Model Adaptation by State Splitting of HMM for Long Reverberation

Chandra Kant Raut, Takuya Nishimoto, Shigeki Sagayama

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

10.21437/Interspeech.2005-157

Abstract

In environment with considerably long reverberation time, each frame of speech is affected by reflected energy components from the preceding frames. Therefore to adapt model parameters of a state, it becomes necessary to consider these frames, and compute their contributions to current state. However, these clean speech frames preceding to a state of HMM are not known during adaptation of the models. This paper describes a method to estimate the preceding frames for a state in HMM, by splitting the state into a number of substates. The estimated sequence of frames can then be used to find reflected energy component for the state and compensate its parameters. The effectiveness of the method was confirmed by the experimental results on an isolated-word recognition task.

1. Introduction

Automatic speech recognition (ASR) systems, though usually trained with clean speech, have to operate under real-life environment for any practical purpose. 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. 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, 4]), channel normalization techniques (e.g., [2, 5]) including Cepstrum Mean Subtraction (CMS) [6] and RASTA [7]. Though these methods have been proved to improve the performance of ASRs, most of them cannot perform well when reverberation time is too long and additive noise is also present.

Model-based approaches like [8, 9, 10], 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 [11]. 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.

In our previous works [12, 13], we proposed a state splitting approach to deal with such long reverberation, that estimates preceding frames for a state of HMM and finds the compensated parameters by convolution of distributions. In this paper, we limit the discussion to mean-only adaptation, and for experiments, we use reflection coefficients estimated from adaptation data rather than using explicitly given channel parameters as in our previous experimental frameworks.

As the method works in model domain and HMMs are (re)adapted only when the channel characteristics changes significantly in contrary to other decoding-time framewise adaptation and front-end methods that work on the frame-by-frame basis, the method has low computational cost. Further, working in model domain makes the method less sensitive to deviation in channel parameters used for compensating the models than it would be in the feature domain. The power of the method lies in the fact that it allows compensation for additive noise, in any (feature or model) domain or both. Therefore once the model is adapted for reverberation, one or more of the enhancement or compensation techniques can be simultaneously applied for additive noise.

2. Effect of Long Reverberation

The effect of reverberation with reverberation time (T_{60}) longer than the analysis window-length, on the short-time Fourier transform of speech is approximated by

$$O(w_i, t) \approx H(w_i, t) * S(w_i, t) \tag{1}$$

where t is frame number, w_i is discrete frequency and * represents convolution along frame. Parameters $S(w_i, t)$, $H(w_i, t)$ and $O(w_i, t)$ are STFTs of clean speech $s[m]$, impulse response $h[m]$ characterizing reverberation/convolutional noise and distorted speech $o[m] = h[m] * s[m]$, respectively.

As impulse response of the environment is not directly given and its spectral parameters are not known, we represent mel-domain parameters for reverberant speech as

$$\begin{aligned} \hat{O}_k(t) &= \alpha_{k,0} S_k(t) + \alpha_{k,1} S_k(t-1) + \alpha_{k,2} S_k(t-2) \\ &+ \ldots + \alpha_{k,N-1} S_k(t-N+1) \end{aligned} \tag{2}$$

and estimate $\alpha_{k,i}$ for each mel filter-bank k from few seconds of adaptation data using minimum mean-squared error (MMSE) technique, by minimizing mean square error

$$\|e_k(t)\|^2 = E\Big(O_k(t) - \hat{O}_k(t)\Big)^2. \quad (3)$$

We call $\alpha_{k,i}$ as reflection coefficients and N is selected depending upon the reverberation time T_{60} of the testing environment. This estimation requires mel spectrum of speech signal and its reverberated counterpart. In practical case, adaptation data required for such estimation can be obtained through small amount of stereo recordings from close-talking and far field microphone. Besides, depending upon the situation, these coefficients can be estimated through other supervised or unsupervised methods as well.

This formulation for the effect of preceding frames on current one will be used for the model adaptation purpose in the current experimental framework.

3. Problems with Conventional HMM for Adaptation

Equations 1 and 2 show 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 on.

