ENHANCING AUTOMATIC SPEECH RECOGNITION WITH AN ULTRASONIC LIP MOTION DETECTOR

David L. Jennings and Dennis W. Ruck

Department of Electrical and Computer Engineering,
Air Force Institute of Technology,
Wright-Patterson AFB, OH, USA 45433-7765
djenning@afit.af.mil druck@afit.af.mil

ABSTRACT

This paper presents the results of experimentation with a simple ultrasonic lip motion detector or "Ultrasonic Mike" in automatic speech recognition. The device is tested in a speaker dependent isolated word recognition task with a vocabulary consisting of the spoken digits from zero to nine. The "Ultrasonic Mike" is used as input to an automatic lip reader. The automatic lip reader uses template matching and dynamic time warping to determine the best candidate for a given test utterance. The device is first tested as a stand alone automatic lip reader achieving accuracy as high as 89%. Next the automatic lip reader is combined with a conventional automatic speech recognizer. Classifier fusion is based on a pseudo probability mass function derived from the dynamic time warping distances. The combined system is tested with various levels of acoustic noise added. In a typical example, at 0dB, the acoustic recognizer's accuracy was 78%, the lip reader accuracy was at 69%, but the combined accuracy was 93%. This experiment demonstrates that this simple ultrasonic lip motion detector, that has an output data rate 12,500 times less than a typical video camera, can improve automatic speech recognition in noisy environments. This experiment also demonstrates an effective classifier fusion algorithm based on dynamic time warping distances.

1. INTRODUCTION

Many potential applications in automatic speech recognition require a system that can operate in high or varied noise situations, such as the cockpit of an aircraft or a noisy office. Unfortunately, the accuracy of current automatic speech recognition systems falls off drastically as the training noise environment differs from the actual environment. In the past many attempts have been made to reduce the effects of noise [1]. One way to reduce the effects of noise is to add an estimate of the noise environment to the clean templates recorded during training. Using these noisy templates increases accuracy in noise, but the accuracy still falls to unacceptable levels as the signal to noise approaches zero. Another way to deal with noise is to use knowledge sources such as syntax and context. One lesser investigated knowledge source that humans use, if available, is lip reading. In a study done with partially deaf people, Walden [2] demonstrates the benefit humans derive from lip reading. In a consonant recognition task with audio only the average accuracy was 48% and with visual only, around 45%, but with both audio and visual the subjects achieved 85% accuracy. Clearly, including lip information in an automatic speech recognition system could enhance accuracy. The human ability to lip read has inspired a number of past attempts to include lip reading in automatic speech recognition.

2. BACKGROUND

Most of the past efforts to include lip information in an automatic speech recognition system have relied on video images of the mouth. In 1984, Petajan [3], using video images, showed improved results in automatic speech recognition. Since then, researchers have tried many pattern recognition techniques including hidden Markov models and neural networks [4, 5, 6, 7, 8] to incorporate lip information into an automatic speech recognition system. Another approach, in acquiring the lip information, concentrates less on static images of the face and more on motion. In 1991, Mase and Pentland [9] demonstrated that good results could be obtained by using optical flow analysis of the video images. All of these attempts have demonstrated improved results by including the additional information from the lips, but at a cost. A typical video camera outputs five megabytes of information per second, which

0-7803-2431-5/95 $4.00 © 1995 IEEE

represents a heavy computational burden for the recognizer. This paper will investigate the use of an "Ultrasonic Mike" (UM) in acquiring the lip information. The output rate of the UM is 400 bytes per second, which is 12,500 times less than a typical video camera. All of the past efforts have also demonstrated difficulty in fusing the lip information in the recognition system. In most of the past work fusion was based on heuristic rules. This paper will explore an analytic approach to fusion based on a pseudo probability mass function derived from the dynamic time warping distances.

