EXPERIMENTS IN MULTI-MICROPHONE SPEECH ENHANCEMENT FOR RECOGNITION

Michael. S. Scordilis† and Stuart Adams‡

† Department of Electrical and Electronic Engineering
The University of Melbourne

‡ Sun Microsystems Laboratories

ABSTRACT - Operating on high quality speech data is important in present speech recognition systems, and any variation has quite detrimental effects on system recognition performance. Head-mounted, close-talk microphones are usually required, but are too restraining. As an alternative, this paper investigates speech processing for several arbitrarily positioned microphones, and evaluates the performance of a given recognition system for a number of different arrangements. Results indicate that while beamforming improves performance, it does not match the performance of more traditional close-talk setups unless additional signal enhancement is performed.

INTRODUCTION

Speech recognition technology has made considerable progress over the past few years. This has become possible with the introduction of new algorithms and through the development of faster digital signal processors and economical, high performance computer systems.

The current state-of-the-art in speech recognition stands at very high accuracy in continuously spoken, small size word vocabularies, both in speaker-dependent and independent applications; very high accuracy in isolated-word, large vocabulary, speaker-dependent applications; high accuracy in medium size vocabularies, continuous speech applications.

Such performance is achieved on high resolution data of large bandwidth speech, operating in low noise conditions. However, system performance degrades usually rapidly as signal quality deteriorates. In the case of the IBM Tangora, which is a 20,000 isolated-word, speaker-dependent system, word error rates varied from 0.8% when training and testing on "quiet" speech, to more than 50% when training on "quiet" speech and testing on "noisy" speech, or when training on "noisy" speech and testing on "quiet" speech (Das et al, 1993). These results were collected with a high quality, head-mounted, unidirectional microphone. Such studies indicate that any change to the quality of the testing data, either for the better or for the worse in relation to the training data, results in severe degradation of system performance.

A highly desired feature for a functional and user-friendly speech recognition system is that performance comparable to head-mounted microphones be achieved even for stand-mounted or clip-on microphones. This is typically not the case with present systems where the best results are provided with uncomfortable and constraining head-mounted microphones. It is therefore evident that when compared to the remarkable human abilities of acoustic processing of speech in diverse listening environments even at considerable distance from the talker, the performance of current speech recognition systems is far from satisfactory, limiting the application of large vocabulary, continuous speech recognition systems to demonstration purposes and their development only as prototypes in laboratory-based studies. All these factors have limited the robustness and therefore the level of acceptance of speech recognisers.

This paper addresses the problem of enhancing the recognition performance of an available speech recognition system, BBN's HARK, when up to four microphones are used to collect data. The operating environment was a fairly typical office, with hard walls, and the tests were conducted on a high performance workstation, as shown in Figure 1. The enhancement method used was a classical delay-and-add beamforming (Flanagan et al, 1991), with inputs consisting of one, two or four channels.

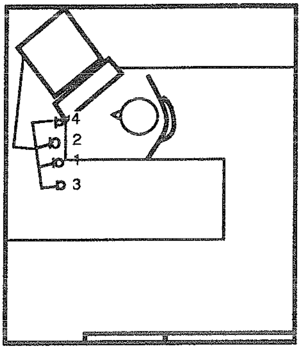

Figure 1. Experimental Setup

SPEECH ENHANCEMENT FOR RECOGNITION

All undesirable signal components are classed as noise. In the speech recognition task, any factor contributing to a decline in system performance may be classified as introducing noise into the process. Signal quality and recognition performance are usually positively correlated, except in the case where a system was trained with noisy speech and tested on clean speech resulting in worse performance. In a typical, single-user, quiet office working environment there are two main noise-contributing factors: additive background noise and reverbation. In the general case, background noise can be viewed as statistically independent from the speech signal and as occupying potentially the same spectrum as the signal. In such a case, linear filtering is usually not effective in removing it. On the other hand, reverbation is the sum of scaled and delayed replicas of the signal, and as such, it is correlated with the produced speech. The total power of the reverbant signals depends on the geometry and the type of material used in the operating enclosure. The accumulative effect of reverbation is non-linear spectral distortion on the original speech signal. Reverbation conveys important spatial information and it is often desirable and useful to human listeners. However, by colouring the speech spectrum it contributes to lowering the performance of speech recognisers. Other typical sources of signal degradation and subsequent increase in recognition error rate are the contribution of speech from non-relevant talkers and the variability in microphone properties. In the case of recognition over the telephone factors reducing recognition performance are the presence and variability in degradation introduced by the channel, such as bandwidth limitation and phase jitter, line echoes and noise introduced by the coding scheme in digital lines.

