ROBUST SPEECH RECOGNITION IN THE AUTOMOBILE

Nobutoshi Hanai and Richard M. Stern*

Department of Electrical and Computer Engineering
and School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

ABSTRACT

In this paper we discuss a number of the ways in which the recognition accuracy of automatic speech recognition systems is affected by ambient noise in the automobile, along with the extent to which various techniques for robust speech recognition can provide for more robust recognition. We consider separately the effects of engine noise, interference by turbulent air outside the car, interference by sounds from the car's radio, and interference by the sounds of the car's windshield wipers. Recognition accuracy was compared using baseline processing, cepstral mean normalization (CMN), and codeword-dependent cepstral normalization (CDCN). The greatest degradation in recognition accuracy was produced by interference from AM-radio talk shows. The use of CMN and especially CDCN was found to be significantly improve recognition accuracy, except for the effects of interference from radio talk shows at low car speeds. This type of interference is effectively suppressed through the use of adaptive noise cancellation techniques.

1. INTRODUCTION

The need for robustness in speech recognition accuracy in real applications environments such as long-distance telephone lines, automobiles, aircraft cockpits, offices, and factory floors is becoming increasingly important as speech recognition is becoming more successful. This paper concerns speech recognition accuracy in the automobile, which is a critical factor in the development of hands-free cellular telephony. Major factors that impede recognition accuracy in the automobile include noise sources such as tire and wind noise while the vehicle is in motion, engine noise, and noise produced by the car radio, fan, windshield wipers, horn, turn signals, etc.

A number of researchers have considered the problem of robust recognition in the automobile previously. Their approaches include adaptive noise cancelling techniques (*e.g.* [1, 2]), spectral transformation [3], the use of microphone arrays [*e.g.* 4], and multi-dimensional HMMs [5]. For the most part these studies dealt only with "running noise" sources such as tire, engine, and wind noise, and they did not consider "functional noise" caused by functional components such as the car radio, fan, and windshield wipers. In this paper we consider the effects of all of these sources of degradation, and we compare the extent to which these effects are ameliorated by the compensation techniques of *cepstral mean normalization* (CMN) [6], and *codeword-dependent cepstral normalization* (CDCN) [7].

2. DATABASES

The experimental results in this paper were obtained by training the CMU SPHINX-I system [8] on the previously-described census database [7] and tested using a database of speech recorded in automobiles recorded by and provided by the Motorola Corporation. In this section we describe the Motorola automotive database which was used to evaluate effects of the noise in the automobile on the SPHINX system. We also briefly review the contents of the census database.

2.1. The Motorola Automotive Database

The Motorola automotive database consists of 12 speakers: 9 males and 3 females in their 20s and 30s. Each speaker uttered six 7-digit strings at three driving speeds: 0 (with engine idling), 30, and 55 m.p.h., and the following six conditions in the vehicle: (1) baseline (windows up, fan, radio, and windshield wipers off), (2) driver's window down, (3) fan on, (4) FM radio playing music, (5) AM radio playing a talk show, (6) windshield wipers on (recorded at 0 m.p.h. only). The digit strings were read from a script with equal probabilities for all digits. The digit, '0', had two pronunciations, "zero" and "oh".

Speech was recorded on a DAT recorder in various automobiles using 2 microphones located on the driver's visor. The microphone used for our data was a high-fidelity Sony ECM-959DT, which uses an electret element and has a flat bandpass response over 50 - 18,000 Hz. The data were lowpass filtered to about 6,720 Hz before sampling at 16 kHz using the line inputs of an Ariel Digital Microphone.

Since the goal for collecting the database was to make it as realistic as possible, the recording conditions were somewhat variable and reflected what an untrained population of users might produce. Some of the files for various speakers were missing due to recording problems which were not noticed until the data were reviewed.

* Currently at Mitsubishi Heavy Industries, Ltd.

2.2. The Census Database

The census database was used to train the system because the Motorola automotive database had too few speech samples for training. The training component of the census database is composed of 74 speakers (53 males and 21 females), and the utterances consist of strings of letters, numbers, and a few control words. The training set consists of 1018 utterances, recorded using a Sennheiser HMD224 close-talking microphone in an office environment.

No attempt was made to optimize the SPHINX-I system for the 11-digit Motorola database. For example, the the census database has a larger vocabulary size than the automotive database, and performance could have been improved by recomputing the phonetic models for this far more constrained task.

