Combining MLLR Adaptation and Feature Extraction for Robust Speech Recognition in Reverberant Environments

Aik Ming Toh[1], Roberto Togneri[1], Sven Nordholm[2]

School of Electrical, Electronic and Computer Engineering
[1]The University of Western Australian, Australia
[2]Western Australian Telecommunication Research Institute
aikming@ee.uwa.edu.au, roberto@ee.uwa.edu.au, sven@watri.org.au

Abstract

This paper presents an investigation on speech recognition performance in reverberant environments. Reverberant noise has been a major concern in speech recognition systems. Many speech recognition systems, even with state-of-art features, fail to respond to reverberant effects and the recognition rate deteriorates. This shows the limitations of robust feature extraction in reverberant environment. The maximum likelihood linear regression (MLLR) adaptation scheme is adopted for reverberant speech recognition on the TI-DIGIT database. The use of adaptation data improved the recognition performance significantly especially for strong reverberations. The performance of the MFCC_0_D_A features improved by more than 15% for reverberation level greater than 0.4s. The recognition performance is maintained above 90% up to reverberation time of 0.5s. This paper demonstrates the optimal strength of both robust feature extraction and adaptation scheme for reverberant speech recognition.

1. Introduction

Current state-of-the-art speech recognition systems show impressive performance for clean acoustic environments. However, the performance degrades when there is a mismatch between the training data and the testing data. Large amount of training data is required to retrain speech recognition systems to a new environment. This alternative is not feasible as it is both computationally expensive and time consuming. Therefore, it is desirable to be able to improve the performance of the system while using a small amount of adaptation data. Several adaptation schemes have been introduced into speech recognition systems to enhance the performance and robustness in noisy environments (Giuliani, Omologo, and Svaizer 1996). However, most of these researches focus on additive noise (Kristjansson, Frey, Deng, and Acero 2001), (Yao, Yu, and Huang 1996).

Several adaptation techniques have been adopted for reverberant speech recognition (Raghavan, Renomeron, Che, Yuk, and Flanagan 1999),(Couvreur, Boite, Dupont, Ris, and Couvreur 2001) but most of the research focus on the use of model estimation (Couvreur, Ris, and Couvreur 2001) and microphone array processing (Giuliani, Matassoni, Omologo, and Svaizer 1999) in reverberant speech recognition. This paper focuses on the use of adaptation data and speech features to improve the recognition accuracy and robustness.

From previous work (Toh, Togneri, and Nordholm 2005), it has been shown that feature extraction alone is not enough to alleviate reverberant noise. No single feature set performed best for all different levels of reverberation. While one feature may be robust in additive noise, it may not be robust in reverberant noise. Even the state-of-the-art feature Mel-frequency cepstral coefficient with delta and double delta features (MFCC_0_D_A) suffers in reverberant environments.

There are papers which have reported that even with reverberant matched models, the recognition rate cannot be improved sufficiently when the reverberation time is greater than 0.4s (Baba, Lee, Saruwatari, and Shikano 2002), (Yamamoto, Nishimoto, and Sagayama 2004). The results presented in this paper demonstrate a different perspective. It is shown that by using a small portion of the training data as adaptation data, the recognition rate can be significantly improved.

In this paper, the performance of the adaptation scheme in reverberant speech recognition is compared with the performance of the acoustic matching scheme. In addition, the significance of using both MFCC_0_D_A features and maximum likelihood linear regression (MLLR) adaptation scheme in reverberant speech recognition is also demonstrated. Significant improvements have been achieved for severe reverberation time.

The paper is organized as follows: The next section explores the effects of reverberation on speech signals. Section 2 provides information on the state-of-the art features. The MLLR adaptation scheme is presented in section 3. Section 4 specifies the experimental setup followed by the results in section 5. The last section comprises the conclusions.

2. Reverberant effects on speech signals

Human auditory systems are accustomed to speech in moderate reverberant environments. However, the performance of a speech recognition system degrades dramatically in reverberant conditions. Reverberation is caused by

This research is partly funded by the National ICT Australia (NICTA). National ICT Australia is funded through the Australian's Government Backing Australia's Ability initiative, in part through the Australian Research Council (ARC).

Proceedings of the 11th Australian International Conference on Speech Science & Technology, ed. Paul Warren & Catherine I. Watson. ISBN 0 9581946 2 9
University of Auckland, New Zealand. December 6-8, 2006.