Therefore, to adapt the output distribution $\boldsymbol{b}_j$ at state $q_t = j$ of given HMM [Fig. 1], the frames occured at time $t-1$, $t-2$ and so on should be considered.

However, with such the conventional HMM used in most of speech recognition systems, nothing can be inferred deterministically about the observations, and even the state sequence preceding to a given state cannot be known (as adaptation under consideration is *not* decoding-time framewise adaptation and therefore does not use the observations, and in such case even most likely state sequence preceding to a state cannot be estimated).

Therefore, the philosophy adopted in the method is: (a) to use expected number of occupation of preceding states to estimate the preceding state sequence, and (b) use composite mean from the output distribution of the occured state as the observation for the frame. With this principle, if we are adapting state 3 and suppose expected occupations of state 1 and state 2 are four and two respectively, then preceding state sequence fot state 3 will be taken as $\{1, 1, 1, 1, 2, 2\}$ and preceding observations will be $\{\bar{\mu}_1, \bar{\mu}_1, \bar{\mu}_1, \bar{\mu}_1, \bar{\mu}_2, \bar{\mu}_2\}$, where $\bar{\mu}_i$ is the overall mean of the output density of state i.

However, there will be frames coming from own state as well, which are more important than farther frames. For example, when state 3 repeats for fifth time, there will be already four frames coming out of state 3 as preceding frames. These frames, as they are closer to current frame, are vital for estimating total reflected component, and should be included in the frame sequence.

Further, compensation required for the same state j in such HMM will be different at time $t, t+1, t+2$ and so on, as self-transition loop is executed repetitively and state sequence changes. In the same example of adapting state 3, the preceding state sequences for state 3 at different times are:

state sequence	time	repetition
1,1,1,1,2,2	t	1st occurance of state 3
1,1,1,1,2,2,3	t+1	2nd occurance
1,1,1,1,2,2,3,3	t+2	3rd occurance
1,1,1,1,2,2,3,3,3	t+3	4th occurance
1,1,1,1,2,2,3,3,3,3	t+4	5th occurance, and so on.

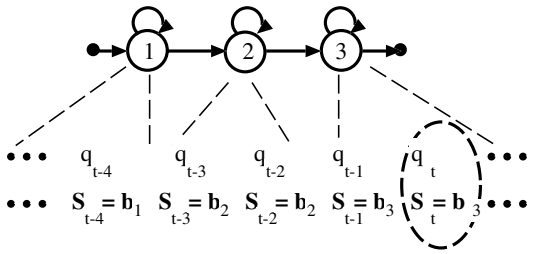

Figure 1: Problems with conventional HMM structure: To compensate state output, say, S_t for long reverberation by adding the reflected energy components from previous frames, knowledge of clean observations at time $t-1$, $t-2$ and so on is required, which cannot be inferred with this configuration of HMM. Further, compensations for same state, e.g., state 3 at time $t-1$ and t will be different, which cannot be modeled separately by static output distribution function used in such HMMs.

As with the every repeated occurance of state 3, preceding state sequences are different, there will be different values of compensation required at these different times. There is no way to accomodate these different compensations into the single state of traditional HMM with static output distributions.

Therefore, besides the propositions made, two other basic problems need to be addressed to estimate the reflected components from preceding frames effectively and adapt the models:

1. Accouting the frames coming from own state into the estimated sequence.
2. Accomodating different compensation values required for the single state.

4. HMM State Splitting

To address issues mentioned in Section 3, the conventional HMM of Fig. 1 is transformed into a split-state HMM as shown in Fig. 2 by splitting each states into a number of substates. It should noted that only the last substate has self-transition loop in the trasformed HMM, and it 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 for all the paths of the split-state HMM can be computed. The output distribution 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 occupation $\bar{d}_i$ of the state in the way to minimize execution of self-transition loop, and is constrained by compromise between complexity/speed and accuracy. When duration of states has not been explicitly modeled, the inherent duration density is used for computing the expected occupation of a state. The inherent duration probability $p_i(d)$ associated with state i, with self-transition coefficient a_{ii} is

