FRAME-BY-FRAME HMM ADAPTATION FOR REVERBERANT SPEECH RECOGNITION

Hitoshi Yamamoto, Takuya Nishimoto, and Shigeki Sagayama

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

ABSTRACT

This paper describes a technique for robust speech recognition under a reverberant environment. Acoustic signals are distorted by their reflection inside a room, and this reverberation degrades the performance of automatic speech recognition system. When the effect of reverberation spans for several frames, dealing such the distortion due to reverberation simply as convolutive distortion is not sufficient. In such cases, it becomes necessary to consider the reverberation effect of other frames over current frame. In this paper, frame-synchronous estimation of reverberation component and HMM combination are proposed to deal with such distortion. The proposed technique was evaluated with an isolated word reverberant speech recognition task, and the result showed improvement in recognition performance. Moreover, use of Jacobian adaptation for a frame-synchronous model adaptation is investigated.

1. INTRODUCTION

In adverse environments, existence of noise and reverberation causes acoustical mismatch between the training and the testing conditions of hidden Markov model (HMM), and the recognition performance of automatic speech recognition (ASR) system degrades [1, 2]. For additive noise, speech enhancement techniques such as spectral subtraction and model adaptation techniques such as parallel model combination (PMC) [3] are well known. In case of convolutive distortion such as microphone or channel distortion, several techniques, for example cepstrum mean subtraction (CMS) [4], exist. However, in reverberant environment, it is difficult to handle the distortion caused by reverberation simply as convolutive one. The reason is that there is reflection from walls and objects inside a room in addition to the direct sound from sound source, and the reflection (reverberation component) at one frame differs from others'. The reverberant effect often spanning for several frames, and therefore the reverberation component from outside frames have to be considered.

There are many signal processing techniques for estimating reverberant-free speech from reverberant speech, but when the features are based on short-time spectra, the features of reverberant-free speech estimated by these methods will be distorted. Feature compensation method using multi-resolution analysis [5] sometimes distort features of clean speech. With of acoustic model compensation methods [6, 7], it is difficult to cope with the distortion when the reverberation time is too long.

This paper addresses the reverberant speech recognition problem when only the speech models and the acoustical transfer characteristics of room are given to ASR system. In this problem, if an acoustic model that matched with reverberant speech is given at every frame, the ASR system will have robustness to the reverberant room environment. This paper elaborates convolutive and additive components of distortion due to reverberation. Further, dynamic adaptation of phoneme models to reverberant environment is also investigated.

2. REVERBERANT SPEECH RECOGNITION

Here we consider reverberant speech recognition task using features based on the short-time spectra and phoneme models trained with the clean speech.

2.1. Reverberant Speech

Acoustic signals are distorted by reflection from wall inside a room as depicted in figure 1. This phenomenon is observed as reverberation or echo. The reflection of one speech masks the following, and this effect degrades the accuracy of ASR system. In a continuous speech recognition task, it has been reported that when the reverberation time of a room is longer than 0.4 second, even reverberation-matched models cannot improve the recognition rate sufficiently [6].

In many conventional research, the speech enhancement techniques, for example inverse filtering or microphone arrays, are used to reduce the effect of reverberation. However, these techniques mainly focus on recovering the speech signal with good perceptual quality and intelligibility [5], and the short-time spectra of recovered speech are distorted.

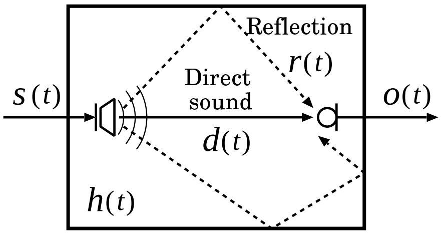


Fig. 1. Acoustical path inside a room.

So these techniques are inappropriate for improving the recognition performance significantly as a front-end of ASR system.

There are some researches that focus on reverberant speech recognition. Feature compensation method based on multi-resolution analysis [5] is able to cope with reverberant speech, but this technique distorts clean speech in contrast. Model selection from several reverberation-matched model [6] improves recognition performance well, but plural models have to be trained and there is possibility of selecting wrong models. HMM separation method [7] that is based on the modeling of acoustic transfer characteristics doesn't deal with inter-frame effect, and it doesn't work well when the reverberation time is long. Recent researches describe the effect of outside frames on current frame by linear prediction from preceding observation sequence [8].