3. EXPERIMENTAL METHODS

Figure 1 gives a pictorial representation of how the UM works. The "Ultrasonic Mike" uses a piezoelectric material and a 40KHz oscillator to create a continuous wave ultrasonic signal. The ultrasonic signal is transmitted from a headset boom mount located in front of the speaker's face. The signal reflects off the speaker's mouth and returns to an ultrasonic receiver. Repeated reflections between the speaker's mouth and the device establish a standing wave. As the speaker's mouth moves, the standing wave changes, resulting in changes in the magnitude of the received signal. The device then processes the received signal with an envelope detector and AC coupling, resulting in a low frequency output signal that changes in response to motion of the speaker's mouth. Sampling at 16KHz, the acoustic signal is collected on one channel and the UM signal is simultaneously collected on the second channel. Figure 2 gives a block diagram of the overall combined system. After sampling, the acoustic signal is sent through a dynamic time warping based classifier using LPC analysis, resulting in 10 LPC cepstral coefficients per step. Analysis of the UM signal indicated no significant frequency components above 50Hz, therefore the signal could be significantly down sampled. To coincide with the number of windows analyzed by the acoustic system, the UM signal was down sampled by 80, resulting in an effective sampling rate of 200Hz. Both the acoustic and ultrasonic classifiers are based on dynamic time warping distance between the test utterance and training templates. Each classifier outputs a ranked list of the candidates and their DTW distances. Classifier fusion is accomplished by converting the ranked distances to a pseudo probability mass function. One way to convert the distances to probabilities is to let the probability for each class, k, be defined as

$$p_k = \frac{\frac{1}{d_k}}{\sum_{i=1}^{M} \frac{1}{d_i}}, \qquad (1)$$

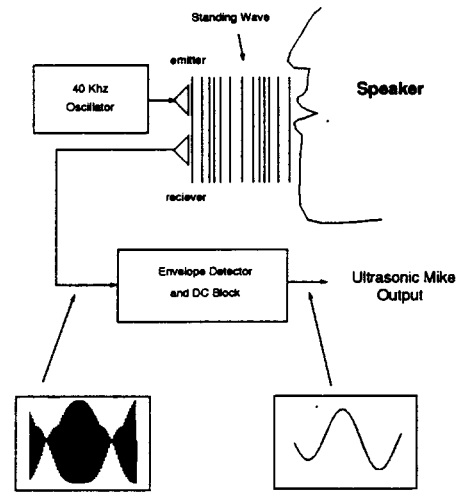

Figure 1: Operation of the Ultrasonic

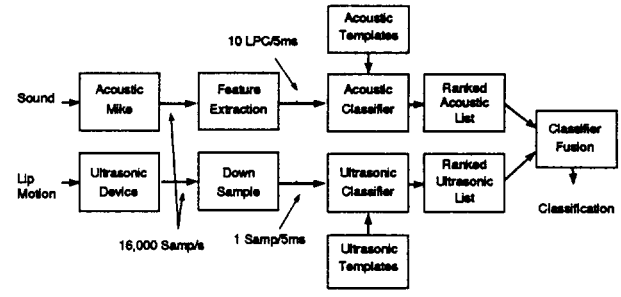

Figure 2: Block Diagram of Combined System

where d_k is the distance between the class k template and the test utterance and M is the total number of classes [10]. The resulting pseudo probability mass functions are combined into one, using a scale factor to compensate for the different reliabilities of the two classifiers. The combined probability mass function, P, is

$$P_k = \lambda * p_{k,ac} + (1 - \lambda) * p_{k,um}, \qquad (2)$$

where $p_{k,ac}$ and $p_{k,um}$ are the probability mass functions for the acoustic and ultrasonic classifiers and λ is the scale factor.

4. RESULTS

The following results are based on speaker dependent tests in a isolated digit recognition task. In each test 100 test utterances were used and up to four templates per class. In the following sections results of the automatic speech recognizer alone, the automatic lip reader alone, and a combined recognizer are presented.

