社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

状態分割と分布の畳み込みを用いる残響音声のためのモデル適応

ラウト　チャンドラ カント　　西本　卓也　　嵯峨山茂樹

東京大学大学院 情報理工学系研究科
〒 113–8656 東京都文京区本郷 7–3–1
E-mail: {raut,nishi,sagayama}@hil.t.u-tokyo.ac.jp

あらまし　本稿は、残響音声に頑健な音声認識を実現するためのモデル適応法について報告する。あるクリーンな HMM の状態をいくつかのサブ状態に分割して、先行する状態出力系列がよりよく推定できるような新しい HMM に変換する。そして、推定した先行する状態系列を残響特徴量のパラメータに畳み込んで残響音響モデルを合成する。本手法の評価として残響下音声の特定話者孤立単語音声認識実験を行ない効果を確認した。残響時間が 120ms の残響環境では単語認識率は、クリーン HMM では 30.1%、CMS では 39.8%であったのに対して、提案手法では 52.1%に上昇した。
キーワード　残響，畳み込み雑音，モデル適応，状態分割

Acoustic Model Adaptation for Reverberant Speech by State Splitting of HMM and Convolution of Distributions

Chandra Kant RAUT, Takuya NISHIMOTO, and Shigeki SAGAYAMA

Graduate School of Information Science and Technology, The University of Tokyo
7–3–1 Hongo, Bunkyo-ku, Tokyo, 113–8656 Japan
E-mail: {raut,nishi,sagayama}@hil.t.u-tokyo.ac.jp

Abstract　This paper describes a model adaptation technique for robust recognition of reverberant speech. The states of given clean HMM are split into a number of substates that enables close estimation of preceding state sequence and thus the preceding output distributions for a given state. This estimated sequence of speech densities, occurred before the state, is then convolved with the spectral parameters of impulse response in appropriate domain to find the distribution for reverberant speech for the state. The experimental results showed significant improvement in recognition rate of ASRs with this method; typically, for convolutional noise with reverberation time (T_{60}) of 120 ms, the recognition rate increased from 30.1% by clean model to 52.1% by current method compared to 39.8% by Cepstrum Mean Subtraction (CMS).
Key words　Reverberation, Convolutional Noise, Model Adaptation, State Splitting

1. Introduction

Automatic speech recognition (ASR) systems, though usually trained with clean speech, have to operate under real-life environment for any practical purposes. But the speech signal in real life is always distorted by additive and convolutional noise, and speech recognition system trained with clean speech performs poorly under such condition. Additive noise is caused by sounds coming from other interfering sources active at the same time, e.g., fan or car or other speakers. Its effect on the input speech signal appears as addition in the waveform and linear-spectral domain. The convolutional noise in speech is usually caused by channel effects, microphone characteristics and reverberation of a room, and is usually characterized by reverberation time (T_{60}) of impulse response (acoustic transfer function) of the transmitting medium. Reverberation time longer than 100 ms are not uncommon for office rooms [1]. The effect of such convolutional noise on the input speech signal appears as convolution in waveform domain, and can severely degrade the performance of ASRs. For example, word accuracy of the SPHINX speech recognition system has been reported to drop from 85% to 20% when a desktop microphone was substituted for close-talking microphone used for training [2].

There are varieties of techniques to deal with

such convolutional noise. These techniques can be broadly categorized into two classes depending upon where they are applied in the recognition system and whether they attempt to restore the clean speech signal or compensate for the distortion: feature-based techniques and model-based approaches.

Feature-based techniques attempt to enhance the perceived quality of speech or feature at the front-end, and include inverse filtering (e.g., [3]), microphone array based techniques (e.g., [3] and [4]), channel normalization techniques (e.g., [5]) including Cepstrum Mean Subtraction (CMS) [6] and RASTA [7]. Though these methods have been proved to improve the performance of ASRs, they cannot perform well when additive noise is also present or when reverberation time is too long. Further, these methods usually have high computational cost.