$$p_i(d) \quad = \quad (a_{ii})^{d-1}(1 - a_{ii}). \quad (4)$$

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

$$\bar{d}_i = \frac{1}{1 - a_{ii}}. \quad (5)$$

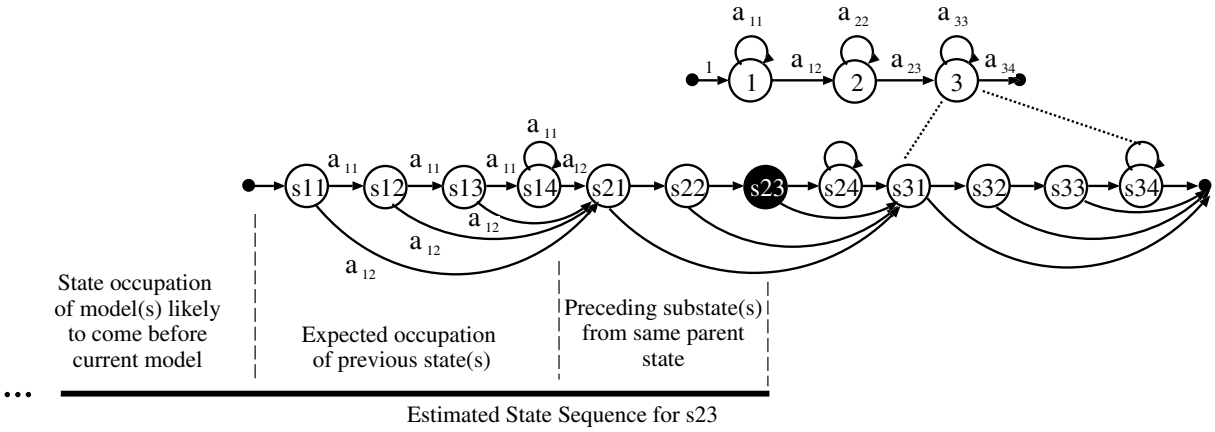

Figure 2: State Splitting of HMM: Once states are split into such configuration, number of frames coming from own state can be exactly known (except for last substate). Further, each substate acts as storage for different compensation values required for the same state.

In such split-state HMM, both of the problems mentioned in Section 3 have been avoided, except at the last substate, which still has self-transition loop. However, the state is expanded into substates in such a way that the probility of repeated occurances of last substate is minimized and the error caused by it will be very low.

First, due to the structure of split-state HMM and avoidance of self-transition loops from states, there is no more ambiguity for number of frames coming from the same state. The preceding frames generated by the same state is essentially modeled by substates and is easily accounted for the sequence. For example, for substate 23 in Fig 2, the preceding frames generated by same (parent) state are $s22$ and $s21$. In this way, the frames contributed from the same state can be accounted, in definite way, for estimation of preceding sequence. With this model, now the estimated state sequence for adapting substates, e.g. $s21$, $s22$ and $s23$, will be $\{1,1,1,1\}$, $\{1,1,1,1,2\}$, and $\{1,1,1,1,2,2\}$, respectively. For $s24$, though not alway true due to self-transition loop, it is still taken as $\{1,1,1,1,2,2,2\}$.

Further, the frames contributed by preceding *model* can be also accounted by taking expected occupation of its states in the estimated state sequence. The knowledge about the preceding models can be obtained by using context-dependent models.

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. Each substate essentially provides a way to store different compensations required for the same state.

With this framework, for an estimated state sequence of length N, the mean of a state is adapted by

$$\begin{aligned} \mu_O^{lin}(t) &= \alpha_{k,0}\mu_S^{lin}(t) + \alpha_{k,1}\bar{\mu}_S^{lin}(t-1) \\ &+ \ldots + \alpha_{k,N-1}\bar{\mu}_S^{lin}(t-N+1) \end{aligned} \quad (6)$$

where superscript lin represents linear mel-domain parameters and $\bar{\mu}_S^{lin}$ represents composite mean (for multimixture case) corresponding to distribution of the occured state in the sequence.

