

NOISE-IMMUNE MULTISENSOR SPEECH INPUT: FORMAL SUBJECTIVE TESTING IN OPERATIONAL CONDITIONS

V. Viswanathan and C. Henry

BBN Systems and Technologies Corporation
Cambridge, MA 02138

ABSTRACT

In prior research, we have successfully demonstrated the feasibility of using multisensor configurations for transducing speech with enhanced immunity to high levels of acoustic background noise. The goal of the current work has been to evaluate multisensor speech input in operational conditions including an oxygen facemask, different noise conditions, and use with two speech coders and a speech enhancement system. We developed a neckstrap for mounting accelerometers. We selected the best multisensor system and fabricated an Exploratory Development Model (EDM). Results from formal speech intelligibility and quality tests in simulated fighter aircraft noise show that the EDM provides consistently both higher intelligibility and quality, as well as more robust performance under all tested operating conditions, than the microphone alone.

1. INTRODUCTION

The goal of our research in multisensor speech input has been to transduce speech that is more immune to high levels of acoustic background noise than is possible with any single microphone. In our ICASSP-85 paper [1], we presented a two-sensor system involving an accelerometer mounted on the speaker's throat and a noise-cancelling microphone located close to the lips. The accelerometer is relatively insensitive to acoustic noise. The two-sensor signal is the sum of filtered and amplitude-adjusted versions of the two individual sensor signals, with the accelerometer providing the low-frequency information and the microphone providing the high-frequency information. Results from formal speech intelligibility and quality tests in simulated F-15 aircraft cockpit noise showed clearly that the two-sensor signal outperformed the signal from the noise-cancelling microphone alone and that the performance improvement generally increased with the noise level.

In the work we reported in ICASSP-86 [2, 3], we successfully demonstrated the feasibility of high-performance speech recognition, using a parallel-input multisensor system in which the individual sensor signals are not combined to produce a single signal, but instead are used directly to extract features for use in recognition.

With the feasibility of multisensor speech input demonstrated in prior work, the goal of the current work has been to investigate the use of multisensor speech input under a number of operational conditions including an active oxygen mask, different noise conditions, and tandem operations with three real-time speech processors: the 2.4 kbit/s LPC-10 coder, the 16 kbit/s CVSD coder, and the Speech Enhancement Unit (SEU) developed by the Air Force.

2. NECKSTRAP FOR ACCELEROMETER MOUNTING

In our prior multisensor work [1-3], we attached the accelerometer (Vibro-meter Model 501) to the speaker's throat with double-stick tape. This "sticker mount", though adequate for experimental purposes, is not appropriate in an operational situation. A neckstrap mount is expected to be more robust for operational use. Initially, we investigated a commercially available neckstrap-mounted accelerometer, the Controlonics Model MI-1. In our experimental comparisons of the MI-1 with the Vibro-meter 501, we found that the MI-1 picked up more acoustic background noise than the 501, and that the 501's frequency response was significantly flatter than the MI-1's response. Therefore, we developed a neckstrap for mounting the 501 accelerometer. The neckstrap allows mounting of two accelerometers, symmetrically located on each side of the throat. The signals from the two accelerometers are added together, and the sum is then combined with the microphone signal to produce the multisensor output. The rationale for the use of the two-accelerometer mix is to minimize the effects due to head movements; this was verified experimentally.

During the development process, we collected speech in acoustic background noise using the neckstrap-mounted accelerometer and the sticker-mounted accelerometer, and compared the data using listening tests and spectral analyses. The primary neckstrap design goals were to maximize the naturalness of the transduced speech and to minimize the level of the acoustic noise transduced by the accelerometer. Additional design goals included minimizing the slippage of the strap and the accelerometers against the skin as well as any discomfort to the wearer.