Accepted after full paper review

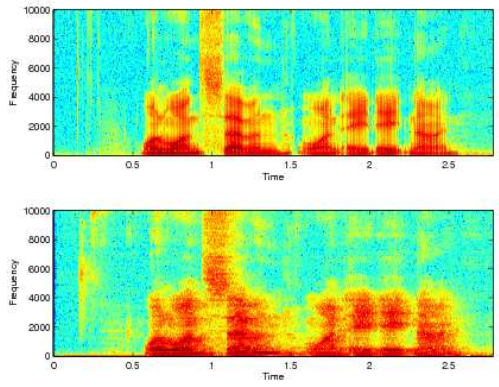

Figure 1: Spectrogram for clean connected digit utterance ”1-1-7-1-8-8-9” and the utterance under the influence of reverberation time 0.2s

the superposition of an acoustic signal and its reflected signals of different delays and amplitudes. It introduces a convolutional interference that comprises both spectral distortion and additive noise.

Reverberation typically increases the loudness at a given location because the energy generated over a range of time in the past is received in the present. The syllable onsets and identities of the speech can be masked by the decaying energy from previous syllables when the reverberation time, RT60 is long (Gold and Morgan 2000). This can hurt the intelligibility of the speech severely. RT 60 is used to characterise the reverberation time. It is the time interval in which the reverberation level decays by 60dB.

An obvious effect of reverberation is temporal smearing on the acoustic signal. This effect is also known as overlap masking in which, segments of an acoustic signal are affected by the reverberant components of previous segments. Although this adds richness of sound to music, it makes the speech loses its intelligibility. Human listeners do not have much trouble understanding reverberant speech because most of the basic energy features are still in intact even under the influence of reverberations. However, it is difficult for speech recognition systems since they are trained in clean models. The smearing effects are evident in the temporal resolution of both spectral and cepstral features. Figure 1 shows the spectrogram of a speech utterance in clean and reverberant condition. The spectrogram of the reverberant utterance clearly depicts the temporal smearing effect.

In addition, reverberation affects the statistical properties such as the mean and variance of speech features. The shifts in these parameters can be compensated with normalization strategies such as mean and variance normalization. However, the temporal smearing effect cannot be effectively compensated in the feature context. Thus, adaptation is proposed to improve the performance of speech recognition in reverberant environment.

3. Feature extraction

The speech signal is first segmented into frames and each frame is windowed and transformed into a power spectrum using the short-time Fourier transform. The short term power spectrum is then warped by Mel-scale using the Mel-scale filter bank. Logarithmic compression is applied to each of the Mel spectral vectors to approximate the human auditory processing. The log Mel spectrum is then converted into MFCC vectors via Discrete Cosine Transform (DCT). Apart from that, the MFCC_0 coefficients constitute the zeroth cepstral coefficient which represents the energy of the MFCC feature vectors.

The MFCC vectors are appended by dynamic features which indicate the rate of change, delta (D) or an acceleration of the cepstral components, double delta (A). The dynamic features describe the trajectories of speech parameters in the vicinity of a given speech vector (Furui 1981). Dynamic features typically perform poorly on their own due to deemphasis of features with lower frequency which also carry important linguistic information. Therefore, dynamic and acceleration features are mostly used in conjunction with static features.

4. Maximum likelihood linear regression

The maximum likelihood linear regression adaptation scheme is used in this work. This method is simple and known to be robust for unsupervised adaptation as well as effective for small amount of adaptation data. MLLR was initially developed for speaker adaptation (Legetter and Woodland 1995). The aim of MLLR is to obtain a set of transformation matrices that maximizes the likelihood of the adaptation data. The transformation sets are relatively small compared to the total number of Gaussian parameters computed from the training data.

A single global transformation is used for the small amount of data. In this paper, both the mean parameter estimation and Gaussian variance are updated and adapted in two separate stages. The variances are updated after the new mean is derived. The transformation matrices are tied across a number of Gaussians to ensure robust estimation of transformation parameters. This set of Gaussians, which share a transform, is referred to as a regression class.