There exist two general approaches dealing with the presence of noise which may be either correlated or uncorrelated with the original speech signal. The first approach focuses on minimising noise effects implicitly in the design of the recognition system. One method used, concerns the development of techniques for front-end processing, which yield robust features for recognition under acoustically adverse conditions. Although the methods for extracting such features are generally independent on the noise properties, and therefore not optimal, they nevertheless improve overall recognition. Some of the most effective features are short-term modified coherence representation, SMC (Mansur & Huang, 1988), and relative spectral processing RASTA (Hermansky et al, 1991). The other method concerns with transforming the models developed during training on clean speech, so as to better account for the effects of noise present during recognition. Such techniques are mainly applied to HMM-based recognisers and are used to alter the parameters of the models when training and testing signal qualities are detected to be sufficiently different (Gong & Treurniet, 1993).

The second approach , usually termed as speech enhancement, aims at processing the speech signal explicitly for the purpose of improving quality, expressed as signal-to-noise ratio, or intelligibility or lowering listener's fatigue. These techniques are not restricted to speech recognition but apply to all applications concerned with the collection of high performance data remotely, at some distance from the source or generally when signal quality is poor due to adverse acoustic conditions. Methods in this category are divided into two groups: single-channel and multi-channel, depending on the number of microphones used for data collection.

In single-channel enhancement, a single microphone is used for data collection and noise properties are estimated during non-speech intervals. Three general methods exist for the suppression of additive noise: spectral subtraction, optimum Wiener filtering (Deller et al, 1993) and acoustic echo cancellation (Murano et al, 1990). In the first method, long-term spectral averages of the noise estimated during silence are subtracted from short-term spectra during speech activity. A number of problems result from this method, one of them being the generation of negative spectral components and subsequent distortion introduced by forcing such components to zero. The second method is based on statistical noise estimates during silence and on underlying assumptions about the speech process itself (all-pole, etc). The third method is based on estimating the room acoustic impulse response and uses it to construct an FIR filter to remove the effects of reverbation. The major limitation of this method is that it needs to compute the output a filter of the order of 4,000 taps in real-time. Current efforts are concentrated on the use of IIR filters, which would provide substantial computational savings (Chao & Tsujii, 1991).

In multi-channel enhancement, two methods have been applied. The first, using two microphones, is based on adequate physical separation between each sensor, so as to ensure that the primary sensor provides speech corrupted by additive noise, while the other sensor provides just noise which is highly correlated with the noise in the primary sensor. These estimates are used to suppress long-term as well as short-term noise phenomena (Deller et al, 1993). However, the requirements for setting up such systems are excessively constraining and reduce their value to controlled experimental studies. In the second method, multi-microphone arrangements are used to enhance the signal through beamforming and since it presents a number advantages, it was also the choice for the work reported here.

The experimental setup used for this work came about in an effort to overcome a number of the limitations imposed by the techniques highlighted previously, and to evaluate a practical and less constraining senario for data collection. Sennheiser HMD-414 microphones were used on a Sun Sparc 10 workstation with 4-channel audio capabilities as shown on Figure 1.

MULTI-MICROPHONE SPEECH RECOGNITION

Typically, commercially available speech recognisers are provided as plug-and-play systems, where the user has no access to the various system modules, and in the case of systems built to operate as truly speaker independent (such as BBN's HARK) no speaker adaptation phase is provided. In such case, where the system is used more or less as a "black box", the only preprocessing the user can perform in an effort to enhance recognition performance is speech enhancement. The requirements employed here were that a flexible microphone array with up to four elements be used for data collection, based on the principle of delay-and-add beamforming for signal enhancement. If the number of microphones is I and the produced speech signal is s(t), then the sum of all I channels, $s_i(t)$ is:

$$s_T(t) = \sum_{i=1}^{I} s_i(t) = \sum_{i=1}^{I} s(t - \tau_i) = \sum_{i=1}^{I} s(t - \frac{d_i}{c}) \ ,$$

where τ_i is the time required for the signal to reach microphone i, d_i is the distance of microphone i from the talker and c is the speech of sound in air. In discrete-time, the signal resulting from adding all contributing channels is:

$$s_T(n) = \sum_{i=1}^{I} s_i(n) = \sum_{i=1}^{I} s(n - m_i) = \sum_{i=1}^{I} s(n - \frac{d_i}{c}) ,$$

In this case, the ratio d_i/c was rounded to the nearest integer, m_i, to give delays in integer multiples of the sampling period.