The neckstrap we developed is an elastic strap, with strips of Velcro sewn to its ends. The strap is placed around the speaker's neck and held in place by joining the Velcro strips together. The neckstrap holds two identical mounting assemblies, one for each accelerometer. The 501 accelerometer is placed inside an aluminum cylinder lined with a thin layer of fiberglass insulation, which significantly reduces the influence of resonant frequencies in the mount on the accelerometer's noise response. Four strap guides are affixed to the outside of each assembly to allow the assembly to be positioned along the strap. A thin membrane is glued to the inner edge of the cylinder facing towards the speaker; the accelerometer, whose cable passes through a slot in the cylinder, is slid partway through a small washer at the center of the membrane and glued in place. The membrane limits the mechanical coupling between the accelerometer and the cylinder; this helps to isolate the accelerometer from vibrations of the cylinder caused by high background noise. The accelerometer extends slightly past the membrane so that when the mounting assembly is pressed against the throat, the accelerometer is pushed slightly outward. The resulting

tension on the membrane keeps the accelerometer pressed firmly against the throat.

3. EXPERIMENTAL INVESTIGATION

We collected speech data from four male and two female speakers in a noise-free environment and in several simulated F-15 aircraft cockpit noise conditions, using neckstrap-mounted accelerometers and several noise-cancelling microphones, including the following: the microphone M101, which is mounted in a fixed position in an MBU-12P oxygen mask; a prototype in-mask microphone M169, whose position inside the mask can be adjusted to be located close to the lips; and a prototype second-order gradient microphone EV990. Measurements were made with the subject wearing an Air Force helmet, and with and without an oxygen mask. A number of multisensor signals were generated [1]. We performed informal listening tests, spectral analyses, and limited intelligibility tests using the Diagnostic Rhyme Test (DRT). Also, we investigated the use of multisensor speech input with three real-time speech processors: the 2.4 kbit/s LPC-10 coder, the 16 kbit/s CVSD coder, and the Speech Enhancement Unit (SEU) developed by the Air Force. We provide a brief summary of the results below.

3.1 Selection of Best Multisensor System

From comparisons of a number of multisensor systems, we selected the best system for use with the oxygen mask and hence for our EDM design and fabrication. The selected system, denoted as (ACC, M), uses the two-accelerator mix ACC for the low-frequency region and the in-mask noise-cancelling microphone M (M101 or M169) for the high-frequency region, above approximately 1 kHz. The multisensor signal is obtained by summing together the highpass-filtered and gain-adjusted microphone signal with the two-accelerator mix, which has been processed through an optional highpass filter with a cutoff frequency of about 250 Hz. There is no need to lowpass filter ACC, because it is naturally attenuated at high frequencies by skin conduction losses. The optional 250-Hz highpass filtering of ACC was found to reduce the perceived "boominess" of the multisensor signal. The highpass filter cutoff frequency and the relative gain of the microphone signal are the parameters of the multisensor system. We developed guidelines for adjusting these parameters, for use in EDM design and subsequent EDM testing.

Informal listening tests showed that the overall speech quality was clearly better for the multisensor signal than for the in-mask microphone signal. Quality improvements included a reduction in the level of breathing noise and fuller sounding speech.

3.2 Use of Multisensor Speech Input with Three Speech Processors

For this investigation, we used two real-time Department of Defense (DoD) standard speech coders, 2.4 kbit/s LPC-10 (Version 45) and 16 kbit/s CVSD, available at Hanscom Air Force Base, MA, and the real-time Speech Enhancement Unit (SEU) developed by the Air Force [4]. For the speech data we collected in high-noise conditions, the SEU was found to produce the best signal quality when we applied its noise reduction algorithm using the low threshold in adaptive mode [4].

We processed the microphone and multisensor signals with each of the foregoing three speech processors. Also, we processed the SEU-processed microphone and multisensor signals with each of the two speech coders. A number of comparisons were performed via informal listening tests. We mention here two important results. First, previously optimized multisensor system parameters continued to work well for the various multisensor/speech processor tandem operations, with one exception. For processing with LPC-10, the optional 250-Hz highpass filter for the accelerometer must not be used; otherwise, the LPC-10 pitch extraction algorithm produces a large number of pitch and voicing errors, which yield poor speech quality and intelligibility. Second, for all speech processor cases investigated, the processor output had better overall speech quality for the multisensor system than for the microphone.