Model-based approaches like [8]〜[11], on the other hand, operate to reduce the mismatch between the trained model and working environment. Though the assumptions of different techniques vary, depending upon the computational complexity, domain of applicability and performance, current tendency seems in favor of model-based approach than noise removal approach [12]. However, most of the current model adaptation approaches work well only with short reverberation, and are unable to account the effect of preceding frames of speech effectively, resulting in degraded performance when convolutional noise spans over several frames and additive noise is also present.

This work considers the case with long reverberation time and gives a way to account the effect of preceding frames for compensating the model parameters. As the compensation is done in model domain, the method has much less computational cost compared to the methods working at the front-end. Further, the method is less sensitive to deviation in channel parameters used for compensating the models than it would have been in the feature domain. The method also gives flexibility for compensating additive noise, in any (feature or model) domain or both. Therefore once the model is adapted for convolutional noise, one or more of the enhancement or compensation techniques can be applied for additive noise.

The paper is organized as follows: Section 2 describes effect of long convolutional noise on speech and formulates the model convolution approach, Section 3 describes the state splitting technique to obtain the preceding state sequence and corresponding output densities for convolution, and Section 5 summarizes the algorithm. The result of the evaluation of the method and future work are discussed in the subsequent sections.

2. Effect of Long Reverberation and Model Convolution

Reverberation of a room can be modeled by passing clean speech signal through a filter with impulse response of the propagation channel as transfer function, such that reverberant speech is given by

$$o[m] = h[m] * s[m] \tag{1}$$

where $s[m]$ is clean speech, $h[m]$ is impulse response, $o[m]$ is reverberant speech, m is sample number and $*$ represents convolution in time domain.

This method assumes that impulse response $h[m]$ of the propagation channel or room is given. The impulse response of a room under consideration can be found by playing with white noise [1], analyzing stereo recordings from close-talking and far field microphone [1] or playing with sine waves of different frequencies or time-stretched pulse (TSP) [13].

Taking the short-time Fourier transform (STFT) of Eq. 1 gives

$$O(w_i, t) \approx H(w_i, t)S(w_i, t) \tag{2}$$

where t is frame number and w_i is discrete frequency. Parameters $S(w_i, t)$, $O(w_i, t)$ and $H(w_i, t)$ are STFTs of clean speech $s[m]$, distorted speech $o[m]$ and impulse response $h[m]$, respectively.

However, when impulse response is longer than the analysis window-length, Eq. 2 no longer holds; and the effect of long convolutional noise on the short-time Fourier transform of speech is given by

$$O(w_i, t) \approx H(w_i, t) * S(w_i, t) \tag{3}$$

where $*$ represents convolution along frame (time).

From Eq. 3, the kth mel filterbank output is given as

$$O_k(t) = \sum_{\forall w_i} m_k(w_i)\big[H(w_i, t) * S(w_i, t)\big] \tag{4}$$

where $m_k(w_i)$ is the filter gain for kth filterbank. Further analysis gives

$$\begin{aligned} O_k(t) &= \sum_{\forall w_i} m_k(w_i)\big[H(w_i, 0)S(w_i, t) \\ &\quad + H(w_i, 1)S(w_i, t-1) + \dots\big] && (5) \\ &\approx \bar{H}_k(0) \sum_{\forall w_i} m_k(w_i)S(w_i, t) \\ &\quad + \bar{H}_k(1) \sum_{\forall w_i} m_k(w_i)S(w_i, t-1) + \dots && (6) \\ &= \bar{H}_k(0)S_k(t) + \bar{H}_k(1)S_k(t-1) + \dots && (7) \\ &= \bar{H}_k(t) * S_k(t). && (8) \end{aligned}$$