Computational techniques

To avoid high amplitude transient acoustic noise, such as clicks and pops occuring before or after the spoken utterance, simple but conservative, energy-based endpoint detection was performed on each channel. Following that, and in order to reduce computation, processing was focused on the region around point M, the location of the frame where signal energy peaked. This was typically the case of some stressed vowel sound. In stereo setups the system provided for the two channels to be recorded synchronously. To reduce the search, and to avoid misalignment in the event of a sustained vowel sound with very similar looking periods, the microphones were located so that their relative delays $\Delta d = d_i - d_j$, were less than half a pitch period. Assuming a possible maximum pitch period of 400 Hz results in a relative shift of 1.25 ms. For the 16 kHz sampling rate used, this corresponded to maximum relative shifts of 20 sample periods. If the speech of sound in air is $c = 340$ m/s, the maximum allowed difference in the distance from the talker to each microphone is 0.425 m. Beamforming was achieved by computing the relative delay between every channel pair, aligning by shifting one channel per pair by its computed relative delay, and summing the two signals. The computed delay equals the shift that maximises the normalised cross-correlation, $R_{1,2}$, defined as

$$R_{1,2}(d) = \frac{1}{N}\left(\sum_{k=M-N+1}^{M} s_1(n)s_2(n-d)\right) ,$$

over N=300 points of each channel pair. The beamforming algorithm is summarised in Figure 2.

In the case of four-channel sound, the system could not provide synchronisation between all channels, but only pair-wise. After the channel delays were computed as described, the overall delay was derived by searching over a much larger portion of the signals. The computed delay compared very well with the differences in the locations given by the endpoint detection algorithm. As a result, after the ability of the algorithm to cope with large temporal differences introduced by the lack of full of channel synchronisation was established, all spatial restriction on the microphones were lifted, thus allowing for completely arbitrary sensor positioning.

Recognition task and performance for different microphone setups

The recognition task in these experiments consisted of spoken inquiries of calendar entries. The recognition engine was BBN's HARK, a large vocabulary, speaker-independent, continuous speech recognition system. The language and general system interfaces were designed by Sun Microsystems Laboratories. Twelve sentences were used for recognition. A head-mounted Sennheiser HMD-414 microphone was used to set the reference recognition performance of the system. In that case, the achieved word recognition was 60.4%, with substitution, deletion and insertion rates of 34.1%, 5.5% and 2.2% respectively.

For multi-microphone speech, results indicated that recognition performance was dependent to a large degree on the location of the microphones with respect to the fan of the workstation. Also, the processing of the signals in pairs often tended to favour one pair over another depending on positioning, as described. The best recognition performance was achieved with the arrangement shown on Figure 1. The average talker-to-microphone distance was 1.2 m. The word recognition performance for a single microphone was 18.7%, while for the two-microphone setup it was 30.8% representing a 64.7% improvement. The addition of two extra microphones in these experiments did not provide any additional benefit. The recognition performance is summarised on Figure 3, showing one, two and four microphone setups, plus the results for the head-mounted, close-talk mic.

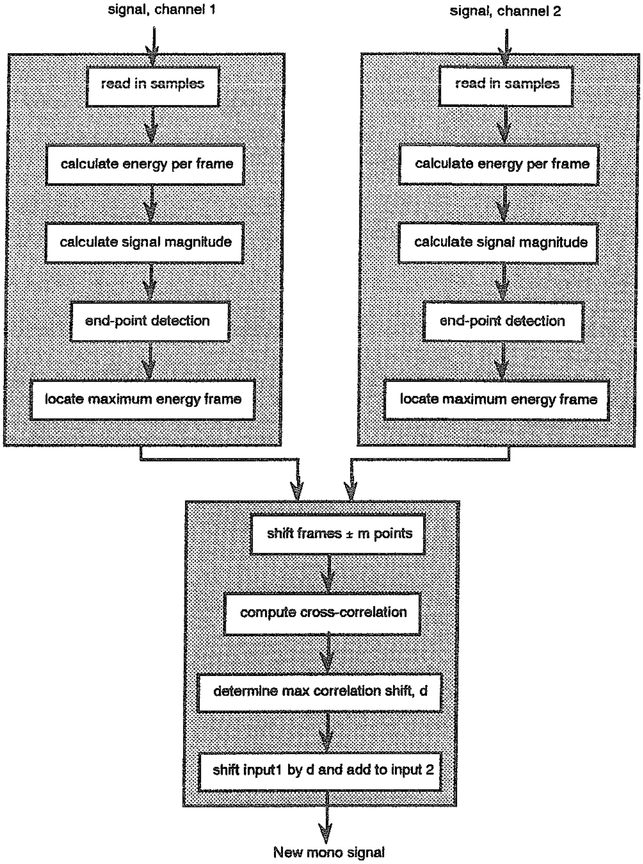
signal, channel 1
signal, channel 2
read in samples
calculate energy per frame
calculate signal magnitude
end-point detection
locate maximum energy frame
read in samples
calculate energy per frame
calculate signal magnitude
end-point detection
locate maximum energy frame
shift frames ± m points
compute cross-correlation
determine max correlation shift, d
shift input1 by d and add to input 2
New mono signal

