

NOISE-IMMUNE SPEECH TRANSDUCTION USING MULTIPLE SENSORS

Vishu R. Viswanathan, Claudia M. Henry,
and Alan G. Derr

Bolt Beranek and Newman Inc.
Cambridge, MA 02238

ABSTRACT

In this ongoing work, we have developed and tested configurations of multiple sensors for transducing speech with enhanced immunity to acoustic background noise. In this paper, we describe our work to improve upon our previously reported two-sensor configuration consisting of an accelerometer and a gradient microphone, and also present several additional multisensor configurations including one that uses a second-order gradient microphone for low frequencies and a first-order gradient microphone for high frequencies. Results from formal speech intelligibility and quality tests in simulated fighter aircraft cockpit noise show that each of the multisensor configurations tested outperforms the constituent individual sensors.

1. INTRODUCTION

The overall objective of this ongoing research is to develop and test multisensor configurations of existing sensors for transducing speech that is more immune to acoustic background noise than is possible with any single microphone. Speech input from a multisensor configuration is the sum of filtered and amplitude-adjusted versions of the individual sensor signals.

In the work we reported in ICASSP-84 [1,2], we performed detailed measurements of the sound field in the vicinity of the mouth and neck during speech using pressure and pressure gradient (noise-cancelling) microphones and an accelerometer that measures the skin vibrations. We used the first-order pressure gradient output (denoted in this paper as M12) from a three-microphone array we fabricated, ElectroVoice's first-order gradient microphone EV 985, and Vought's prototype second-order gradient microphone. Compared to the microphones we used, the accelerometer is essentially insensitive to acoustic noise, especially at low frequencies. Through long-term and short-term spectral analyses and articulation index study of the measured sound field data, we determined the best position for each sensor; also, we developed a two-sensor configuration involving the accelerometer attached to the speaker's throat and one of the gradient microphones located in front of the lips. Results from formal speech intelligibility and quality tests in simulated fighter aircraft cockpit noise showed clearly that each of the two-sensor signals under test outperformed the signal from the gradient

microphone alone and that the performance improvement generally increased with the noise level.

In this paper, we report on a method we investigated for improving the foregoing two-sensor configuration, and describe several additional multisensor configurations we developed. Also, we present the results of formal speech intelligibility and quality tests comparing the various multisensor configurations and the individual sensors in simulated fighter aircraft cockpit noise.

2. SPECTRAL SHAPING OF ACCELEROMETER SIGNAL

The accelerometer signal is severely attenuated at high frequencies because of skin conduction losses. However, results from informal listening tests on the accelerometer signal processed through a variable-cutoff highpass filter showed that it contains useful information (intelligible speech), though low in level, even at frequencies above 3 kHz. Spectral analysis of the accelerometer signal also indicated the existence of high-frequency signal amplitudes larger than the accelerometer's noise floor, for voiced sounds. Encouraged by these results, we investigated a method of spectrally shaping the accelerometer signal; this method emphasizes the high-frequency amplitudes relative to the low-frequency amplitudes, to enhance the low-level high-frequency information and to reduce the unnatural quality (e.g., boominess) due to the exaggerated low-frequency amplitudes. The spectrally shaped accelerometer signal would then be combined with a gradient microphone's signal to obtain the two-sensor signal mentioned above.

We determined the spectral shaping function through informal listening tests on the accelerometer signal processed using an adjustable one-third octave filterbank. The shaping we chose provides a sharp highpass at about 800 Hz, a 5 dB/octave boost over 800-2800 Hz, a flat response over 2.8-4.7 kHz, and a sharp lowpass at 4.7 kHz. The shaped version of the accelerometer signal recorded in 107 dB SPL cockpit noise contained a fair amount of high-frequency noise; the noise level was reduced appreciably when we lowpass filtered the shaped signal at about 3.2 kHz. In subsequent tests, we used the shaper-lowpass filter cascade.