The hidden Markov models (HMM) are modified such that (Legetter and Woodland 1996)

$$\mathcal{L}(\mathcal{O}_\mathcal{T}|\check{\mathcal{M}}) \geq \mathcal{L}(\mathcal{O}_\mathcal{T}|\hat{\mathcal{M}}) \geq \mathcal{L}(\mathcal{O}_\mathcal{T}|\mathcal{M}) \qquad (1)$$

where $\mathcal{L}$ is the likelihood, $\mathcal{M}$ is the original model set, $\hat{\mathcal{M}}$ has the updated mean parameters, $\check{\mathcal{M}}$ set has updated both means and variances and $\mathcal{O}_\mathcal{T}= \{o(1), o(2), ..., o(T)\}$ is the adaptation data.

The transformation matrix used to produce a new estimate of the adapted mean is given by (Gales, Pye, and Woodland 1996),(Legetter and Woodland 1995)

$$\hat{\mu} = W\varepsilon \qquad (2)$$

$$\varepsilon = [\omega\ \mu_1\ \mu_2\ ...\ \mu_n]^T \qquad (3)$$

where W is the $n \times (n + 1)$ transformation matrix (for n dimensional data), ε is the extended mean vector and ω represent a bias offset of value 1.

Proceedings of the 11th Australian International Conference on Speech Science & Technology, ed. Paul Warren & Catherine I. Watson. ISBN 0 9581946 2 9
University of Auckland, New Zealand. December 6-8, 2006.

Accepted after full paper review

The Gaussian variance vectors or the covariance matrices are updated using the transformation (Gales, Pye, and Woodland 1996);

$$\hat{\sum_{m}} = B_m^T \hat{H}_m B_m \tag{4}$$

where $\hat{H}_m$ is the linear transformation to be estimated and B_m is the inverse of the Cholesky factor of $\sum_m^{-1}$, such that$\sum_m^{-1} = C_m C_m^T$ and $B_m = C_m^{-1}$.

The variance transformation is shared over a number of Gaussians and the maximum likelihood estimate is given by

$$\hat{H}_m = \frac{\sum_{r=1}^{R} C_r^T \left[\sum_{r=1}^{T} L_r(\tau)(o(\tau) - \hat{\mu}_r)(o(\tau) - \hat{\mu}_r)^T \right] C_r}{\sum_{r=1}^{R} \sum_{r=1}^{T} L_r(\tau)} \tag{5}$$

where $\hat{\mu}_r$ is the previously calculated mean.

The formulations for the transformation matrix used to give a new estimate of the adapted mean and variance based on (2) have been well reported (Legetter and Woodland 1996).

5. Experimental setup

The adult portion of the TI-Digit database was used in this work. The database comprised both the isolated and connected digit utterances. The training data contained utterances of 55 male and 57 female speakers. There were also 23 male and 27 female speakers for the testing data. Each speaker subset composed of 77 digit utterances.

Several sets of adaptation data have been extracted from the training data. The adaptation data sets comprised the eleven isolated digit utterances from randomly chosen speakers in the training data. These adaptation data were corrupted to match the reverberant condition of the testing data described below. Table 1 provides a summary of the database used.

Reverberant effects were captured by estimating the impulse response of the room environment from long segments of speech. The experiment used the room impulse response designed to match the characteristic of a 2.2m high, 3.1m wide and 3.5m long room. The microphone and the speakers were localized 0.5m from the wall at opposite ends. The speech was then convolved with the RT60 room impulse response. RT60 is the time interval in which the reverberation level decays by 60dB. The number of filter coefficients was adjusted according to the reverberation time.

All the speech files were pre-emphasized and windowed with a Hamming window. The speech signal was analyzed every 10ms with a frame width of 25ms. A Mel-scale triangular filterbank with 26 filterbank channels was used to generate the Mel-frequency cepstral coefficients (MFCC) features. The MFCC_0 coefficients constituted 12 static MFCC vectors and the zeroth cepstral coefficients. The hidden Markov model (HMM) used 15 states and 5 mixtures for the connected digit recognition.

Table 1: Details of the database

	Size (Mb)	Speakers	Utterances
Train data	637	112	8624
Test data	277	50	3850
Adapt 50.0Mb	50.0	112	1232
Adapt 25.0Mb	25.0	55	605
Adapt 12.5Mb	12.5	27	297
Adapt 6.3Mb	6.3	14	154
Adapt 3.0Mb	3.0	7	77
Adapt 1.2Mb	1.2	3	33

6. Experimental results

In the author's previous work on robust features for speech recognition in hostile environments (Toh, Togneri, and Nordholm 2005), it has been shown that different features showed robustness in different levels of reverberation. Preliminary experiments with baseline MFCC_0 features have shown that MFCC_0 with MLLR adaptation surpassed the performance of the state-of-the art feature MFCC_0_D_A in reverberant environments. Significant improvements were achieved for severe reverberant condition. Thus, the work was extended to the use of MFCC_0_D_A features and its recognition using the full TI-digit database.