The algorithm for the method is depicted in Fig. 3. Once the split-state HMM is composed, and its distribution and transition matrix are initialized, preceding state sequence for each state is estimated. The composite means corresponding to output distribution of preceding state sequence are transformed to mel-domain, and adapted value for the state is computed. The adapted parameters are transformed back to cepstral domain by transforming to log-domain and applying discrete Cosine transform (DCT).

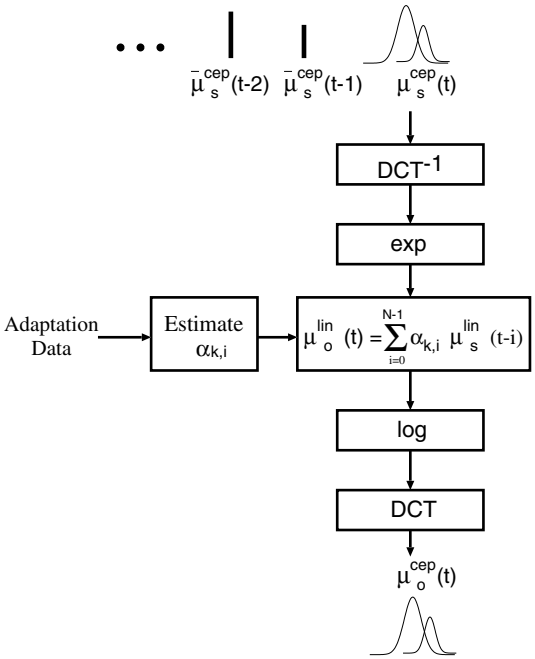

Figure 3: Adaptation of model parameters after estimating state sequence and taking composite means of corresponding output densities as preceding frames

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, each with three emitting states single mixture Gaussian model

Table 1: Experimental Results (Word Recognition Rate %)

Model	Clean	CMS	MASS
Clean Speech	93.1	–	–
Reverberated Speech	30.1	39.8	63.5

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 using 24 mel filter-banks. The test set consisted of 655 words of the same speaker taken exclusively from the ATR speech database A-set and Julian3.3p3 Multipath version [14] was used as decoder.

For testing, reverberant speech was simulated by a linear convolution of clean speech and impulse response (E1A) with reverberation time of 120 ms taken from RWCP Sound Scene Database in Real Environment. The performance for the reverberant speech with clean model degraded as listed under "clean" in Table 1. 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 four substates and transition probabilities were updated as described in Section 4 and implicit duration density was used for estimating average state occupations. About 12 s of speech signal and its simulated reverberant signal were analyzed into frames of 24 mel filter-bank parameters (with same windowing and frame-shift) and used for estimating coefficients $\alpha_{k,i}$ with $N = 10$. For state sequence, frames coming from preceding models were not considered. Further, only mean vectors were adapted by this approach. The result of model adaptation by state splitting (MASS) approach is listed in Table 1 under "MASS".

The method has better performance than clean model and CMS that proves its effectiveness, however, there is still much room for the improvement. The long reverberation time requires to consider the masking effect of preceding phonemes on following ones as well and compensate for it. Further, the accurate estimate of reflection coefficients $\alpha_{k,i}$ is very important for modeling the effect of reverberation on speech parameters. Given the better estimate of reflection coefficients $\alpha_{k,i}$, and by considering effect of preceding models as well, the performance of ASRs can be further improved.

6. 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.

Future work includes effective estimation of past frames contributed by preceding models using context-dependent models and accurate estimation of reflection coefficients, as well as evaluation of the method for joint compensation for additive noise and reverberation.

7. 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 derevereration 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] Gales, M. J. F., *Model-Based Techniques for Noise Robust Speech Recognition*, Ph. D. Thesis, Cambridge University, 1995.

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

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

[12] Raut, C. K., Nishimoto, T., Sagayama, S., "Model convolution by state splitting of HMM for robust speech recognition in presence of convolutional Noise," in *Proc. ASJ*, 3-5-5, pp. 85-86, 2005.

[13] Raut, C. K., Nishimoto, T., Sagayama, S., "Model adaptation for reverberant speech by HMM state splitting and convolution of distributions," IEICE Technical Report, vol. 104, no. 631, SP2004-151, pp. 37-42, 2005.

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