4. EXPLORATORY DEVELOPMENT MODEL

For the multisensor system selected, we designed and fabricated an Exploratory Development Model (EDM). The EDM accepts three inputs: two accelerometer signals and one microphone signal. The input signals are filtered, amplified (or attenuated), and combined to produce a single output. The EDM was constructed using a Printed Circuit (PC) board; low-power devices and commercially available components were used wherever possible. The EDM provides built-in flexibility so that circuit parameters can be easily changed. The EDM features AGC and squelch control.

5. FORMAL SUBJECTIVE TESTING IN NOISE

The overall test objective was to evaluate the extent of speech intelligibility and quality improvements that the EDM multisensor system provides over a DoD standard aircraft microphone (M101) under a number of operational conditions including an active oxygen mask, different high-noise conditions, and post-processing with each of the three speech processors: the 2.4 kbit/s LPC-10, the 16 kbit/s CVSD, and the SEU. We used three-speaker Diagnostic Rhyme Tests for evaluating speech intelligibility and 10-point rating tests for evaluating speech quality.

5.1 Test Conditions

The number of possible test conditions was determined by the following factors: sensors (EDM, M101, M169, ACC); noise conditions (at least two); speech coders (none, LPC-10, CVSD); an active oxygen mask (in, out); SEU-processing (in, out); and AGC (in, out). To stay within the available testing resources, we chose a total of 36 test conditions. We note that two noise conditions were used for testing: average F-15 cockpit noise at 110 dB (SPL) and at 115 dB [1]. We chose to use these high noise levels because our earlier investigations have shown that with the oxygen mask providing substantial noise attenuation, the in-mask microphone M101 alone performs adequately for lower noise levels such as 105 dB; any performance gains provided by the EDM over M101 at these lower noise levels may not be significant.

5.2 Recording of Test Data in Noise

The cockpit noise was simulated in a reverberant room to ensure a stable, diffuse field. We recorded speech from three speakers, two male and one female. For test conditions involving an oxygen mask, each speaker wore an aircraft

helmet; the mask, attached to the helmet to hold it in position, was connected to a pressurized oxygen cylinder via standard regulators. Speakers used expandable foam-type earplugs to protect against hearing damage at high noise levels.

Speakers wore the neckstrap containing the two accelerometers, which were located symmetrically on either side of the throat, above the Adam's apple, and at the outer edges of the laryngeal cartilage; this location was found to yield the best results in our earlier investigations. For test conditions involving the M169, which has an adjustable in-mask mount, the microphone was located directly in front of the lips such that when the speaker pursed the lips, as during the vowel /u/, the lips touched the microphone surface lightly.

As test speech materials, we used eleven 232-word DRT lists (each test used a randomly selected list), for speech intelligibility evaluation and a set of six sentences that is representative of running speech, for speech quality evaluation. Depending on the number of sensor signals recorded during a particular run, either one or two Kudelski 2-channel tape recorders were used for that run. For each run, the EDM mix parameter values were found for each speaker through informal listening. Very little adjustment of mix parameters was required; an M101 or M169 cutoff frequency of 1 kHz for males and 1.2 kHz for the female was satisfactory for all runs, and the relative gain knob was set to either 0 dB or -3 dB for all runs.

5.3 Generation and Scoring of Test Tapes

One three-speaker DRT tape was generated for each of the 36 test conditions from the data recorded in noise, after necessary post-processing was performed with LPC-10, CVSD, and SEU. The DRT tests were administered and scored at Dynastat, Austin, TX, using eight listeners.

For speech quality evaluation, we prepared a test tape containing a randomized order of 108 sets (=36 conditions x 3 speakers) of six sentences. The data on the speech quality test tape was tested at BBN with six listeners in three sessions, with each session run on a different day. The listeners were asked to rate the overall quality of each sentence set on a 10-point scale, with 1 being the worst quality and 10 being the best quality. The testing procedure yielded a total of 18 (6 listeners x 3 sessions) quality ratings for each of the 108 sentence sets.