It should be noted that impulse response spectrum has to be assumed constant along frequency within

the band of kth filterbank, in order to convolve with clean speech filterbank parameters $S_k(t)$. We take weighted average of $H(w_i, t)$ over the band of kth filterbank, and assume it constant along the frequency of the band, such that

$$\bar{H}_k(t) = \frac{\sum_{\forall i} m_k(w_i) H(w_i, t)}{\sum_{\forall i} m_k(w_i)}. \quad (9)$$

Mel-frequency Cepstral Coefficients (MFCCs) are generally used for HMM parameters. To compensate the model parameters, they are transformed from cepstral domain to mel-domain, and then convolved with the channel transfer function parameters and transformed back to cepstral domain. Therefore, given clean speech HMM with cepstral domain parameters $\boldsymbol{S}^c$, the HMM parameters for corrupted speech is given as

$$\boldsymbol{O}^c(t) \approx \mathcal{F}\left(\log\left(\exp\left(\mathcal{F}^{-1}\left(\boldsymbol{S}^c(t)\right)\right) * \bar{\boldsymbol{H}}(t)\right)\right) \quad (10)$$

where $\mathcal{F}$ is Discrete Cosine Transform (DCT). Rewriting Eq. 10 for mel filterbank (linear) domain gives

$$O_k^{lin}(t) = S_k^{lin}(t) * \bar{H}_k(t) \quad (11)$$

$$= S_k^{lin}(t)\bar{H}_k(0) + {S_k}^{lin}(t-1)\bar{H}_k(1) + S_k^{lin}(t-2)\bar{H}_k(2) + \ldots \quad (12)$$

where superscript *lin* refers to linear mel-domain parameters.

During model adaptation, as observations are not available, the philosophy adopted in the method is to use observation density function instead, and therefore Eq. 12 represents the convolution of speech densities with averaged spectral parameters of impulse response, $\bar{H}_k(t)$, in mel filterbank (linear) domain. But clean speech densities are no longer Gaussian in linear domain (rather they are log-normally distributed), and finding distribution for corrupted speech parameters will not be straightforward anymore. Similar approximations as in Parallel Model Combination [14] to find the distribution of corrupted speech can be applied in this case too. For example, log-normal approximation [14] can be applied, with the assumption that sum of two or more log-normally distributed variables are still log-normal. In that way, both mean and covariance matrix for corrupted speech can be estimated. However, this may result in poor estimation of covariance matrix. Alternatively, instead of using output densities, only mean vectors of the estimated sequence of densities can be used to estimate the mean vector for reverberant speech, retaining the covariance matrix same as clean speech.

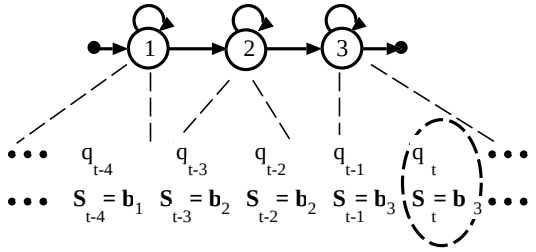

Fig. 1 Problems with conventional HMM: To compensate state output, say, S_t for long reverberation, the preceding sequence of outputs $(\ldots \boldsymbol{S}_{t-4}, \boldsymbol{S}_{t-3}, \boldsymbol{S}_{t-2}, \boldsymbol{S}_{t-1}, \boldsymbol{S}_t)$ is required, which cannot be known with this configuration of HMM. Further, compensations for same state, e. g., 3 at time $t-1$ and t will be different, which cannot be modeled by static output distribution function used in this HMM.

Eq. 12 shows that the spectral parameters of corrupted speech at frame t do not depend only upon this frame, but also upon the preceding frames at $t-1$, $t-2$ and so forth. Therefore, to adapt the output distribution $\boldsymbol{b}_j$ at state $q_t = j$ of given HMM, the frames occured at time $t-1$, $t-2$ and so forth should be considered. However, with the conventional HMM [Fig. 1] used in most of speech recognition systems, state sequence is hidden, and the preceding occurrences of states are not known for a given state (The adaptation being considered is before recognition begins and observations become available; this excludes even the possibility of estimating the most likely state sequence). In such condition, possible preceding observations or output densities cannot be known. Further, compensation required for the same state j will be different at time $t, t+1, t+2$ etc., as self-transition loop is executed repetitively and state sequence changes.