Figure 2. The delay-and-add correlation algorithm used for beamforming

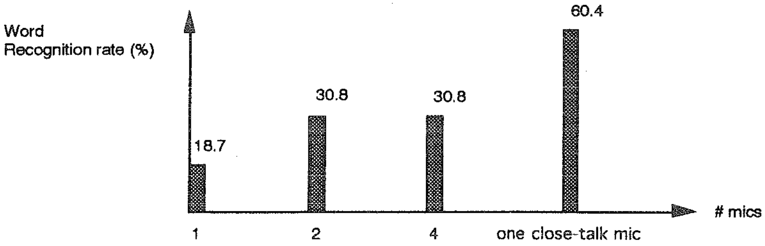

Figure 3. Summary of recognition performance

DISCUSSION AND CONCLUSION

The motivation for this series of experiments was to investigate the feasibility of abandoning the requirement for head-mounted microphone data collection and replacing it with a less restraining, flexible multi-microphone setup. The processing described in this paper, is based on delay-and-sum beamforming techniques which aim at enhancing similar signals while suppressing dissimilar signal components. The attractive features of the presented technique are its low computational requirements and the ability to process inputs from sensors located arbitrarily in space. There are two main limitations of this technique, first the sensitivity of the results to "bad locations", which are close to main noise sources, and the assumption that noise components are uncorrelated with the main signal, which is not the case when significant reverbant components are present.

The processing method presented here indicates that arbitrary microphone setups can be easily handled and shows promise in allowing for a more user-friendly speech-recognition environment. When coupled with a single-channel enhancement technique and with filtering to remove acoustic echoes, it could present an attractive and effective alternative to present data collection senarios for speech recognition.

ACKNOWLEDGMENT

This work was conducted at Sun Microsystems Laboratories, Chelmsford, Massachusetts, and the authors' appreciation is extended for the opportunity provided to conduct this investigation.

REFERENCES

Chao, J. and Tsujii, S. (1991) *A stable and distortion-free echo and howling canceller,* IEEE Transactions on Signal Processing, Vol. 39, No. 8, August 1991, pp. 1812-1820.

Das, S., Bakis, R., Nadas, A., Nahamoo, D., Picheny, M. (1993) *Influence of background noise and microphone on the performance of the IBM Tangora speech recognition system,* Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, pp. II.71-II.74.

Deller, J.R., Proakis, J.G. and Hansen, J.H.L., (1993) *Discrete-time processing of speech signals,* (Macmillan, New York).

Flanagan, J.L., Berkley, D.A., Elko, G.W. and Sondhi, M.M. (1991) *Autodirective microphone systems,* ACUSTICA, Vol 73 (1991), pp. 58-71.

Gong, Y. and Treurniet, W.C., (1993) *Speech recognition in noisy environments: a survey.* Technical report, CRC-TN 93-002, Broadcast Technologies Research Branch, Communications Research Centre, Department of Communications, Ottawa, Canada.

Hermansky, H., Morgan, N., Bayya, A., and Kohn, P., (1991)*Compensation for the effect of communication channel in auditory-like analysis of speech (RASTA-PLP),* Proceedings of European Conference on Speech Technology, pp. 1367-1370.

Mansur, D. and Juang, B.H., (1988) *The short-time modified coherence representation and its application for noisy speech recognition.* Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 525-528.

Murano, K., Unagami, S. and Amano, F., (1990) *Echo Cancellation and Applications.* IEEE Communications Magazine, January 1990, pp. 49-55.