2.2. Length of Reverberation

ASR problem under reverberant environments can be divided according to the acoustic transfer characteristics from sound source (speaker) to observation point (microphone). Acoustic characteristics of a room are assumed as a linear system here. Then two categories are considered according to the relationship between the length of impulse response (reverberation time (RT)) and the width of analysis frame (or the duration time of HMMs' states or phonemes):

1. **Short RT**: The effect of transfer characteristics of room appears in the variation of spectra within each frame. In this case, the distortion can be handled by conventional techniques such as channel adaptation.

2. **Long RT**: The effect spans several frames. The reflection of preceding frames' speech effect on current frame's speech, so mismatch between input speech and trained acoustic model appears. At this time, it is difficult to adapt the models for the reverberant environment.

Generally, it is adequate to consider the reverberant room environments as the latter case. For example, RT of a meeting room is about 0.6 second while the analysis frame width is of 10ms order.

2.3. Separation of Distortion

Here we again consider the distortion of speech features caused by reverberation. Both direct sound and its reflection exist in reverberant environment. So it is reasonable to cope with each distortion separately. For the direct speech, there is no necessity of compensation when only time delay is assumed. Even when there are distortions of spectra, conventional techniques such as channel adaptation are suitable to recognize the direct speech. On the other hand, the reverberation component varies momentarily according to preceding signal and affect on current frame additively. Therefore, these distortion cannot be regarded as constant one for each frames. To accomplish a model adaptation in such condition, it is necessary to consider the context and estimate reverberation component dynamically.

In this work, the reverberation component from outside frames is considered as an additive noise, and HMMs are composed at every frame.

3. FRAME-SYNCHRONOUS HMM COMPOSITION

3.1. Expression for Acoustic Transfer Characteristics in Terms of Short-time Power Spectra

Reverberation of a room can be described with an impulse response. In figure 1, observed speech signal at time t, $o(t)$, is expressed as

$$o(t) = h(t) * s(t) \tag{1}$$

where $s(t)$ is the speech signal, and $h(t)$ is the impulse response of the room.

When RT is relatively short compared with the frame length, the short-time power spectra of $o(t)$ at frame n, $O(\omega, n)$, is expressed as

$$O(\omega, n) = H(\omega) \cdot S(\omega, n) \tag{2}$$

where $S(\omega, n)$ and $H(\omega)$ represent the short-time power spectra of $s(t)$ and $h(t)$ respectively, at frequency ω. In case of MFCC, ω specifies the band in filter-banks, i.e. it can be filter number or the center frequency of specified spectra band. Most conventional techniques to cope with convolutive distortion are based on this equation and regard such distortion as unchanged one for all of the frames.

However, when RT is relatively long compared to the frame length, the speech of current frame, $S(\omega, n)$, is affected by preceding speech, $S(\omega, n-1), S(\omega, n-2), \cdots$. In such case, $O(\omega, n)$ can be expressed as

$$O(\omega, n) = H(\omega, n) * S(\omega, n) \tag{3}$$

This short-time power spectra response H expresses a power attenuation of every frequency-band (figure 2). Here it is assumed that H has a finite response length of L frames and the acoustic characteristics of room are described by H. This L corresponds to the reverberation time.

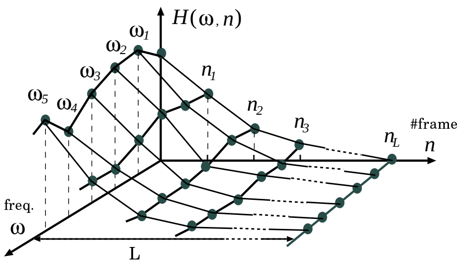


Fig. 2. Characteristics of reverberation expressed as sequence of short-time spectra.

3.2. HMM Composition with Reverberation Component

As mentioned in last section, reverberation component of one frame is due to the preceding frames' speech. The acoustic transfer characteristics H are assumed to be separated into two parts as

$$H(\omega,n) = D(\omega,n) + R(\omega,n) \tag{4}$$

where D and R are transfer characteristics of direct signal and reflected signal, respectively. Now, eq. (3) can be rewritten as

$$O(\omega,n) = D(\omega,n) * S(\omega,n) + R(\omega,n) * S(\omega,n) \tag{5}$$

The term of direct signal $D * S$ is clean speech affected by convolutive distortion D. When D is flat, clean speech model is fit for the direct signal. On the other hand, the term of reverberation component $R * S$ is convolution of R and the preceding speech S. This reverberation component denoted by N is considered as an additive noise that affects direct signal. This is expressed as

$$N(\omega,n) = \sum_{l=0}^{L} R(\omega,l)S(\omega,n-l) \tag{6}$$

In this paper, a method is proposed for combining the reverberation component N with speech HMMs at every frame. This frame-synchronous model composition, as depicted in figure 3, carries out dynamic model adaptation according to speech signal context.