3. NOISE CHARACTERISTICS

As noted above, we distinguish between the “running noise” caused by window, engine, and tire noise that primarily depends on vehicle speed and “functional noise” which depends on operator- controlled functions such as the heater fan, radio, and windshield wipers. Spectral analysis of the Motorola automotive database reveals a peak at very low frequencies in the idling condition. This peak becomes smaller and the spectrum of the noise broadens as vehicle speed is increased. With the windows down, wind noise becomes more intense and the distribution of the noise power becomes broader, as shown in Figure 1.

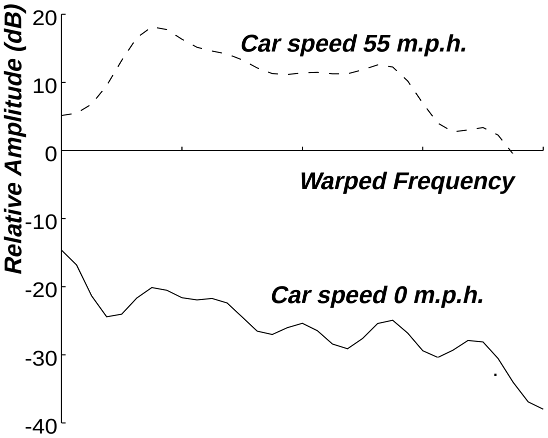

Figure 1. Typical noise spectra in the “windows down” condition at 0 and 55 m.p.h. The horizontal axis represents frequency after the nonlinear warping used by SPHINX.

The car radio causes significant interference to speech recognition. It is not very clear how the spectral shape between music and talk shows differs, but some transient patterns appear in the noise region in both cases. The heater fan produces broadband noise in the region of 2 – 7 kHz. This characteristic is similar to the wind noise observed at high vehicle speeds. Windshield wipers are intermittent noise sources which generate transient patterns. Other functional components of the vehicle such as the horn and turn signals also produce transient noise.

4. EXPERIMENTAL RESULTS

We performed speech recognition experiments with the SPHINX recognition system using the census database for training and the Motorola automotive database for testing. Speech was sampled at 16 kHz, and conventional Mel-frequency cepstral coefficients (MFCC) were used as the baseline parametric representation of speech frames [9]. We made use of previously-trained phonetic models for SPHINX-I which consisted of 400 generalized triphone models for a vocabulary size of 104. The language model and pronunciation dictionary were restricted to the 11 words (“one” through “nine”, “oh” and “zero”) that are present in the 7-digit strings in the Motorola automotive database.

We compared recognition accuracy obtained using three types of signal processing: the baseline MFCC representation, cepstral mean normalization (CMN) [*e.g.* 6], and codeword-dependent cepstral normalization (CDCN) [7]. The simple CMN method compensates primarily for differences in the frequency response of each channel, while CDCN compensates simultaneously for the effects of linear filtering and additive noise. The main source of variation in the car environment is additive noise. However, there are also differences in channel frequency response between the training and testing data, since different databases with different microphones were used for training and testing. It was also hoped that CMN would eliminate some of the effects of speaker variability due to differences in vocal tract.

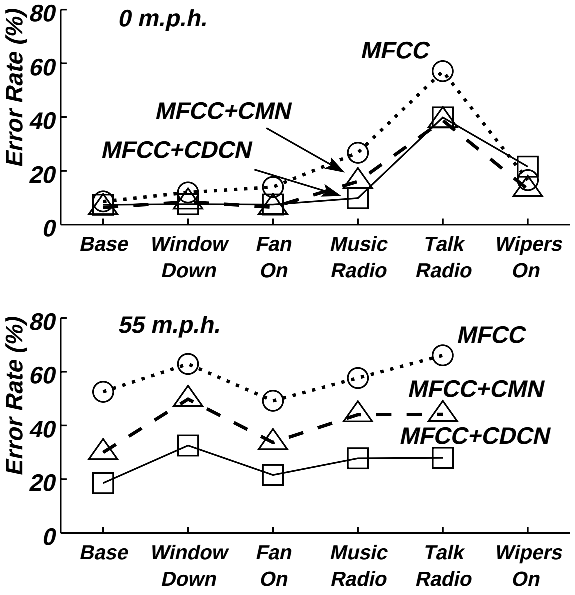

Figure 2. Digit error rates obtained using the Motorola automotive database for two car speeds: 0 m.p.h. (upper panel) and 55 m.p.h. (lower panel). Three different compensation conditions were used: MFCC (circles), CMN (triangles), and CDCN (squares). Error rates are plotted for each of six different functional configurations of the automobile (see text.)

Figure 2 summarizes the digit error rates obtained at two different vehicle speeds, comparing the three types of signal processing for the six conditions of the Motorola database described in Sec. 2.1. Results obtained at 30 m.p.h. almost always produced error rates that fell between the rates observed at 0 and 55 m.p.h. Recordings with the windshield wipers on were not obtained at 30 or 55 m.p.h.