Therefore, two basic problems need to be addressed for such a model convolution:

1. Estimating state sequence of *sufficient* length.
2. Way of updating output probability distribution for a state.

We present the solution in next section, along with the required approximations.

3. HMM State Splitting

To enable the prediction of preceding states, the conventional HMM of Fig. 1 is transformed into a split-state HMM as shown in Fig. 2 by splitting each states into optimum number of substates. It is to be noted that only the last substate has self-transition loop now, and the HMM turns into a multipath one as well. The transition probability from a substate to itself or another substate of its own parent state i is taken equal to self-transition probability a_{ii}, whereas from a substate of state i to a substate of state j, it is taken as a_{ij}. In this way, transition probabilities

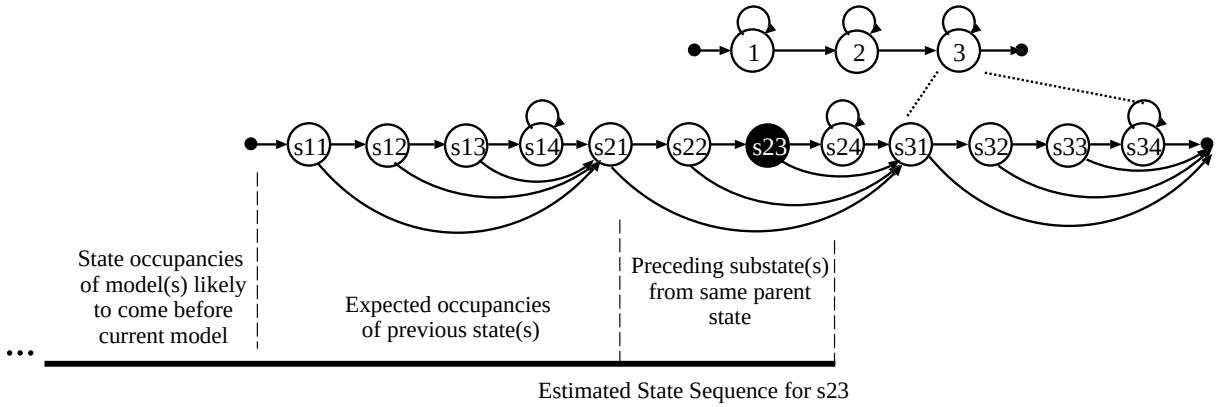

Fig. 2 State Splitting of HMM: Once states are split into such configuration, the preceding state sequence can be partially known. For example, while considering substate $s23$, we know that substates $s22$ and $s21$ must precede. Beyond that, substates of state 1 occurs, however as the exact number of occurrence of state 1 cannot be be found due to self-transition loop, we take average number of occupancy of state 1, as given by its duration density.

for all the paths of the split-state HMM can be computed. The output distributions of each substate is initialized to be equal to that of its parent state.

The number of substates under state i is taken proportional to expected duration $\bar{d}_i$ of the state, and are not necessarily kept equal. When duration of states has not been explicitly modeled, the inherent duration density is used for computing the expected duration under a state. The inherent duration probability $p_i(d)$ associated with state i, with self-transition coefficient a_{ii} is

$$p_i(d) = (a_{ii})^{d-1}(1 - a_{ii}). \tag{13}$$

With such exponential duration density, the expected number of observation under state i is

$$\bar{d}_i = \frac{1}{1 - a_{ii}}. \tag{14}$$

Once the split-state HMM is obtained, it can provide solutions to both of the problems mentioned in Section 2, up to some extent. First, the structure of split-state HMM enables the estimation of preceding states up to some finite length. For example, for substate 23 in Fig 2, the last two states must be $s22$ and $s21$. Beyond that, substates of state 1 occurs, but their numbers cannot be exactly ascertained, and average number of occupancy of state 1 is taken. If average occupancy of state 1 is 6, then preceding state sequence for substate 23 is (6 number of state 1, 2 number of state 2). During model-convolution, the states occurring at nearer position, say at $t-1$, $t-2$, $t-3$ etc., to current state at t are crucial than the states at farther position; and their accurate prediction is very important. State splitting does provide nearer state occupancies exactly, whereas at farther frames, it takes the average number of occupancies of states.