5.4 Discussion of Test Results

For all test conditions considered here, we observe that except where noted below, the in-mask microphone M101 was used in the EDM multisensor system; an oxygen mask was used; and the 250-Hz highpass filter was applied to the accelerometer mix, except for cases involving LPC-10 post-processing of the EDM signal.

Table 1 gives three-speaker average DRT and quality rating scores for the EDM and the microphone. In the "Processing/System ID" column, the entry SEU, Coder (LPC or CVSD) refers to the case in which the sensor signal was first SEU-processed and then processed through the coder. Of the 36 total test conditions used, the table gives results for 22 conditions (=11 "Processing/System ID" cases x 2 sensors). Below, we consider each of a number of important comparisons and discuss the test results.

Comparing EDM with the microphone, we see from

Table 1 that EDM is better than the microphone in *both* speech intelligibility and quality, over all 11 test conditions. Except for test conditions 9 and 10, the quality improvements provided by the EDM over the microphone are all statistically quite significant ($P < 0.01$ or better). Several of these improvements (e.g., 1, 3, 5, and 7) are extremely significant ($P < 0.000001$); all of these cases involved 115 dB noise and no SEU-processing. For conditions 9 and 10, the noise level was lower and SEU-processing was used; not surprisingly, the quality improvements due to the EDM were small and statistically not significant. Relatively large speech quality improvements due to the EDM occurred for the CVSD and no-mask conditions in

No.	Processing/ System ID	DRT		Quality	
		EDM	M101	EDM	M101
	115 dB Noise				
1	None	89.7	87.7	7.8	6.6
2	SEU	82.6	79.0	7.6	6.8
3	LPC	63.3	48.7	2.4	1.4
4	SEU, LPC	61.3	54.9	2.7	2.0
5	CVSD	76.9	69.4	4.4	2.6
6	SEU, CVSD	76.6	69.6	4.9	3.6
7	No Mask	88.6	81.2	7.7	6.1
8	M169, SEU	86.7	86.4 (M169)	8.9	8.1 (M169)
	110 dB Noise				
9	SEU	86.5	84.4	8.2	8.1
10	SEU, LPC	67.6	55.1	3.0	2.7
11	SEU, CVSD	80.7	75.3	5.5	4.5

Table 1 Three-speaker DRT and quality scores for EDM and the microphone

115 dB noise. Considering speech intelligibility, we see from Table 1 that improvements provided by the EDM over M101 range from 2.0 to 14.6. The improvements for conditions 1 and 9 are only marginally significant ($P < 0.1$). All other improvements (except the one for No. 8) are quite ($P < 0.01$) or extremely ($P < 0.000001$) significant. Considering condition 8, where we used M169 and SEU-processing, the improvement is small and not statistically significant. The improvement may have been significantly larger had we not used SEU-processing.

To evaluate the effect of noise level on sensor performance, we compare the test conditions given in Table 1 in a pairwise manner (115 dB vs 110 dB) as follows: 2 vs 9, 4 vs 10, and 6 vs 11. Considering speech quality, we observe that performance degradation due to the increase in acoustic background noise level is smaller for the EDM than for M101, for all three comparisons. The same result is true for speech intelligibility for the SEU and SEU, CVSD conditions; for the SEU, LPC condition, however, M101 provides a smaller degradation. Thus, the EDM provides relatively stable performance as compared to M101.

To evaluate the effect of LPC coding, we make the following pairwise comparisons of the test conditions in Table 1: 1 vs 3, 2 vs 4, and 9 vs 10. Similarly, the comparisons for CVSD coding are as follows: 1 vs 5, 2 vs 6, and 9 vs 11. Both sets of comparisons show clearly that both intelligibility and quality losses due to coding are consistently smaller for EDM than for M101, once again demonstrating the robust performance of the EDM.