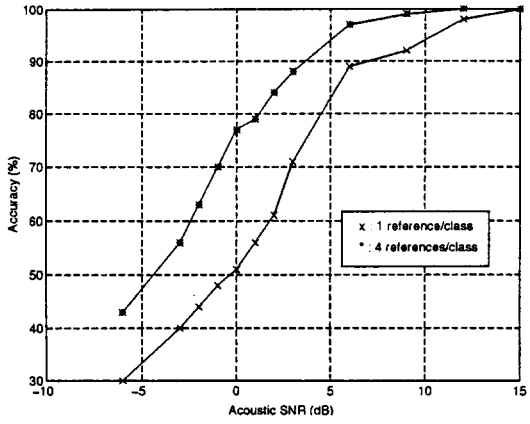

Figure 3: Typical results of the acoustic classifier alone. *: 4 reference templates per class,x: 1 reference Template per class.

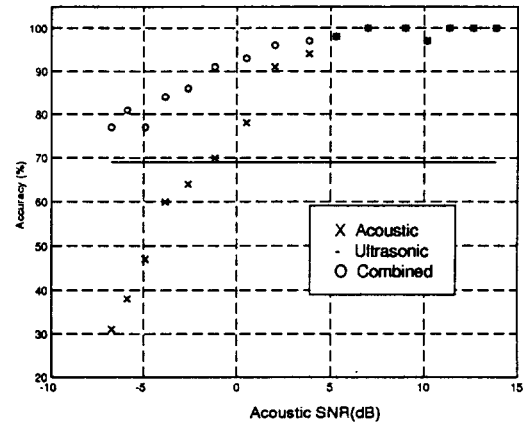

Figure 4: Results using 4 acoustic templates and 1 UM template

Table 1: Average automatic lip reader accuracies over 5 tests in percentages within one session. The Ultrasonic recordings were made sequentially.

Number of Templates	Ultrasonic Mike
1	75.2
2	80.2
3	86.2
4	89.2

4.1. Automatic Speech Recognizer

Figure 3 demonstrates the typical results of the automatic speech recognizer, with various level of white Gaussian noise added. For these results, the same level of white noise is added to the templates and the test utterances. As can be seen from the figure increasing the number of templates increases the accuracy, but it also increases the computations required for classification.

4.2. Automatic Lip Reader

The initial tests with the UM demonstrated the device was sensitive to position, resulting in an average cross session accuracy of 12.6%. To reduce the effect of movements, recordings were made in sequences such as (0,0,0,...,1,1,...). Recording in this manner reduced the chances of moving the device between utterances. The average accuracy of the automatic lip reader alone is presented in Table 1. Again increasing the number of reference templates improves the overall accuracy.

4.3. Combined Recognizer

Figures 4 and 5 summarize the results of this experiment. Two different runs of the experiment are presented. The first run uses a single template per class and the second run is based on four templates per class. From the first experiment, the flat line at 69% is the accuracy achieved for the lip reader alone using the first utterance as the reference template. Depending on the reference template selected, single template results varied from 65-82%. Various levels of white Gaussian noise were added to the acoustic utterances and the same level of noise was added to the acoustic templates. As the SNR fell below 0dB the accuracy of the acoustic system drops off rapidly. Note that the combined results are always equal to or better than the best of the individual classifiers. The best improvement was at 0dB, the acoustic recognizer's accuracy was 78%, the lip reader accuracy was at 69%, but the combined accuracy was 93%. The experiment using four ultrasonic templates resulted in better lip reader performance, 85%, and the combined performance at 0dB was 90%, approximately the same as before.

