SPEECH RECOGNITION IN A REVERBERANT ENVIRONMENT USING MATCHED FILTER ARRAY (MFA) PROCESSING AND LINGUISTIC-TREE MAXIMUM LIKELIHOOD LINEAR REGRESSION (LT-MLLR) ADAPTATION

P Raghavan[†], RJ Renomeron[‡], C Che[*], D-S Yuk and JL Flanagan

CAIP Center, Rutgers University, Piscataway, NJ 08854

ABSTRACT

Performance of automatic speech recognition systems trained on close-talking data suffers when used in a distant-talking environment due to the mismatch in training and testing conditions. Microphone array sound capture can reduce some mismatch by removing ambient noise and reverberation but offers insufficient improvement in performance. However, using array signal capture in conjunction with Hidden Markov Model (HMM) adaptation on the clean-speech models can result in improved recognition accuracy. This paper describes an experiment in which the output of an 8-element microphone array system using MFA processing is used for speech recognition with LT-MLLR adaptation. The recognition is done in two passes. In the first pass, an HMM trained on clean data is used to recognize the speech. Using the results of this pass, the HMM model is adapted to the environment using the LT-MLLR algorithm. This adapted model, a product of MFA and LT-MLLR, results in improved recognition performance[1].

1. INTRODUCTION

Reverberation is a major cause of degradation in performance in distant-talking speech recognition. Microphone arrays have been used for quite sometime to mitigate this effect and to obtain better quality speech. Though the improvement in recognition accuracy using this speech is significant, the performance is still far from satisfactory.

It has been shown that additional HMM adaptation results in improved recognition accuracies [2, 1, 4]. In [2], Omologo et al used Time Delay Compensation (TDC) techniques to improve the quality of the speech signal while using an algorithm similar to Maximum Likelihood Linear Regression [5] called the Phone Dependent Linear Regression (PDLR) to adapt the models to the environment. The Maximum a posteriori (MAP) adaptation technique was used with array-beamforming by Omologo et al in [1] and by Silverman et al in [4]. The Speaker Independent (SI) model trained on clean speech was adapted to microphone array speech in supervised mode using MAP. The performance in [1] was seen to degrade heavily with increase in reverberation, despite adaptation.

This paper presents a techinique to combine Matched Filter Array Processing [8] and the Linguistic Tree MLLR [3] adaptation algorithm for improved recognition even at high reverberation levels [7]. This approach compensates for reverberation in cases with and without prior training data from the environment.

2. MICROPHONE ARRAY WITH MATCHED FILTER

Sound captured through a microphone, in any environment, can be modeled as application of a transfer function on the sound. If $s(t)$ be the sound, then the actual sound captured by the microphone, $m(t)$, is expressed as

$$m(t) = s(t) \otimes h(t), \quad (1)$$

where $h(t)$ is the composite transfer function affecting the sound from the sound source to the microphone. The '$\otimes$' denotes time-domain convolution.

The MFA algorithm [8] consists of filtering the input signal obtained from each microphone with the time reverse of the focus-to-sensor impulse response, where the focus is the focal point of the microphone array. For the sound source located at the focus, the effect of the matched filter is to convolve the undistorted signal with the autocorrelation of the focus-to-sensor response,

$$y_i(t) = m_i(t) \otimes h_{fi}(-t) = s(t) \otimes h(t) \otimes h_{fi}(-t), \quad (2)$$

where $m_i(t)$ denotes the signal at sensor i.

[1] This work was supported by DARPA Contract DABT63-93-C-0037.

[†] Currently with Lucent Technologies, Holmdel, NJ

[‡] Currently with Raytheon Systems, Falls Church, VA

[*] Currently with Phillips Research, Taipei, Taiwan

For a single sound source at the focus, the result of the MFA processing is

$$\begin{aligned} y_{on}(t) &= \sum_{i=0}^{N} \{s_{on}(t) \otimes h_{fi}(t)\} \otimes h_{fi}(-t), \\ &= s_{on}(t) \sum_{i=0}^{N} h_{fi}(t) \otimes h_{fi}(-t), \end{aligned} \quad (3)$$

where s_{on} is the signal originating from the focal region, $h_{fi}(t)$ is the impulse response from the focal point to the microphone i, and N is the number of sensors.

It has been shown [8] that the MFA has a distinct advantage over simple beam-forming in that the MFA can remove reverberation from a captured signal. The SNR improvement has been shown to be proportional to the number of sensors instead of the number of reflections in the case of beam forming. If K be the number of reflections, the signal to reverberant energy can be expressed as, $SNR = NK/K-1$, which is independent of the number of reflections for $K \gg 1$. For comparison, the signal to reverberant energy for delay-sum beam-forming is, $SNR = N/K - 1$, which is monotonically decreasing when K increases.