6.1. Matched acoustic conditions

A number of slightly mismatched speech recognition experiments were performed to show the effectiveness of acoustic matching scheme in reverberant environments. A reverberation time such as 0.4s is both clearly and audibly perceived by human ears. This is the reverberation level where the performance of speech recognition system starts to deteriorate significantly in mismatched environments. Reverberation time of 0.5s onwards can be regarded as severe reverberation and the recognition performance degrades rapidly in these regions.

Table 2 shows the recognition accuracy of reverberant testing data with training data matched to different reverberation time. The mismatched row records the results for speech recognition with clean training data and reverberant testing data. The columns represent the reverberant testing environments while the subsequent rows represent the reverberant training data. Matched acoustic matching scheme is illustrated by recognition accuracy in bold. This

Table 2: Speech recognition in slightly mismatched environments

RT(s)	0.1s	0.2s	0.3s	0.4s
Mismatched	98.71	98.09	95.86	86.06
Train 0.1s	**99.19**	99.17	98.39	95.24
Train 0.2s	99.03	**99.22**	98.66	96.01
Train 0.3s	98.48	98.88	**98.68**	97.96
Train 0.4	93.10	95.95	98.31	**98.08**

Proceedings of the 11th Australian International Conference on Speech Science & Technology, ed. Paul Warren & Catherine I. Watson. ISBN 0 9581946 2 9
University of Auckland, New Zealand. December 6-8, 2006.

Accepted after full paper review

Table 3: Speech recognition with isolated digits adaptation (50Mb) in reverberant environments

Test data RT(s)	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8
Matched	99.19	99.22	98.68	98.08	97.11	96.35	91.35	84.25
Mismatched	98.71	98.09	95.86	86.06	73.51	57.95	47.03	38.98
Adapt 0.1s	**98.91**	**98.39**	96.40	93.81	87.30	75.63	59.63	47.99
Adapt 0.2s	98.56	**98.39**	96.76	93.82	87.94	77.57	63.20	50.94
Adapt 0.3s	98.35	98.12	**97.02**	**95.14**	**90.73**	81.31	68.19	55.15
Adapt 0.4s	97.95	97.56	96.61	94.65	90.28	**82.16**	**70.90**	58.78
Adapt 0.5s	97.14	96.93	96.07	93.83	89.28	81.41	69.85	**59.42**
Adapt 0.6s	96.31	95.95	94.83	92.60	87.68	80.60	69.28	59.21
Adapt 0.7s	95.63	95.40	93.78	91.92	87.42	79.68	69.14	58.03
Adapt 0.8s	94.94	94.95	92.99	91.71	87.31	78.10	68.85	57.97

table shows that training in even a slightly mismatched environment such as RT 0.1s would improve the recognition accuracy for RT 0.4s by 9.18%. The use of better matched conditions would further improve the recognition accuracy.

Optimal performance in the recognition accuracy was achieved with matched reverberation time for each reverberant conditions. Recognition rates of more than 90% were achieved up to a reverberation time of 0.7s. Significant improvements were more evident for severe reverberations such as RT 0.5s and RT 0.6s.

6.2. MFCC_0_D_A and MLLR adaptation

The use of acoustic matching scheme shows optimal results but this requires the training data be corrupted to match the acoustic conditions of the reverberant testing data. Such an option is computationally expensive as well as time consuming, considering the vast amount of training data required. Thus, the use of adaptation scheme such as MLLR is recommended to adapt the clean speech models to the corrupted speech models.

The MLLR scheme was used to adapt clean speech models to the reverberant adaptation data in this work. The aim was to keep the adaptation data to a minimum while optimizing the improvements in the recognition accuracy.

Table 3 and Table 4 depict the results of speech recognition with MLLR adaptation scheme in reverberant environments. The results show that significant improvements were achieved in the recognition accuracy by adapting the HMM models to noisy data. With adaptation, even slightly mismatched adaptation data contributes to improvement in the recognition performance.