This technique doesn't change the analysis procedure of speech, so there is little spectral distortion such as caused by conventional speech enhancement techniques and robust feature extraction techniques [5]. Further it is able to handle any fluctuation in reverberation component while HMM-separation method [7] doesn't account it. This method estimates reverberation component by convolution rather than by linear prediction [8]. However, the computational cost of proposed method is high.

3.3. Practical Algorithm

In practical situations, even if $H(\omega,n)$ can be known through a possible measuring method for transfer function from the sound source (speaker) to the observation point (microphone), it is often difficult to separate it into the direct path D and reflective (reverberant) path R components. Here we simply assume that non-zero parts of D and R do not overlap along time, and then approximate D by first few non-zero frames of the impulse response H and R by the rest of them. This assumption implies that the effect of D is limited within few frames and $D * S$ does not affect much the distant frames over phoneme boundaries.

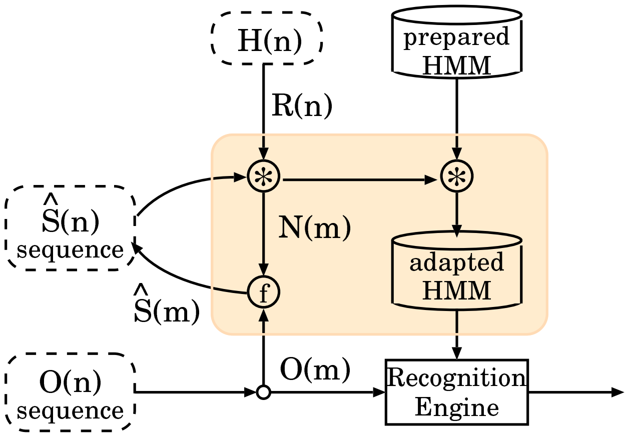


Fig. 3. Framework of frame-synchronous model adaptation technique.

In the beginning of an utterance, as there is no reverberant component from the preceding frames overlapping the present frame because of causality of R (i.e. $N = 0$), the observation O is regarded as $D * S$. Since D represents the direct path, $D * S$ may be considered to be identical to S with some time delay. We consider two cases: when D can be ignored ("proposed1") and when it can not be ("proposed2") where D is replaced by a direct path characteristic measured in an anechoic chamber.

Later at n-th frame, the additive reverberant component $N = R * S$ is removed from the observation O to estimate S by $\hat{S} = O - N$. Starting from the first frame with $N = 0$, this recursion is performed for every frame to estimate $\hat{S}$ by $O - R * \hat{S}$ using past estimate of S. Even though this calculation does not give an accurate estimate of S, it is still accurate enough for estimating the reverberation component $R * S$.

The reverberation component $R * S$ is combined frame by frame with the clean speech models ("proposed1") or with the direct path models ("proposed2") using a model composition method such as Parallel Model Combination (PMC).

The algorithm for MFCC-based recognition of reverberant speech is as follows:

Preparation :

1. Prepare HMMs of speech and H

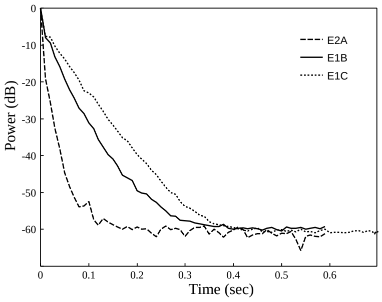


Fig. 4. Characteristics of impulse response used in evaluation.

Initialization :

1. Set D and R
2. Transform model parameters into filter-bank domain
3. Set $N(\omega, 0) = 0$, $S(\omega, 0) = O(\omega, 0)$

Composition : For each frame n do:

1. For each mel filter-bank ω do:
   (a) Estimate $N(\omega, n)$ (eq. (6))
   (b) Calculate $\hat{S}(\omega, n) = O(\omega, n) - N(\omega, n)$
   (c) Add $N(\omega, n)$ to the mean of each Gaussian
2. Transform model parameters into MFCC domain
3. Use the adapted HMMs to calculate likelihood of O

4. EVALUATION

To evaluate the proposed technique, we tested it on Japanese language based speaker-dependent isolated word speech recognition task in reverberant environments

4.1. Experimental Conditions

The test data was 655 words of each speaker taken from ATR Speech Database A-set. To simulate reverberant environment, three types of impulse responses from RWCP Sound Scene Database in Real Acoustical Environment were convoluted to speech data artificially. The distance between sound source and microphone was 2 meter in the database. The impulse responses used in the experiments are shown in figure 4.