To evaluate the effect of the above spectral shaping method, we performed informal speech

19.1.1

LGE EXHIBIT NO. 1020

Amazon v. Jawbone
U.S. Patent 8,467,543
Amazon Ex. 1020

quality tests on the accelerometer signal and on the two-sensor signal (we used the first-order gradient microphone M12), with and without spectral shaping and in 100 dB, 107 dB, and 114 dB levels of cockpit noise. At all three noise levels, the quality of the accelerometer signal alone was improved significantly by shaping. The two-sensor signal with shaping showed a slight but clearly audible improvement in quality over the unshaped case.

3. A TWO-MICROPHONE CONFIGURATION

As part of our long-term spectral analysis work, we compared the long-term speech spectrum with the long-term noise spectrum for each of our three gradient microphones. This study showed that Vought provides a higher signal-to-noise ratio (SNR) at low frequencies than M12 and EV 985, and that both M12 and EV 985 provide a higher SNR at high frequencies than Vought, with the high-frequency SNR being significantly higher for M12 than for EV 985. Fig. 1(a) shows the long-term speech and noise spectra for Vought located in front of the lips (Position 7, see Fig. 1 in [1]), and Fig. 1(b) shows the corresponding spectra for M12 located 45 degrees to one side of the mouth (Position 3). (The speech spectra shown are peak speech spectra since we have added 12 dB to the RMS spectra, and the noise spectra correspond to the responses of the microphones to cockpit noise at about 102 dB SPL.) Figs. 1(a) and 1(b) show clearly the superiority (in terms of the SNR performance) of Vought at low frequencies and the superiority of M12 at high frequencies. The reason for the superior high-frequency response of M12 is that we located the microphone array substantially closer to the mouth than we could locate the bulkier Vought [2].

More important, the above result suggests a two-microphone configuration in which a lowpass filtered Vought microphone signal is combined with a highpass filtered and amplitude adjusted M12 signal, with the same cutoff frequencies for the two filters. The amplitude adjustment is performed to make the high-frequency energy of the M12 signal equal to that of the Vought signal, which ensures

the proper energy balance in the spectrum. We determine the value of the cutoff frequency, using an exhaustive procedure, as that which maximizes the overall SNR (or alternatively the overall articulation index) of the two-microphone configuration. The result is shown in Fig. 1(c), where the cutoff frequency is indicated by the dashed vertical line. The overall SNR of the two-microphone configuration was increased in this example by about 3.7 dB as compared to the Vought microphone alone. The articulation index score (which is a measure of speech intelligibility) was also increased significantly.

The two-microphone system just described is a desirable alternative to the earlier two-sensor system since the accelerometer used in the latter can pick up unwanted vibrations caused, for example, by the movement of the user.

4. ADDITIONAL MULTISENSOR CONFIGURATIONS

Results from short-term spectral analysis of the previously measured sound-field data indicated that the spectral amplitudes of nasal murmur (in relation to an adjacent vowel) were substantially weaker for all three gradient microphones we used than for a pressure microphone [1,2]. We, therefore, investigated the use of a nasal microphone as part of a multisensor configuration. Using the EV 985 for this purpose, we found that the best way of locating it is to place directly under one nostril and not use a puffscren, which allows the microphone to be very close to the sound source. Initial experiments in simulated noise yielded two results: 1) It is necessary to highpass filter the nasal microphone signal at about 500 Hz prior to combining with a lip microphone to reduce the level of the perceived noise; and 2) the (M12, EV 985) combination, with EV 985 being the nasal microphone, produced a slight improvement in the transduction of nasal sounds over M12 alone.