The MFA has also been shown to reject signals which are not on the focal point. The MFA algorithm was found to return average improvements as much as 6.39dB, in experiments performed in [8], over different levels of reverberation.

3. REGRESSION TREE MLLR HMM ADAPTATION

Various types of adaptation algorithms have been investigated in the past few years, but one of the most successful algorithms is the Maximum Likelihood Linear Regression (MLLR). The MLLR [5] computes an Affine transformation, in an ML sense, to move model parameters so that the resulting model achieves better performance for that speaker. The transformation matrix can be computed for a specific phone class or for a group of phone classes by pooling their adaptation data.

The Regression Tree MLLR [3] is better suited to adaptation as it selects the optimal number of transforms, based upon the amount of data. This optimized adaptation is achieved by clustering all the mixtures of HMMs of the model in the form of a tree, where the root contains all the mixtures. The tree-node grouping may be acoustic or linguistic in nature. The tree is grown till the number of mixtures in the leaf node reaches a certain predefined level in the former case or the phone level in the latter case. The leaf nodes are termed "base classes" while the higher nodes are termed "regression classes". This is usually performed off-line. In this paper we use a linguistic tree for adaptation.

4. EXPERIMENTS AND RESULTS

4.1. HMM-based Recognition System

The baseline HMM model was trained on the 991 word (excluding silence) DARPA Resource Management task [6]. The input speech was preemphasized and windowed at a frame size of 25ms with a frame rate of 10ms. From each frame, 12 dimension Mel-frequency Cepstral Coefficient with Log Energy were extracted. The Δ and $\Delta\Delta$ were appended to result in a total dimension of 39. Cross-word triphone models were used which were generated using tree-based clustering for unseen triphones. There were 6,966 Physical HMMs for 112,849 cross-word triphone contexts with 3 emitting states per triphone and 4 mixtures per state. These were generated using the HTK HMM Toolkit [9]. The language model was word-pair grammar.

4.2. Reverberant Speech Corpus

The reverberant data is picked up from a distance talking speaker using two different types of microphones – one a single directional microphone and the other an 8-element microphone array. The microphone array was backed by the Matched Filter Array (MFA) processing [8].

Speech data for this experiment was collected, by Renomeron [8], in the Varechoic Chamber at Lucent Technologies in Murray Hill, NJ. The Varechoic Chamber is a $6.7 \times 6.1 \times 2.9$ meter room (Fig. 1) with double wall construction for insulation from the outside environment. The chamber has a mechanism of controlling the reverberation level by adjusting panels on the chamber walls. It allows for variation of reflection levels for a reverberation time of 0.1–0.9 seconds. Speech sentences from the RM database [6] were played through loudspeakers and captured by an 8-element microphone array and a single directional microphone. The reverberant data was generated at 4 different reverberation levels of 0.1s, 0.2s, 0.5s and 0.9s. The data was captured at 0 and 3 in Fig (1), i.e., at (2.07,2.20) and at (4.51, 4.54). The loudspeaker was at a height of z=1.4 meters.

4.3. Data Experiments

All evaluations are based on the 12 speaker SD (Speaker Dependent) Eval set of the RM database with 100 sentences per speaker. The performances for a speaker

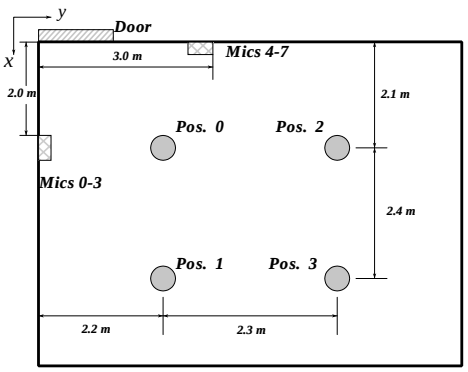


Figure 1: Varechoic Chamber recording configuration.

Position	Microphone Type	Reverberation Level	
		0.9s	0.5s
2.0m	Single	70.6 (36.38)	76.6 (51.03)
2.0m	MFA	85.4 (74.85)	88.5 (85.12)
4.5m	Single	68.1 (31.29)	78.2 (47.53)
4.5m	MFA	87.0 (80.59)	87.3 (83.00)

Table 1: Environment Dependent Training Recognition Accuracy (with MLLR adapted results in brackets).

at a distance of 2m can be seen in Fig (2) and in Fig (3) for the speaker at a distance of 4.5m. An improvement of almost 100%, relative, in recognition performance is seen in many cases. At a distance of 4.5m from the mic, with reverberation at 0.9s, the performance is improved to 80.59% from the baseline performance of 42.11%, a gain of over 90%. In almost all cases the MFA with MLLR processing (MFA-MLLR) returns acceptable performance levels. Furthermore applying MLLR to the original directional microphone speech is seen to be of little value. The MFA-MLLR can be seen to be far superior to using a single directional microphone.