The results of Fig. 2 indicate that with the vehicle stopped with the engine idling, recognition accuracy is degraded primarily by functional noise sources from the automobile such as the heating fan, windshield wipers, and the car radio. Recognition accuracy is especially poor with the radio playing talk shows because the speech from radio produces many insertion errors. The intermittent nature of the noise caused by the windshield wipers also contributes to insertion errors. As the vehicle speed is increased from 0 to 55 m.p.h., recognition accuracy in all conditions worsens because of masking introduced by the running noise from turbulent air, tires, and the engine. The talk-show signal from the radio does not introduce as much degradation in recognition accuracy at higher vehicle speeds as it had at lower speeds due to masking of the radio signal by these running noise sources. The effect of the fan is relatively small compared to the other conditions.

It is also seen in Fig. 2 that both CMN and CDCN can substantially improve recognition accuracy. At 0 m.p.h. the recognition accuracy obtained with both CMN and CDCN is similar, suggesting that at this speed the algorithms are primarily compensating for differences in spectral shape between the training and testing conditions (despite the fact that the sources of degradation are additive in nature). It is clear that the presence of a talk radio signal remains a serious problem at low speeds, even with CMN or CDCN. At 55 m.p.h. the effects of additive noise become far more important, and the recognition accuracy obtained with CDCN is clearly superior to results obtained using CMN.

Because both CMN and CDCN can improve recognition performance, we also measured recognition accuracy with the two methods combined. Unfortunately, recognition performance obtained using a combination of the two approaches was no better than that obtained with CDCN alone.

5. ADAPTIVE NOISE CANCELLATION OF SIGNALS FROM THE CAR RADIO

As noted in the previous section, neither CMN nor CDCN can compensate completely for the effects of interference by AM-radio talk shows. Fortunately, the radio signal to the loudspeakers is electrical, and it can be monitored directly at the loudspeaker output as a noise source, to be used as the reference channel for adaptive noise cancellation. We describe in this section the results of a pilot study to confirm the utility of adaptive noise cancellation of radio signals.

Since the Motorola automotive database does not have a separate channel with the interfering signal from the radio, we collected a new database using two simultaneously-recorded signals. One channel contains speech corrupted by the car radio recorded at three running speeds (0 m.p.h. [idling], 30 m.p.h., and 55 m.p.h.). The other channel contains the input to the loudspeaker for the car radio. The speech contains 7-digit strings as did the Motorola automotive database. The radio was tuned to talk stations on the AM band, which were recorded monophonically.

We applied the LMS (*Least-Mean-Square*) algorithm (e.g. [10]) to cancel the car radio signal. 150 taps were used for the FIR filter, and based on informal listening the value of the step-size parameter was set to 0.05 times the theoretical upper bound for stability which equals the product of the number of taps and the average signal power.

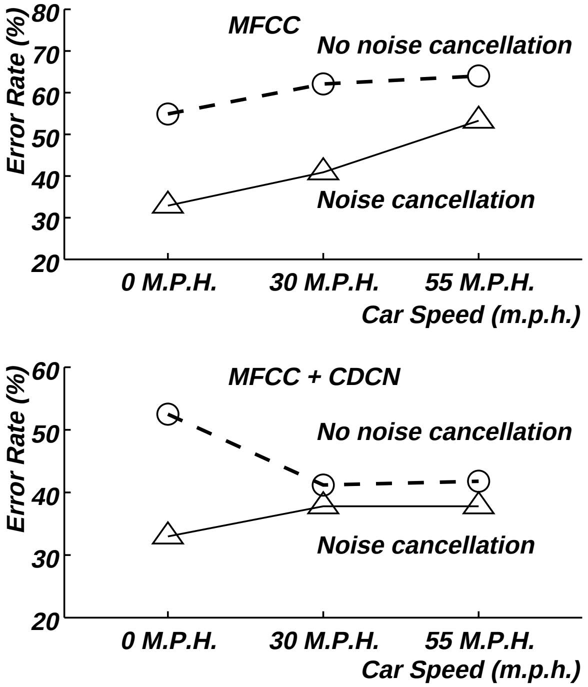

Figure 3. Digit error rates obtained with and without adaptive noise cancellation of AM-radio talk show signals using the LMS algorithm. Data were recorded at CMU using recording conditions similar to those of the Motorola automotive database. Results are shown without CDCN (upper panel) and with CDCN (lower panel).