The results presented above suggest the following configurations involving three or four sensors: 1) (Vought, M12, and EV 985); 2) (accelerometer, Vought, M12, and EV 985); and 3)

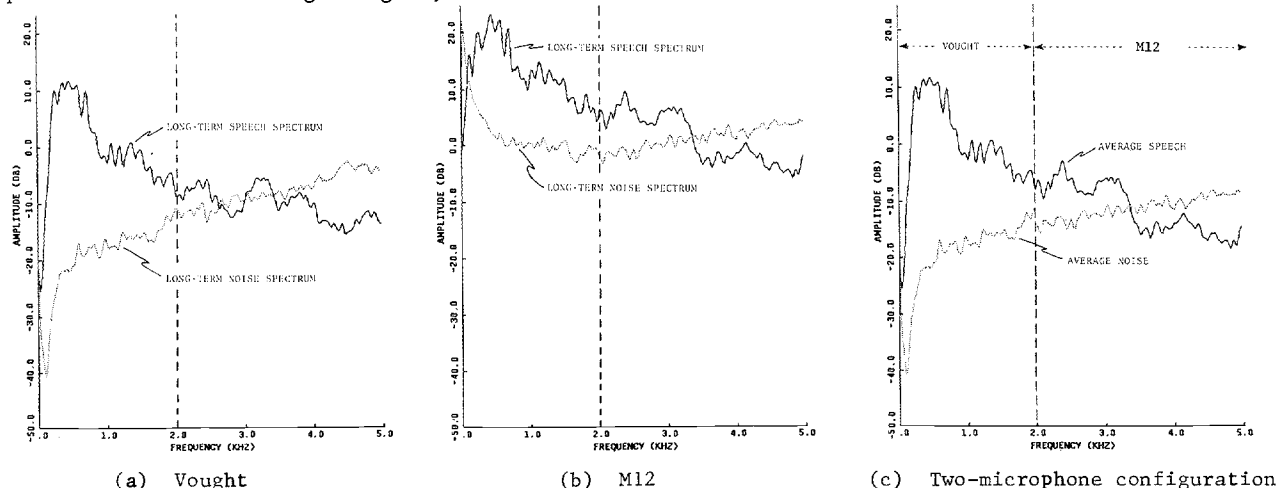


Figure 1. Comparison of the speech and noise responses for three cases.

(accelerometer, Vought, and M12). In cases (1) and (2), EV 985 is included as a nasal microphone. In cases (2) and (3), the Vought signal must be bandpass filtered before combining with the other sensor signals so that the accelerometer contributes information primarily to the low band; the Vought, to the middle band; and the M12, to the high band.

5. SUBJECTIVE TESTING IN NOISE

To compare the performance of the various multisensor configurations in noise, we simulated fighter aircraft cockpit noise at 95 dB and 115 dB SPL, and conducted a two-speaker Diagnostic Rhyme Test (DRT) [3] for speech intelligibility evaluation and a 10-point rating test for speech quality evaluation. The cockpit noise was simulated in a reverberant room to ensure a stable, diffuse field.

5.1 Sensor Positions and Configurations

For each of the sensors, we determined the best position based on the results from articulation index analysis [1,2], short-term spectral analysis, and puffnoise study in which we evaluated, by listening, the level of the puffnoise picked up by a microphone with and without a puffscren. Positions directly in front of the lips (Positions 1 and 7; see Fig.1 in [1]) often led to objectional levels of puffnoise, even though they produced high articulation index scores. The sensor positions we chose are as follows: one side of the throat just above the Adam's apple (Position 10) for accelerometer; 45 degrees to one side of the mouth (Position 3) for M12; 45 degrees to the other side of the mouth for Vought; directly under one nostril for EV 985, as we are using this exclusively as a nasal microphone. We decided to use puffscreens for M12 and Vought and no puffscren for EV 985.

From the results presented in Sections 2-4, we chose to evaluate the following seven multisensor configurations, involving four different sensors:

- C1) Accelerometer and Vought
- C2) Accelerometer and M12
- C3) Accelerometer with shaping and M12
- C4) Vought and M12
- C5) Accelerometer, Vought, and M12
- C6) Vought, M12, and EV 985
- C7) Accelerometer, Vought, M12, and EV 985.

For mounting the multiple microphones, we developed a headpiece from a face shield by attaching metal rods that are adjustable both vertically and horizontally.