Secondly, as each state has been expanded into a number of substates with no self-loop except at last substate, no substate except last one can occur twice, and the need for different compensations or dynamic output distribution function has been eliminated. As probability of occurring last state is very low, the resulted error will be of much small scale. Each substate essentially provides a way to store different compensations required for the same state.

4. Algorithm

The algorithm for model convolution by state splitting (also shown in Fig. 3) is given as:

1. Split states into *optimum* number of substates.
2. Find transition probability matrix for new HMM.
3. For each substate, estimate a sequence of preceding states $\boldsymbol{q} = (\ldots q_{t-3}, q_{t-2}, q_{t-1}, q_t)$.
4. Find output probability distribution for current state by convolving given $\bar{\boldsymbol{H}}$ and output density sequence $(\ldots \boldsymbol{S}_{t-3}, \boldsymbol{S}_{t-2}, \boldsymbol{S}_{t-1}, \boldsymbol{S}_t)$ associated with $\boldsymbol{q}$.
   - (a) Transform clean speech output densities from cepstral domain to mel-filterbank (linear) domain.
   - (b) Take STFT of $h[m]$ and find $\bar{H}_k(t)$.
   - (c) Convolve clean speech densities with $\bar{H}_k(t)$ and find distribution for corrupted speech.
   - (d) Transform corrupted speech parameters back to cepstral domain.

5. Evaluation

For the evaluation of state splitting approach, it was tested on a speaker-dependent isolated word recognition task. The clean speech HMM was trained with 2620 words of the same speaker taken from ATR speech database A-Set. The clean speech HMM comprised of 41 context-independent phoneme models,

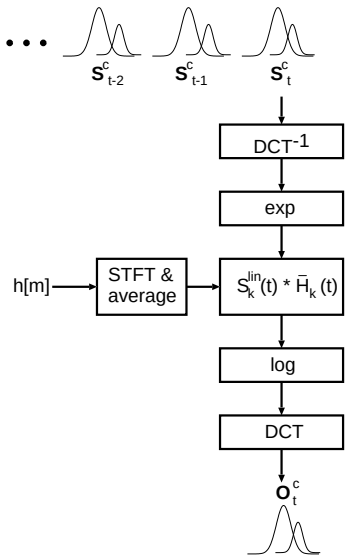

Fig. 3 Adaptation of model parameters by convolution of distributions in split-state HMM: Estimated preceding sequence of output densities are transformed to mel-domain and convolved with averaged spectral parameter $\bar{\boldsymbol{H}}$ of given impulse response and converted back to cepstral domain.

each with three emitting states single mixture Gaussian model initially. The speech signal was single channel with sampling frequency of 16 kHz. The speech signal was analyzed with Hamming window of 25 ms frame length and frame shift of 10 ms into 13-dimensional MFCC_0 feature vectors. The number of mel filterbank was 24. The test set consisted of 655 words of the same speaker taken exclusively from the ATR speech database A-set. The decoder used was Julian3.3p3 Multipath version [15]. The word accuracy for clean speech with the clean model was 93.1%.

To conduct the test for reverberant speech, it was simulated by a linear convolution of clean speech and impulse response. The impulse responses were taken from RWCP Sound Scene Database in Real Environment [16]. The performance for the reverberant speech with clean model degraded as listed under "clean" in Table 1 for different impulse response. The recognition performance of reverberant speech was evaluated with Cepstrum Mean Subtraction (CMS) method also. For this purpose, CMS was performed on the same training set data, and the model was retrained with it. CMS was applied to test set also, and performance was evaluated with the retrained model. The word accuracy for CMS is also shown in Table 1 under "CMS".