The speech was digitized by 16 kHz sampling, and analyzed into 26-dimensional feature vectors with Hamming window of 25ms width with 10ms time-shift. The feature set included 13-dimensional MFCC (including 0'th coefficient) and their first order time derivatives. The number of mel filter-banks were 24. The acoustic models contained 41 context independent phoneme HMMs each having 3-state 4-mixture with continuous density. They were trained with 2,620 words (excluding testing data) of testing speaker. The decoder was Julian 3.3p3 that was modified for implementation of proposed method.

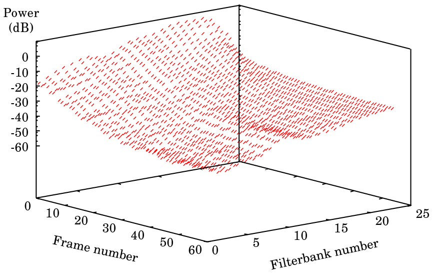


Fig. 5. Characteristics of impulse response in terms of short-time power spectra.

Acoustic characteristics H was short-time spectra sequence of impulse response identical to that convoluted with test speech. The analysis was same as that for speech. D was once assumed as flat and in other case approximated by the characteristics of anechoic chamber. In the former case, R was taken from H after 4th frame to consider frame overlap into account. In the latter case, R was taken from H after 8th frame (the length of D). The characteristics of impulse response "E1B" in terms of short-time power spectra is shown in figure 5.

4.2. Result

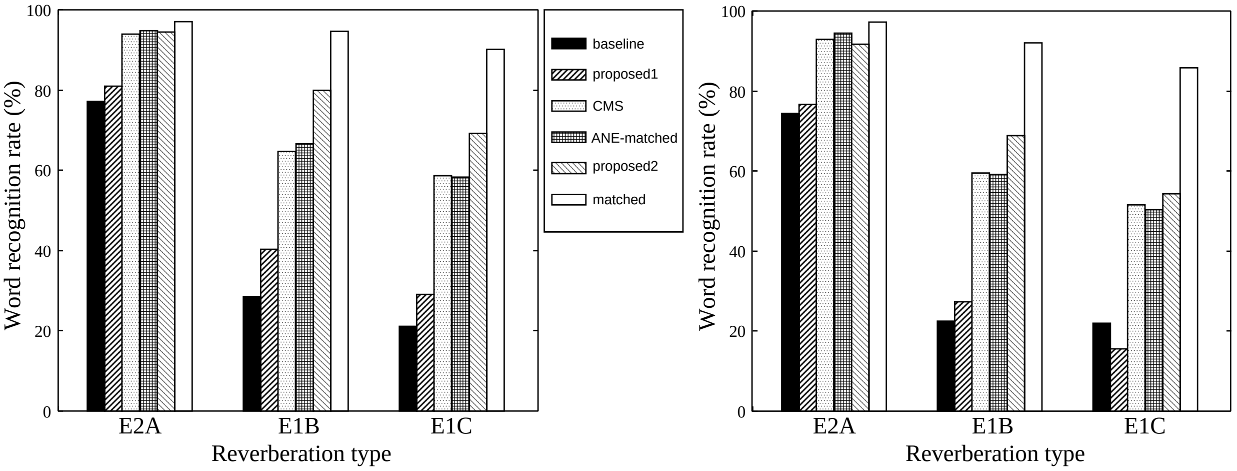
The recognition results are depicted in table 1 and figure 6. The speakers were FFS (female) and MAU (male). In figure 6, "baseline" method used the clean speech HMMs with no adaptation, "proposed1" applied proposed technique to the clean speech HMMs, "CMS" applied cepstrum mean subtraction, "ANE" used the HMMs trained with the data convoluted with the impulse response of anechoic room, "proposed2" applied proposed technique to "ANE", and "matched" used the model trained with data of the same reverberation conditions. In "proposed1" method, the transfer characteristics of direct speech D was approximated flat. In contrast, the impulse response of anechoic room was considered as D in "proposed2" method.