To compare the performance of the MFA-MLLR over Environment Dependent (ED) training, MAP training is performed on SI data generated for the cases where the reverberation level is higher than 0.5s. These conditions have the most serious degradation and the further experiments concentrate to improve them. The training was done on 1200 sentences ($\sim$ 4200s) of speech from the SI set. This avoids improvements that result from seeing a speaker in advance in training thereby biasing the results. The performance of the new models on the ED set is shown in Table (1).

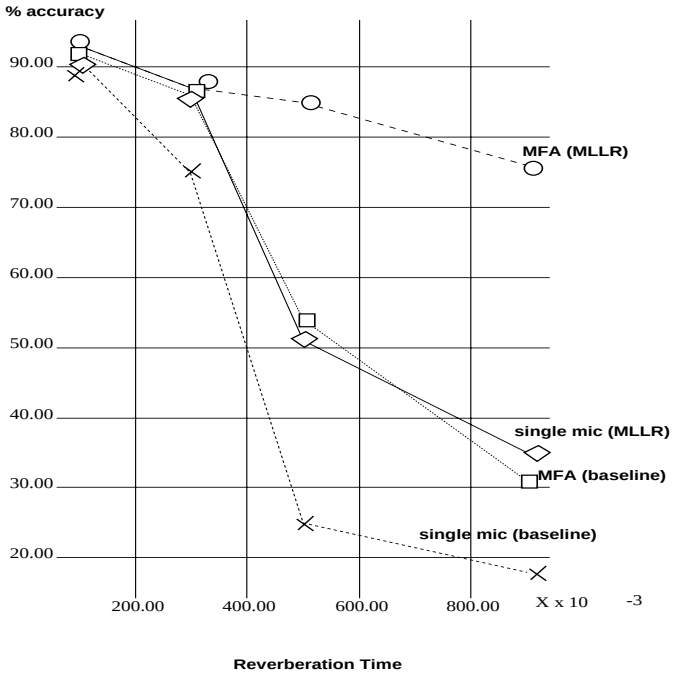


Figure 2: Accuracy versus Reverberation for speakers at a distance of about 2m from the microphone.

The performance is far better than the SI baseline's, which was to be expected. What is remarkable is that the relative improvement for the MFA-MLLR case is comparable to that of Environment dependent training! Though the ED training is still about 5-7% absolute better than MFA-MLLR, the MFA-MLLR achieves this using just $1/12^{\text{th}}$ the data that ED training uses.

In the case that a lot of environment dependent data is available, it would be better to use ED training but in case that the environment conditions keep changing, it would be better to use MFA-MLLR processing to obtain higher performance levels.

If a lot of data from the environment is available prior to the evaluation, as in the above case, then an extra step can be performed. The new ED models can be adapted in unsupervised mode to adapt to the speaker to achieve higher performance levels. This is seen in Fig. (4) for the case when the reverberation level is 0.9s with the microphone at a distance of 4.5m from the microphone (the worst condition). The various strategies used in the figure are the unsupervised adaptation using MFA-MLLR, using ED MAP training and then adaptating to speakers using the ED models (ED+MLLR). The performance of the clean-speech model evaluated on the same 12 speakers for clean speech is 93.2%. It is remarkable that 'ED+MLLR' approaches this.