To evaluate state splitting approach, each emitting states of models were split into 20 substates and transition probabilities were updated as described in Section 3 and implicit duration density was used for estimating average state durations. Same impulse response as used for simulating reverberant speech was used to compute averaged spectral parameters $\bar{H}_k(t)$.

Table 1 Experimental Results (Word Recognition Rate %)

Model	Clean	CMS	SS
E1A ($T_{60} = 0.12s$)	30.1	39.8	52.1
E1B ($T_{60} = 0.31s$)	27.8	16.5	34.6

For state sequence estimation, frames coming from other preceding models were not considered, but only the frames from the same model were taken into account. Further, only mean vectors were adapted by this approach. The experimental results for different impulse responses are listed in Table 1 under "SS".

For E1A, the method has better performance than CMS that proves its effectiveness. For E1B, which has relatively longer reverberation time, the performance of CMS has degraded; and the state-splitting approach also has low improvement in performance. This can be attributed to fact that the estimated speech frames for convolution is not of sufficient length, whereas the long reverberation time requires longer sequence of preceding speech frames to be considered. Given some means to predict or estimate the past frames generated by other models preceding to current model, the performance of ASRs under such case can be improved significantly. Furthermore, with good estimate for state duration, and proper approximation for density estimation of corrupted speech, the technique is likely to perform better. Also, the number of substates can be reduced to large extent with good estimate for model duration and using variable number of substates.

6. Future Work

Current work uses implicit exponential duration density [Eq. 13] for states, however such an exponential duration density is inappropriate [17] for most of the physical signals, and therefore it is preferable to model duration density explicitly in some analytical form, as done in [18] and [19]. Then, the computation of expected number of occupancy of a state will be more accurate than using the inherent exponential distribution to compute it. Furthermore, such explicitly modeled duration density, say, using Gamma distribution can be fit into the split-state HMM by computing transition probabilities using that distribution.

Another viable alternative to state-splitting is to modify duration-density HMM itself, by using different output distribution b_{jk} for each kth occurrence in a particular state j, up to the maximum duration value d_{max} (Fig. 4). Yet another alternative is to eliminate conditional independence assumption of HMM, and use dynamic distributions for states, that will depend not only on the state that generates it, but also on the previous states. The distribution will

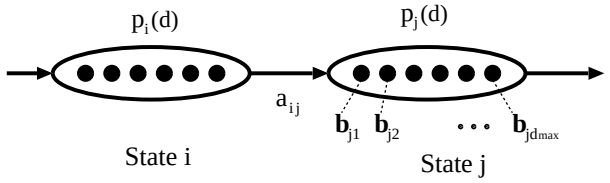

Fig. 4 Duration HMM with duration probability density $p_j(d)$ for state j, and each substate k of state j having different output probability distributions b_{jk}: With such HMM, the preceding substates from own state are available, and beyond that average occupancies of states are taken for the preceding sequence .

also change every time the self-transition loop is executed. These approaches will be presented in other papers.

Though the method has been applied to HMM modeling MFCC parameters with Gaussian mixture models (GMMs), it can be used with parameters other than MFCC and with other mixture models. With GMMs the number of mixtures in the compensated model may be large, and merging of mixtures would prove useful. Further, use of discrete mixture model instead of GMMs can avoid some of the assumptions and complexities involved in handling GMMs.

The state splitting can accurately give some crucial preceding states or output densities, however in case of very long reverberation, still longer output sequence would be necessary, and effects from preceding phones or models should be taken into account (Fig. 2). Use of triphones can predict longer state sequence. For still longer sequence, information from n-gram language models can be used to account the effect from preceding phonemes.