The result shows that word recognition rate of proposed technique is higher than the conventional methods except for E2A, where "CMS" and "ANE" performed well and the improvement of proposed method was little. In case of E2A, direct sound component was dominant as suggested by the reverberation curve depicted in figure 4. For E1B and E1C, "CMS" and "ANE" show almost same performance, and the "proposed2" method has better performance than both of them. The result shows that the proposed method is able to improve the recognition performance significantly against the distortion caused by reverberation.

Though the proposed method performs well, there is still room for improvement compared with "matched". Further works for improvement of techniques and better approximations need to be carried out. Further evaluation for longer reverberation time and continuous speech recognition will be considered in future work.

Table 1. Word recognition rate（%）for reverberant speech when the proposed method was applied.

	(a) Speaker: FFS(female)						(b) Speaker: MAU(male)					
	baseline	proposed1	CMS	ANE	proposed2	matched	baseline	proposed1	CMS	ANE	proposed2	matched
E2A	77.1	81.1	94.1	95.0	94.8	97.1	74.5	76.6	93.1	94.5	91.9	97.6
E1B	28.2	39.9	64.9	66.9	79.9	95.4	22.3	27.4	59.7	59.5	68.9	92.4
E1C	20.8	28.4	58.6	58.3	68.6	90.7	22.1	15.0	51.9	50.7	53.7	86.2



(a) Speaker: FFS(female)

(b) Speaker: MAU(male)

Fig. 6. Word recognition rate (%) for reverberant speech when the proposed method was applied.

5. JACOBIAN ADAPTATION

In this section, frame-synchronous Jacobian adaptation for reverberant speech recognition is investigated.

5.1. Jacobian Adaptation

Jacobian adaptation(JA) [9] is a fast acoustic model adaptation technique for a new acoustical environment. The feature of JA is that the models can be adapted from some noisy(not clean) environment to another noisy environment. In this technique, non-linear transformation of model caused by the fluctuation of environment is approximated as linear, and the model adaptation is carried out in cepstrum domain.

$$\boldsymbol{Y}^c_{tar} \simeq \boldsymbol{Y}^c_{ref} + \boldsymbol{J}_N(\boldsymbol{N}^c_{tar} - \boldsymbol{N}^c_{ref}) \qquad (7)$$

where $\boldsymbol{Y}^c_{ref}$ and $\boldsymbol{N}^c_{ref}$ are speech and noise (source of environment variation) of reference environment, $\boldsymbol{Y}^c_{tar}$ and $\boldsymbol{N}^c_{tar}$ are speech and noise of target environment, respectively. $\boldsymbol{Y}^c$ and $\boldsymbol{Y}^s$ represent vectors of cepstrum domain and spectra domain, respectively. Jacobian matrix $\boldsymbol{J}_N$ can be calculated from next equation,

$$\begin{aligned} \boldsymbol{J}_N &\equiv \frac{\partial \boldsymbol{Y}^c}{\partial \boldsymbol{N}^c} & (8) \\ &= \frac{\partial \boldsymbol{Y}^c}{\partial \log \boldsymbol{Y}^s} \frac{\partial \log \boldsymbol{Y}^s}{\partial \boldsymbol{Y}^s} \frac{\partial \boldsymbol{Y}^s}{\partial \boldsymbol{N}^s} \frac{\partial \boldsymbol{N}^s}{\partial \log \boldsymbol{N}^s} \frac{\partial \log \boldsymbol{N}^s}{\partial \boldsymbol{N}^s} & (9) \\ &= \boldsymbol{C} \frac{\boldsymbol{N}^s}{\boldsymbol{Y}^s} \boldsymbol{C}^{-1} \quad (\boldsymbol{C} : \text{cosine transform matrix}) & (10) \end{aligned}$$

5.2. Jacobian Adaptation for Reverberant Component

Here, JA is considered for proposed framework of reverberant speech recognition described in section 3. Reverberation component is supposed as noise, and model is adapted from some reverberant-matched condition instead of using model combination with clean speech model. Average reverberant component in reference environment $\boldsymbol{N}_{ref}$ and the Jacobian matrix $\boldsymbol{J}_N$ are computed beforehand. Acoustic models are adapted frame by frame according to eq. (7) using estimated reverberant component $\boldsymbol{N}_{tar}$. This technique can be called as frame-synchronous Jacobian adaptation (FSJA).