To evaluate the effect of not using the oxygen mask, which eliminates the benefit of the mask's substantial attenuation, we compare test conditions 1 and 7 in Table 1. The intelligibility and quality differences are, respectively, 1.1 and 0.1 for EDM and 6.5 and 0.5 for M101; this further strengthens the result that EDM is more robust than M101.

To evaluate the effect of SEU-processing, we make the following pairwise comparisons of the test conditions in Table 1: 1 vs 2, 3 vs 4, and 5 vs 6. From these comparisons, we see that, overall, SEU-processing helps M101 more than it helps EDM. Because the EDM provides a relatively noise-immune signal, using SEU-processing on the EDM output improves speech quality slightly but produces, in general, lower DRT scores.

Because our design goal for the EDM was to produce a relatively noise-immune speech signal, it is natural to ask how well the EDM performs relative to a single-sensor speech enhancement technique (SEU-processing is an example). To answer this question, we compare the results given in Table 1 for EDM with those for SEU-processed M101 for the three 115 dB cases: None (no coding), LPC, and CVSD. The improvements provided for the three cases by EDM over SEU at this high noise level are, respectively, 10.7, 8.4, and 7.3 in the DRT score and 1.0, 0.4, and 0.8 in the quality rating score. We note, however, that the SEU was not optimized for the conditions used in the above comparisons.

Comparing test conditions 2 and 8 in Table 1, we note that the prototype in-mask microphone M169 with an adjustable mount is substantially better than the fixed microphone M101, providing an improvement of 7.4 in the DRT score and 1.3 in the quality rating score.

Next, we present test results (not given in Table 1) for three cases. First, use of AGC in the EDM degrades speech quality slightly. Second, we note that the three-speaker average DRT and quality scores for EDM in 115 dB noise (no post-processing) are, respectively, 89.7 and 7.8 for the case involving the accelerometer signal highpass filtered at 250 Hz and 90.3 and 8.6 for the case involving the unfiltered accelerometer signal. EDM without the accelerometer highpass filter is therefore slightly better; this is in contrast to the result, from earlier informal listening tests, that using the highpass filter improves quality by reducing the perception of boominess in the multisensor signal. Third, the DRT and quality scores for a single accelerometer are, respectively, 74.6 and 6.2 for the neckstrap mount we developed and 68.8 and 5.7 for the sticker mount. The corresponding EDM scores are, respectively, 89.7 and 7.8 for the neckstrap mount and 88.2 and 7.3 for the sticker mount. These results demonstrate that the neckstrap mount is better than the sticker mount.

ACKNOWLEDGMENTS

This work was sponsored by the Air Force Systems Command, Rome Air Development Center, Griffiss Air Force Base, New York 13441. The authors would like to thank the following individuals for their help: J. Campbell, J. Cupples, A. Derr, A.W.F. Huggins, F. Iacovino, G. Kang, R. McKinley, H. Mierlak, J. Pianin, W. Russell, G. Spagnuolo, J. Stanley, K. Stevens, and D. Whittemore.

REFERENCES

1. V. R. Viswanathan, C.M. Henry, and A.G. Derr, "Noise-Immune Speech Transduction Using Multiple Sensors", ICASSP-85, Tampa, FL, March 1985, Paper No. 19.1.
2. V. Viswanathan, C. Henry, R. Schwartz, and S. Roucos, "Evaluation of Multisensor Speech Input for Speech Recognition in High Ambient Noise," ICASSP-86, Tokyo, Japan, April 1986, Paper No. 3.2.
3. S. Roucos, V. Viswanathan, C. Henry, and R. Schwartz, "Word Recognition Using Multisensor Speech Input in High Ambient Noise", ICASSP-86, Tokyo, Japan, April 1986, Paper 14.11.
4. M. Weiss and E. Aschkenasy, "Wideband Speech Enhancement Addition", RADC Technical Report RADC-TR-81-53, May 1981.