We also found that stability of the cancelled signal would be improved by the addition of a gain-reduction factor to the weight vector that reduces all components by a factor of 0.1, *i.e.* $W'_{k+1} = 0.1 \times W_{k+1}$ where $\mathbf{W}'_{k+1}$ is the weight vector after gain reduction. Figure 3 summarizes error rates observed at 3 vehicle speeds, with and without CDCN. Adaptive noise cancellation clearly decreases the recognition error rate, especially at low vehicle speeds. A detailed analysis of the results confirmed that insertion errors are reduced dramatically by adaptive noise cancellation.

The CDCN algorithm provides further reductions in recognition error at high vehicle speeds. The use of CDCN, which improves recognition accuracy at high vehicle speeds (when interference is dominated by running noise), complements the use of adaptive noise cancellation, which is most helpful at low vehicle speeds (when interference is dominated by the signal from the radio).

In our experiments, the LMS adaptation continued while speech was input to the system. This produced transients in the weight vectors which, according to informal listening, introduced a needless source of distortion to the compensated signals. We believe that recognition accuracy would be further improved if adaptation were disabled during speech input, although this result has not yet been verified.

6. CONCLUSIONS

We describe the effects of noise sources in the automobile on the recognition accuracy of the SPHINX-I speech recognition system. Significant degradations in recognition accuracy are observed while the automobile is at high speeds caused by quasi-stationary "running noise". In addition, at low speeds recognition accuracy is degraded by transient "functional noise" sources including windshield wipers and the car radio, especially when talk shows are broadcast.

We also compared results obtained using two environmental compensation algorithms, CMN and CDCN. CMN improves recognition accuracy for all conditions, even though it was only expected to reduce channel effects. CDCN provides a further improvement in recognition accuracy under high-noise conditions. The combination of CMN and CDCN did not provide further improvements in recognition accuracy beyond what had been obtained with CDCN.

We also considered the use of adaptive noise cancellation using the extremely simple LMS algorithm to eliminate interference from the radio. As expected, this approach provided significant further reductions in errors caused by radio signals, especially at low speeds. CDCN and adaptive noise cancellation provide complementary benefits.

ACKNOWLEDGEMENTS

This research was sponsored by the Department of the Navy, Naval Research Laboratory, under Grant No. N00014-93-2005. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government. The authors would like to thank Fu-Hua Liu, Pedro Moreno, Yoshiaki Ohshima, Raj Reddy, Tom Sullivan, Tomio Takara, and the rest of the CMU speech group for their contributions to this work. We also thank the Motorola Corporation for providing the automotive database.

REFERENCES

1. Dal Degan, N., and Prati, C., "Acoustic Noise Analysis and Speech Enhancement Techniques for Mobile Radio Applications", *Signal Processing,* **15:** 43-56, 1988.

2. Lockwood, P., Baillargeat, C., Gillot, J.M., Boudy, J.,and Faucon, G., "Noise Reduction for Speech Enhancement In Cars: Non-linear Spectral Subtraction/Kalman Filtering", *EUROSPEECH-91*, **1:** 83-6, 1991.

3. Mokbel, C., and Chollet, G., "Word Recognition in the Car: Speech enhancement / Spectral Transformations", *ICASSP-91*, pp. 925-928, 1991.

4. Oh, S., Viswanathan, V., and Papamichalis, P., "Hands-Free Voice Communication in an Automobile With a Microphone Array", *ICASSP-92*, 1992, pp. I-281-I-284.

5. Gales, M. J. F., and Young, S., "An Improved Approach to the Hidden Markov Model Decomposition of Speech and Noise", *ICASSP-92*, pp. I-233–I-236, 1992.

6. Liu, F.H., Acero, A., and Stern, R.M., "Efficient Joint Compensation of Speech for the Effects of Additive Noise and Linear Filtering", *ICASSP-92*, pages 865-868, March 1992.

7. Acero, A., and Stern, R. M., "Environmental Robustness in Automatic Speech Recognition", *ICASSP-90,* pp. 849-852, 1990.

8. Lee, K.-F., Hon, H.-W., and Reddy, D. R., "An Overview of the SPHINX Speech Recognition System", *IEEE Trans. Acoust. Speech Signal Process.,* **38:** 35-45, 1990.

9. Davis, S. B., and Mermelstein, P., "Comparison of Parametric Representations of Monosyllabic Word Recognition in Continuously Spoken Sentences", *IEEE Trans. Acoust. Speech Signal Process.,* **28:** 357-366, 1980.

10. Widrow, B., and Stearns, S. D., *Adaptive Signal Processing*. Prentice-Hall, Englewood Cliffs, 1985.