Table 3 records the recognition results for speech recognition with MLLR adaptation using only isolated digit utterances. The Matched row records the result for matched acoustic speech recognition while the Mismatched represents the results for speech recognition with clean training data and reverberant testing data. Subsequent rows in the table show the recognition results of the training data adapted to respective reverberant levels. Optimal performances are highlighted in bold for each reverberant testing environments.

It is apparent that the recognition performance degrades rapidly in severe reverberant regions such as reverberation time of 0.5s onwards. However, with adaptation, recognition accuracy of more than 90% could be maintained up to RT 0.5s, with 17.22% improvement over the mismatched recognition. Improvements of at least 14% were achieved for severe reverberation such as 0.5s and above.

The adaptation data was reduced to observe the impact of adaptation data size in relation to speech recognition performance in reverberant environments. The adaptation data was subsequently halved down to 1.2Mb. The smallest set Adapt 1.2Mb would account for just below 0.2% of the training data size. We have performed the recognition experiments and obtained several sets of result similar to Table 2 for each adaptation data set. Table 4 records the optimal recognition accuracy for each reverberant environments with the use of different adaptation data sets. The optimal results refer to the best recognition results obtained using the adaptation data, which could have the reverberation time (RT) different or less than the testing reverberant environments.

It is interesting to observe that significant degradation in recognition accuracy is evident only when the adaptation data is reduced to 6Mb, which is about one percent of the training data. Regardless of the size of the adaptation data, we could still improve the recognition performance by at least 15% in RT 0.7s and RT 0.8s with MLLR adaptation.

Both the training data and the testing data comprised isolated and connected digits data, while the adaptation data only contained isolated digits data. The use of isolated digits in adaptation data has shown favourable improvements in recognition accuracy as shown in Table 1 and Table 2. This refuted the common misconception that adaptation schemes required large amount of adaptation data or acoustically matched data.

Observations from all sets of adaptation data demonstrate that the MLLR adaptation scheme offers more contribution to the recognition in more severe reverberations. The adaptation scheme improved the recognition accuracy by more than 10% for speech recognition in RT60 of 0.5s to 0.8s.

It is also observed that the optimal recognition accuracy in each reverberant environment is not achieved by adapting to the same reverberation level. In Table 3, it can be contemplated that reverberation level of 0.3s to 0.5s can be sufficiently compensated by the adaptation of 0.3s data. Optimal performance for reverberation 0.6s and 0.7s are

Proceedings of the 11th Australian International Conference on Speech Science & Technology, ed. Paul Warren & Catherine I. Watson. ISBN 0 9581946 2 9
University of Auckland, New Zealand. December 6-8, 2006.

Accepted after full paper review

Table 4: Speech recognition accuracy with different sets of adaptation data in reverberant environments

Test data RT(s)	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8
Mismatched	98.71	98.09	95.86	86.06	73.51	57.95	47.03	38.98
Adapt 50.0Mb	98.91	98.39	97.02	95.14	90.73	82.16	70.90	59.42
Adapt 25.0Mb	98.96	98.43	97.01	94.86	90.44	81.80	70.84	59.73
Adapt 12.5Mb	98.90	98.39	97.13	94.69	90.21	81.12	70.35	59.30
Adapt 6.0Mb	98.73	98.35	96.61	94.45	89.66	80.50	69.59	57.86
Adapt 3.0Mb	98.58	98.24	96.76	94.20	88.85	79.40	67.44	55.79
Adapt 1.2Mb	98.61	98.11	96.12	93.12	86.83	77.38	64.63	54.43

achieved with 0.4s adaptation data. In Table 4, most of the optimal recognition accuracies were achieved with adaptation data corrupted to reverberant level which was less than the testing environments.

The experiments have shown that slightly mismatched adaptation would be sufficient to compensate the effects of reverberation on speech recognition. A first order estimate of the room impulse response could be sufficient even if the true reverberant conditions of the testing environment were not known. Adaptation did not experience the constraint imposed by matched acoustic speech recognition. The optimal performance for acoustic matching scheme in each reverberant level was achieved with the use of the matched reverberation time.

The use of MLLR adaptation scheme achieved significant improvements in recognition accuracy, which surpassed the performance of the state-of-the-art features such as MFCC_0_D_A. The experiments have also established the strength of adaptation for speech recognition particularly in severe reverberant environments. The MLLR adaptation method is an effective and feasible solution to reverberant speech recognition since it only requires small amount of adaptation data for substantial improvements in recognition rate.