7. Conclusion

In this paper, we proposed a technique for model adaptation for reverberant speech based on state-splitting of HMM, and presented the expressions and approximations required for it. The experimental results proved the effectiveness and potential of the method.

The method has many possibilities for extensions. Future work includes applying the model convolution approach to explicitly modeled duration density HMM, as well as effective estimation of past frames contributed by preceding models. The use of dynamic output distribution functions will be also investigated.

References

[1] X. Huang, A. Acero, and H.-W. Hon, *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*, Prentice Hall PTR, New Jersey, 1st edition, 2001.

[2] A. Acero, *Acouctical and Environmental Robustness in Automatic Speech Recognition*, Ph.D. thesis, Carnegie Mellon University, 1990.

[3] M. Miyoshi and Y. Kaneda, "Inverse filtering of room acoustics," *IEEE Trans. on ASSP*, vol. 36(2), 1988.

[4] H. Wang and F. Itakura, "An approach of dereverberation using multi-microphone sub-band envelope estimation," in *Proc. ICASSP*, 1991, pp. 953–956.

[5] C. Avendano, S. Tibrewala, and H. Hermansky, "Multiresolution channel normalization for ASR in reverberant environments," in *Proc. Eurospeech*, 1997, pp. 1107–1110.

[6] B. S. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," *J. Acoust. Soc. Am.*, vol. 55, pp. 1304–1312, 1974.

[7] B. E. Kingsbury and N. Morgan, "Recognizing reverberant speech with RASTA-PLP," in *Proc. ICASSP*, Munich, Germany, 1997, pp. 1259–1262.

[8] T. Takiguchi and M. Nishimura, "Acoustic model adaptation using first-order linear prediction for reverberant speech," in *Proc. ICASSP*, 2004, pp. 869– 872.

[9] T. Takiguchi, S. Nakamura, and K. Shikano, "HMM-separation based speech recognition for distant moving speaker," *IEEE Trans. on Speech and Audio Processing*, vol. 9(2), pp. 127–140, 2001.

[10] H. Yamamoto, T. Nishimoto, and S. Sagayama, "Frame-by-frame HMM adaptation for reverberant speech recognition," in *Proc. Special Workshop in Maui (SWIM)*, Jan. 2004.

[11] A. Baba, A. Lee, H. Saruwatari, and K. Shikano, "Speech recognition by reverberation adapted acoustic models," in *Proc. ASJ*, Sep 2002, pp. 27–28.

[12] J.-C. Junqua and J.-P. Haton, *Robustness in Automatic Speech Recognition: Fundamentals and Applications*, Kluwer Academic Publishers, Massachusetts, USA, 1st edition, 1996.

[13] Y. Suzuki, F. Asano, H.-Y. Kim, and Toshio Sone, "An optimum computer-generated pulse signal suitable for the measurement of very long impulse responses," *J. Acoust. Soc. Am.*, vol. 97(2), pp. 1119–1123, 1995.

[14] M. J. F. Gales and S. J. Young, "Robust continuous speech recognition using parallel model combination," *IEEE Trans. on Speech and Audio Processing*, vol. 4, pp. 352–359, 1996.

[15] *Multipurpose Large Vocabulary Continuous Speech Recognition Engine Julius rev. 3.2*, Nara Institute of Science and Technology, 2001, Available: http://julius.sourceforge.jp/.

[16] Real World Computing Partnership and Real Acoustic Environments Working Group, *RWCP Sound Scene Database in Real Acoustical Environments*, Mitsubishi Research Institute, Inc., 2001.

[17] L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*, Prentice Hall International, Inc., 1st edition, 1993.

[18] J. D. Ferguson, "Hidden Markov analysis: An introduction," in *Hidden Markov Models for Speech*, Institute for Defence Analyses, Princeton, NJ, 1980, pp. 8–15.

[19] S. E. Levinson, "Continuously variable duration hidden Markov models for automatic speech recognition," *Computer Speech and Language*, vol. 1(1), pp. 29–45, 1986.