Jacobian adaptation can be applied to “matched” model or multi-condition trained model, so robustness against change in characteristics of reverberation can be expected in addition to good recognition performance. Moreover, the adaptation procedure is carried out in cepstrum domain at the recognition stage, so there is no necessity to transform model parameters to spectral domain as model composition technique, and the computational cost is reduced significantly.

5.3. Reverberation Component of Reference Environment

To adapt one reverberant-matched model to another reverberant environment, it is necessary to know the average reverberant component of reference environment $\boldsymbol{N}_{ref}$ in advance. However, $\boldsymbol{N}_{ref}$ cannot be observed, so it has to be estimated. For instance, if both database of speech and impulse response of room are given, $\boldsymbol{N}_{ref}$ can be simulated by a convolution in waveform domain. The estimation of reverberant component in terms of short-time spectra domain proposed in section 3 also can be applied. In other circumstances, when the reference reverberant component of each phoneme HMMs or each of their states are given, better model will be achieved. For example, difference between matched-model and clean speech model, and statistical information about context of speech given by database can be useful for estimation of reference reverberant component.

5.4. Algorithm

The frame-synchronous Jacobian adaptation algorithm for reverberant speech recognition is described here. Though estimated reverberant component N is transformed from spectra to MFCC, the computational cost of this method is less than that of transformation of model parameters.

Preparation :

1. Prepare speech HMM and acoustic transfer characteristics of room H
2. Calculate a priori reverberant component $\boldsymbol{N}_{ref}$ and Jacobian matrix $\boldsymbol{J}_N$

Initialization :

1. Set D and R (Separate H)
2. Set $N(\omega, 0) = 0$, $S(\omega, 0) = O(\omega, 0)$

Adaptation : For each frame n:

1. For each filter-bank ω:
   (a) Estimate N (eq. (6))
   (b) Calculate $\hat{S}(\omega, n) = O(\omega, n) - N(\omega, n)$
2. Transform N into MFCC domain
3. Model Adaptation (eq. (7))
4. Calculation of likelihood with adapted models

5.5. Preliminary Experiments

The algorithm was tested on the same speech recognition described in section 4. Reference reverberant component was estimated in same way as of N_{tar} and it was assumed to be equal for each phoneme HMMs. When adaptation was applied to matched model (trained by reverberant speech same as test data), the computational cost was largely reduced however sacrificing some accuracy. Therefore, FSJA could be an efficient technique for reverberant speech recognition, provided reference reverberant component can be accurately estimated.

6. CONCLUSION

In this paper, we proposed a new technique to adapt HMMs to reverberant environment. The technique estimates the current reverberation component from the preceding speech and combine them with HMMs at every frame. The proposed method was evaluated with recognition task of speaker-dependent isolated word speech simulated for the reverberant environment, and the results demonstrated improvement of recognition performance. Moreover, the use of Jacobian adaptation was investigated in the same framework.

The future works include the evaluation of proposed technique in various reverberation conditions and continuous speech recognition task. The estimation of reverberation characteristics for unknown environment, and the robustness issue for mismatch in reverberation conditions will be considered. Furthermore, estimete of reference reverberant component in JA framework, and frame-synchronous model adaptation technique for noise rather than reverberation one will be considered.

REFERENCES

[1] Y. Gong, "Speech recognition in noisy environments: A survey," *Speech Communication*, vol. 16, pp. 261–291, 1995.

[2] M. Omologo, "On the future trends of hands-free ASR: variabilities in the environmental conditions and in the acoustic transduction," *ESCA-NATO Workshop on robust speech recognition for unknown communication channels*, pp. 67–74, 1997.

[3] M. J. F. Gales and S. Young, "An improved approach to the hidden Markov model decomposition of speech and noise," *Proc. ICASSP92*, pp. 233–236, 1992.

[4] 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.

[5] C. Avendano, S. Tibrewala and H. Hermansky, "Multiresolution channel normalization for ASR in reverberant environments," *Proc. Eurospeech 97*, pp. 1107–1110, 1997.

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

[7] 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, no. 2, pp. 127–140, 2001.

[8] T. Takiguchi and M. Nishimura, "Reverberant speech recognition using a first-order linear prediction," *Proc. FIT2003*, pp. 111–112, Sep 2003, (in Japanese).

[9] S. Sagayama, Y. Yamaguchi, S. Takahashi and J. Takahashi., "Jacobian approach to fast acoustic model adaptation," *Proc. ICASSP97*, pp. 835–838, 1997.