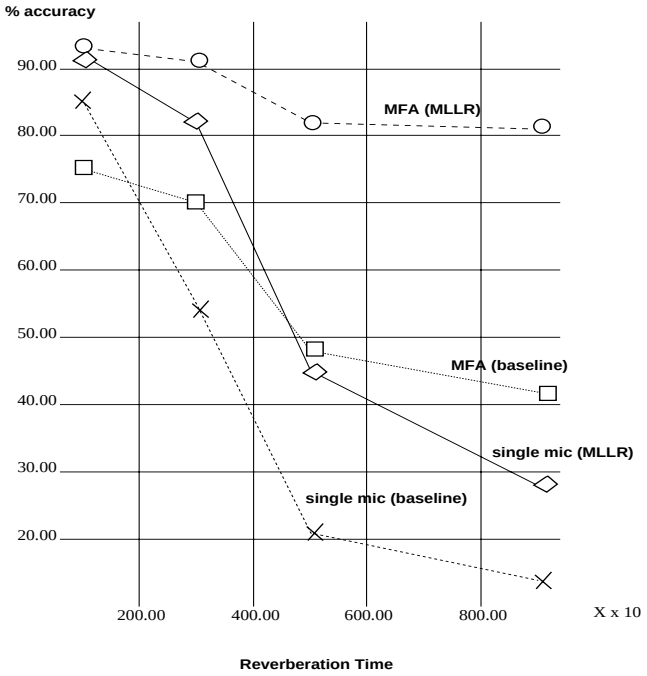
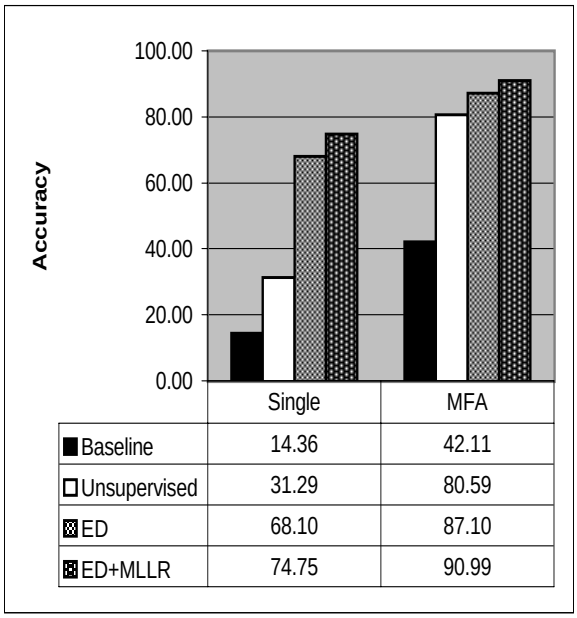


Figure 3: Accuracy versus Reverberation for speakers at a distance of about 4.5m from the microphone.



	Single	MFA
Baseline	14.36	42.11
Unsupervised	31.29	80.59
ED	68.10	87.10
ED+MLLR	74.75	90.99

Figure 4: Recognition Accuracies for Various Strategies, Reverberation = 0.9s, distance from mic = 4.5m.

5. CONCLUSION

MFA-MLLR was found to give substantial improvements over baseline performance. Reason for this is in the fact that MFA achieves higher SNR speech leading to better quality speech. This quality was essential for the MLLR leading to a good unsupervised adaptation performance. Reverberation levels leass than 0.2s affects performance less than reverberation levels greater 0.5s after MFA-MLLR. Effect of distance from microphone greatly reduced after MFA-MLLR. Performance of ED is only marginally better than MFA-MLLR despite using 12 times as much data. Thus, if sufficient data is available, adaptation in two steps is useful – the first step to the environment using all the data available from that environment and in the second, to the speaker.

6. REFERENCES

[1] M. Omologo D. Guiliani and P. Scaizer. Experiments of speech recognition in noisy and reverberant environment using a microphone array and HMM adaptation. In Proc. Eurospeech, 1996.

[2] M. Omologo D. Guiliani, M. Matassoni and P. Scaizer. Robust continuous speech recognition using microphone arrays. In Proc. Eurospeech, volume 3, pages 2021–2024, September 1995.

[3] M. J. F. Gales and S. Young. The generation and use of Regression Class Trees for MLLR adaptation. Technical Report CUED/F-INFENG/TR. 263, Cambridge University, August 1996.

[4] D. J. Mashao J. E. Adcock, Y. Gotoh and H. F. Silverman. Microphone array speech recognition via incremental MAP training. In Proc. ICASSP, 1996.

[5] C. J. Leggetter. Improved Acoustic Modelling for HMMs using Linear Transformations. PhD thesis, Cambridge University, 1995.

[6] J. Bernstien P. Price, W. Fisher and D. Pallett. A database for continuous speech recognition in a 1000-word domain. In Proc. ICASSP, pages 651–654, 1988.

[7] P. Raghavan. Speaker and Environment Adaptation in Continuous Speech Recognition. Master's thesis, Rutgers University, New Brunswick, NJ, 1998.

[8] R. J. Renomeron. Spatially selective sound capture for teleconferencing systems. Master's thesis, Rutgers University, Dept. Electrical and Computer Engineering, New Brunswick, NJ, October 1997.

[9] S. J. Young. The HTK Hidden Markov Model Toolkit V2.0. Cambridge University Engg. Dept.