5.2 Multichannel Taping of Test Data

To guarantee identical speaking conditions for the seven multisensor systems, we mounted all four sensors simultaneously, taped the sensor signals on a multichannel tape recorder, and combined afterwards the individual sensor signals to obtain the multisensor test data. We used two speakers, one male (RS) and one female (CH). Three ambient conditions were considered: quiet or no noise, 95 dB (cockpit) noise, and 115 dB noise. As test

speech materials, we used 1) the 232-word DRT lists, for speech intelligibility evaluation (a different list used for each speaker-background condition) [3]; 2) a set of six sentences that is representative of running speech, for speech quality evaluation [1,2]; and 3) a set of 20 words (10 digits and 10 command words) [4], for future isolated-word recognition tests using a commercial recognition device. (Speakers spoke the 20-word set 30 times to provide training and test tokens.)

For combining the sensor signals in each of the seven multisensor systems given in Section 5.1, the filter cutoff frequencies and the amplitude adjustments for individual sensor signals were "optimized" subjectively through informal listening tests involving three listeners; we considered the two cockpit noise conditions only. For each sensor signal, we used about the same filter cutoff frequency for both speakers and both noise conditions; the required amplitude adjustments, however, varied across speakers and noise conditions.

5.3 Screening Evaluation

We compared the seven multisensor systems through informal listening tests and limited DRT tests, scoring only half the DRT words and with only three listeners. From this evaluation, we made the following observations: 1) Nasal sounds were transduced well by the two-sensor configurations C1, C2, and C3, all involving the accelerometer; 2) the benefit provided by the nasal microphone EV 985 in configuration C7 over C5 was only marginal; 3) the effect of including the nasal microphone in C6 (as compared to C4) was mixed in that perception of nasal sounds was slightly improved and perception of sounds [g, b, v, d] was degraded; and 4) the improvement provided by C5 over C1 or C4 was only slight. Based on these results and our interest to keep the number of sensors in a configuration small, we decided to include only the four two-sensor configurations C1-C4 for formal DRT and quality tests. For comparison purposes, we included also the three individual sensor cases: shaped accelerometer, Vought, and M12.

5.4 Generation and Scoring of Test Tapes

We generated DRT test tapes for a total of 28 conditions (= 7 sensor systems x 2 speakers x 2 noise levels). The tapes were scored at RADC, Hanscom Air Force Base, MA, using a panel of 12 trained listeners. For speech quality tests, we dubbed the six-sentence sets twice more to obtain a total of 84 sets (= 28 conditions x 3 repetitions); dubbing was done to obtain three quality judgments per set from each listener. The test tapes contain a randomized order of these 84 sets. We used six experienced listeners, who rated the overall quality of each six-sentence set on a 10-point scale, with 1 being the worst quality and 10 being the best quality.

5.5 DRT and Quality Test Results

Table 1 gives the two-speaker average DRT scores

and mean speech quality ratings for the three individual sensors and four two-sensor configurations. In the table, we have used the abbreviation ACC to denote accelerometer without spectral shaping and ACC* to denote accelerometer with shaping. Comparing the two gradient microphones M12 and Vought, we see from Table 1 that in 95 dB noise M12 produces a slightly higher DRT score and about the same quality rating; in 115 dB noise, however, Vought produces a noticeably higher DRT score and a slightly higher quality rating. This DRT result is consistent with the findings from our articulation index study. The new two-microphone configuration (Vought, M12), which we introduced in this paper, produces good improvements in DRT and quality scores over Vought and M12 in 115 dB noise, and produces a higher quality rating and essentially the same DRT score in 95 dB noise. Each of the two-sensor configurations (ACC, M12) and (ACC, Vought) produces substantial intelligibility and quality improvements over the respective gradient microphone in 115 dB noise, and produces essentially the same intelligibility and a higher quality rating in 95 dB noise; this result is in agreement with the result reported in [1]. Notice that the two-sensor configurations involving the accelerometer produce huge improvements in the DRT score over the accelerometer alone, even in 115 dB noise. Comparing (ACC, M12) with (ACC*, M12), we see from the table that spectral shaping of the accelerometer signal lowers the DRT and quality scores in both noise levels. In our view, our findings on the existence of useful information at high frequencies of the accelerometer signal are still valid; the inferior results just mentioned show that the particular shaping method we used is not effective. Among the two-sensor configurations, (ACC, M12) seems to have the best overall performance, although the performance differences among them are not large. The quality ratings given in Table 1 suggest that the listeners were primarily judging the level of the background noise in the test sentences.

