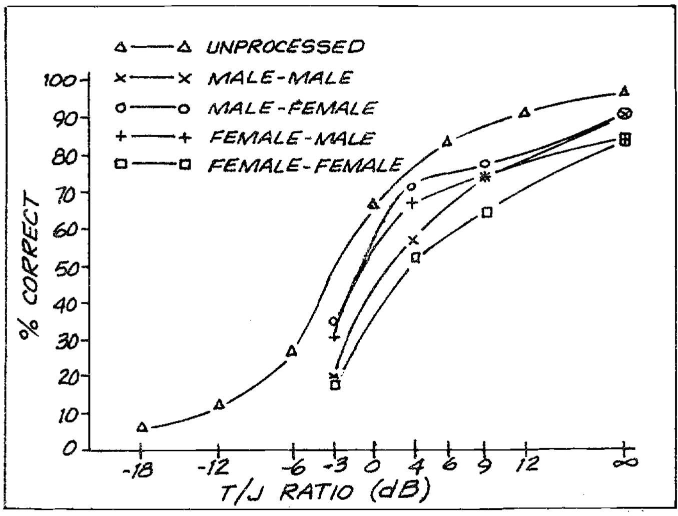

Figure 5: Intelligibility scores for different input speaker combinations.

SUMMARY

Speech intelligibility tests were used to evaluate the performance of an adaptive comb-filtering system for an input signal consisting of two added speech waveforms. The results of these tests indicate that there is no improvement in intelligibility of the desired speech signal for any of the system parameter combinations tested, nor for any of the input conditions used. The results are negative despite the obvious logic of the system; and despite the use of very good pitch information, which would normally not be available to a filtering system.

The reasons for the negative results are at present not understood. The effects of the adaptive comb-filter on speech signals are clearly complex. The degradation of the target signal and the reduction of the jamming signal are the two major results of applying the filter. However, the interactions between these two factors and the overall performance of the system are still unknown; future work is required before they can be explained.

A study of the effects of adaptive comb-filtering on speech enhancement in noise background is currently in progress at the Research Laboratory of Electronics, M.I.T.; the results will be reported in the future.

REFERENCES

1. Shields, V.C., Jr., "Separation of Added Speech Signals by Digital Comb-Filtering," M.I.T. S.M. Thesis, Sept. 1970.

2. Frazier, R.H., "An Adaptive Filtering Approach Toward Speech Enhancement," M.I.T. S.M. and E.E. Thesis, June 1975.

3. Perlmutter, Y.M., "Evaluation of a Speech Enhancement System," M.I.T. S.M. Thesis, Sept. 1976.

4. Henke, W.L., "Signals from External Accelerometers During Phonation: Attributes and Their Internal Correlates," M.I.T. Research Laboratory of Electronics, Quarterly Progress Report, July 15, 1974, pp. 224-231.

5. Nye, P.W., and Gaitenby, J.H., "The Intelligibility of Synthetic Monosyllabic Words in Short, Syntactically Normal Sentences," Haskins Laboratories: Status Report on Speech Research, 1974, pp. 169-190.

6. Miller, G.A., "The Masking of Speech," Psychol. Bull., 44, 1947, pp. 105-129.