7. Conclusion

The amount of adaptation data used is trivial compared to the vast training data used in the experiment. Though the adapted results are not as good as the matched acoustic results, substantial improvements in the recognition accuracy have been achieved with just a small amount of adaptation data. These improvements are significant and it is more apparent for reverberation times such as 0.4s which are common in real-world scenarios.

It has been shown that with matched environments, it is possible to maintain the recognition accuracy above 90% up to severe reverberations such as RT of 0.7s. The use of adaptation scheme with MFCC_0_D_A features has enabled us to maintain the recognition accuracy above 90%, even for reverberant conditions up to 0.5s. Significant improvements have been achieved with the use of adaptation data in compensating for the reverberant effects in speech recognition systems.

It is also important to note that only isolated digit utterances were used for adaptation data and normalization strategies were not implemented. The performance can be further enhanced if the adaptation data comprises some connected digits utterances or more utterances and the features are normalized.

Therefore, we conclude that the combination of robust feature extraction scheme and adaptation scheme offers the optimal performance for speech recognition in reverberant environments.

References

Baba, A., A. Lee, H. Saruwatari, and K. Shikano (2002, Sept). Speech recognition by reverberation adapted acoustic models. In *Proc. ASJ*, pp. 27–28.

Couvreur, L., J. Boite, S. Dupont, C. Ris, and C. Couvreur (2001, Aug.). Fast adaptation for robust speech recognition in reverberant environments. In *Proc. ITRW*, France, pp. 85–88.

Couvreur, L., C. Ris, and C. Couvreur (2001). Model-based blind estimation of reverberation time: application to robust ASR in reverberant environments. In *Proc. Eurospeech*, Denmark, pp. 2635–2638.

Furui, S. (1981). Cepstral analysis technique for automatic speaker verification. *IEEE Transactions ASSP 29*, 254–272.

Gales, M., D. Pye, and P. Woodland (1996). Variance compensation within the MLLR framework for robust speech recognition and speaker adaptation. In *Proc. ICSLP*, pp. 1832–1835.

Giuliani, D., M. Matassoni, M. Omologo, and P. Svaizer (1999). Training of HMM with filtered speech material for hands-free recognition. In *Proc. ICASSP*, Italy.

Giuliani, D., M. Omologo, and P. Svaizer (1996). Experiments of speech recognition in noisy and reverberant environment using a microphone array and HMM adaptation. In *Proc. ICSLP*, USA.

Gold, B. and N. Morgan (2000). *Speech and Audio Signal Processing*. John Wiley & Sons.

Kristjansson, T., B. Frey, L. Deng, and A. Acero (2001, May). Towards non-stationary model-based noise adaptation for large vocabulary speech recognition. In *Proc. ICASSP*.

Legetter, C. and P. Woodland (1995). Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models. *Computer Speech and Language 9*, 171–185.

Proceedings of the 11th Australian International Conference on Speech Science & Technology, ed. Paul Warren & Catherine I. Watson. ISBN 0 9581946 2 9
University of Auckland, New Zealand. December 6-8, 2006.

Accepted after full paper review

Legetter, C. and P. Woodland (1996). Mean and variance adaptation within the MLLR framework. *Computer Speech and Language 10*, 249–264.

Raghavan, P., R. Renomeron, C. Che, D. Yuk, and J. Flanagan (1999). Speech recognition in a reverberant environment using matched filter array MFA processing and linguistic-tree maximum likelihood linear regression LT-MLLR adaptation. In *Proc. ICASSP*, pp. 777–780.

Toh, A., R. Togneri, and S. Nordholm (2005). Investigation of robust features for speech recognition in hostile environments. In *Proc. APCC*, pp. 956–960.

Yamamoto, H., T. Nishimoto, and S. Sagayama (2004, Jan). Frame-by-frame HMM adaptation for reverberant speech recognition. In *Proc. SWIM*.

Yao, L., D. Yu, and T. Huang (1996). An unified spectral transformation adaptation approach for robust speech recognition. In *Proc. ICSLP*, pp. 981–984.

Proceedings of the 11th Australian International Conference on Speech Science & Technology, ed. Paul Warren & Catherine I. Watson. ISBN 0 9581946 2 9
University of Auckland, New Zealand. December 6-8, 2006.

Accepted after full paper review