We investigated in detail the DRT scores for the two individual speakers RS (male) and CH (female) and for each of the six attributes: voicing, nasality, sustention, sibilant, graveness, and compactness [3]. Because of space limitation, we

SENSOR	95 dB		115 dB	
	DRT	QUALITY	DRT	QUALITY
ACC*	82.1	4.2	79.3	4.0
M12	95.5	6.8	85.2	2.3
VOUGHT	94.2	6.7	88.2	2.8
(ACC, M12)	94.8	7.4	92.4	4.1
(ACC*, M12)	94.0	7.0	91.6	3.4
(ACC, VOUGHT)	94.9	7.5	91.4	3.7
(VOUGHT, M12)	94.8	7.2	90.7	3.1

Table 1. Average speech intelligibility and quality results

keep our comments brief. For the accelerometer alone, the DRT score in 115 dB noise is substantially higher for CH than for RS (the difference between their scores in 95 dB noise is only slight); consequently, the accelerometer's contribution to the two-sensor systems in 115 dB noise is higher for CH than for RS. For example, while the DRT score for (ACC, M12) is 4.3 points lower for CH than for RS in 95 dB noise, in 115 dB noise it is, in fact, 3.5 points higher for CH than for RS. Next, let us consider the DRT scores at the attribute level for CH in 115 dB noise. The addition of the accelerometer (no shaping) to M12 in the two-sensor system improves voicing by 13 points, sustention by 20 points, and graveness by 14 points. Considering (ACC, Vought), the major improvement due to the accelerometer is in sustention (15 points). The addition of Vought to M12 in the two-microphone system improves voicing by 13 points, sustention by 7 points, and graveness by 5 points.

6. CONCLUSION

We developed a spectral shaping method for improving the performance of a two-sensor configuration consisting of an accelerometer and a gradient microphone. Also, we developed several additional multisensor configurations including a two-microphone system. A selected set of two-sensor systems and individual sensors were tested in 95 dB and 115 dB levels of simulated fighter aircraft cockpit noise, using formal speech intelligibility (DRT) and quality (10-point rating) tests. The test results show the spectral shaping method to be ineffective. The two-sensor systems tested produce essentially the same DRT scores and quality ratings in 95 dB and much higher DRT scores and quality ratings in 115 dB, as compared to the constituent individual microphones.

ACKNOWLEDGMENT

This work is being 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: M. Alakel, J. Campbell, C. Dietrich, G. Kang, K. Karnofsky, W. Lundgren, S. Koucos, R. Schwartz, A. Segota, K. Stevens, J. Tardelli, and C. Teacher.

REFERENCES

1. V.R. Viswanathan, et al., "Multisensor Speech Input for Enhanced Immunity to Acoustic Background Noise," Proc. ICASSP-84, pp. 18A.3.1-18A.3.4.
2. V.R. Viswanathan, et al., "Multisensor Speech Input," Final Tech. Report No. RADCTR-83-274, Dec. 1983.
3. W.D. Voiers, et al., "Research on Diagnostic Evaluation of Speech Intelligibility," Final Report, Contract No. AF19628-70-C-0182, Air Force Cambridge Research Laboratories, 1973.
4. G.R. Doddington and T.B. Schalk, "Speech Recognition: Turning Theory to Practice," IEEE Spectrum, pp. 26-32, Sept. 1981.