5. CONCLUSIONS

This experiment clearly demonstrates that a simple ultrasonic lip motion detector can provide some lip information without the computational burden of video images. The results also demonstrate an effective way to fuse the ultrasonic classifier with an acoustic classifier, achieving improved performance in noisy situations. Even though the "Ultrasonic Mike" offers a tremendous computational savings over video it has the

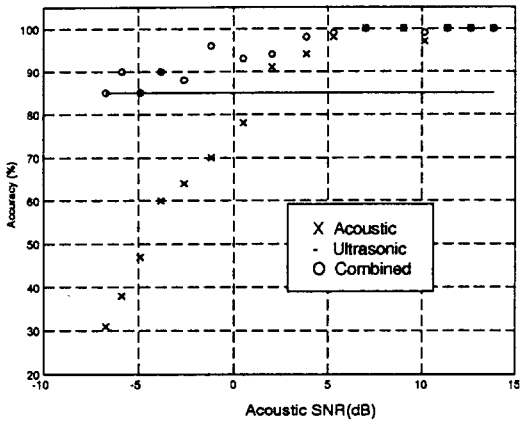

Figure 5: Results using 4 acoustic templates and 4 UM templates

drawback of position sensitivity. This sensitivity can be somewhat mitigated by using a helmet mount, but a more position insensitive design is needed for practical implementation. Ideally a simple device needs to be found that can extract the most important features in lip reading, which based on past studies [11, 12] are vertical lip separation, puckeredness, and horizontal extension of the lips. Clearly automatic lip reading could make applications, in varied and high noise environments, possible.

6. ACKNOWLEDGEMENTS

I would like to sincerely thank Paul Morton and John Schnurer of Armstrong Lab at Wright-Patterson AFB. Without their support and design of the "Ultrasonic Mike" this research would not have been possible.

7. REFERENCES

[1] Lawrence Rabiner and Biing-Hwang Juang. *Fundamentals of Speech Recognition.* Signal Processing Series. Prentice Hall, Englewood Cliffs, New Jersey, 1993.

[2] Brian E. Walden, Debra A. Busacco, and Allen A. Montgomery. Benefit from visual cues in auditory-visual speech recognition by middle-aged and elderly persons. *Journal of Speech and Hearing Research*, pages 431–436, April 1993.

[3] Eric D. Petajan. Automatic lipreading to enhance speech recognition. *Proceedings IEEE Global Telecommunications Conference*, pages 265–272, 1984.

[4] Eric D. Petajan. An improved automatic lipreading system to enhance speech recognition. *Association for Computing Machinery Special Interest Group on Computer and Human Interaction*, pages 19–25, 1988.

[5] Terrence J. Sejnowski and Moise Goldstein. Massively parallel network architectures for automatic recognition of visual speech signals. Final Technical Report DTIC AD-A226 968, The Johns Hopkins University, Baltimore, MD, 1990.

[6] T.J. Sejnowski, B.P. Yuhas, and R.E. Jenkins. Combining visual and acoustic speech signals with a neural network improves intelligibility. In D. Touretzky, editor, *Advances in Neural Information Processing Systems, 2*, pages 232–239. Morgan Kaufmann, San Mateo, California, 1990.

[7] David G. Stork, Greg Wolff, and Earl Levine. Neural network lipreading system for improved speech recognition. *IEEE International Joint Conference on Neural Networks*, 2:285–295, 1992.

[8] Ben P. Yuhas, Jr. Moise H. Goldstein, and Terrence J. Sejnowski. Integration of acoustic and visual speech signals using neural networks. *IEEE Communications Magazine*, pages 65–71, 1989.

[9] Kenji Mase and Alex Pentland. Automatic lipreading by optical-flow analysis. Technical Report 117, MIT Media Lab, Perceptual Computing Group, Cambridge, MA, 1991.

[10] Lei Xu, Adam Kryzak, and Ching Y Suen. Methods of combining multiple classifiers and their applications to handwriting recognition. *IEEE Transactions on Systems, Man, and Cybernetics*, 22:418–435, May 1992.

[11] Kenneth W. Berger. *Speechreading: Principles and Methods.* Herald Publishing House, second edition, 1978.

[12] Pamela L. Jackson, Allen A. Montgomery, and Carl A Binnie. Perceptual dimemsions underlying vowel lipreading performance. *Journal of Speech and Hearing Research*, 19:796–811, 1976.