Model Convolution by State Splitting of HMM for Robust Speech Recognition in Presence of Convolutional Noise *

◎ Chandra Kant Raut Takuya Nishimoto Shigeki Sagayama

Graduate School of Information Science and Technology
The University of Tokyo
{raut, nishi, sagayama}@hil.t.u-tokyo.ac.jp

1 Introduction

Automatic speech recognition (ASR) system trained with clean speech performs poorly when speech is corrupted due to reverberation, channel distortion or microphone characteristics, commonly termed as convolutional noise. Several techniques like inverse filtering, microphone array based techniques, channel normalization techniques like cepstrum mean subtraction (CMS) and relative spectral processing (RASTA), and model-based approaches have been developed to cope with it. Though these methods have been proved to improve the performance of ASRs, not much success has been attained in handling long convolutional noise, and specially when additive noise is also present.

In this paper, we present a state splitting approach to adapt HMMs to reverberant environment with given characteristics. The method considers long reverberation time and allows flexible compensation for additive noise as well.

2 Effect of Long Reverberation

The effect of long convolutional noise on the short-time Fourier transform of speech is given by

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

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. From Eq. 1, the kth mel filterbank output is given as

$$O_k(t) \approx \sum_{\forall w_i} m_k(w_i)\big[H(w_i,t) * S(w_i,t)\big] \tag{2}$$

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

$$\begin{aligned} O_k(t) &\approx \sum_{\forall w_i} m_k(w_i)\big[H(w_i,0)S(w_i,t) \\ &\quad + H(w_i,1)S(w_i,t-1) + \dots\big] \qquad (3)\\ &\approx \bar{H}_k(0)S_k(t) + \bar{H}_k(1)S_k(t-1) + \dots \qquad (4)\\ &\approx \bar{H}_k(t) * S_k(t) \qquad (5)\end{aligned}$$

where we take

$$\bar{H}_k(t) = \frac{\sum_{\forall i} m_k(w_i)H(w_i,t)}{\sum_{\forall i} m_k(w_i)}. \tag{6}$$

*畳み込み残響に頑健な音声認識のための状態分割を用いるモデル適応, チャンドラ カント ラウト, 西本 卓也, 嵯峨山 茂樹 (東大・情報理工)

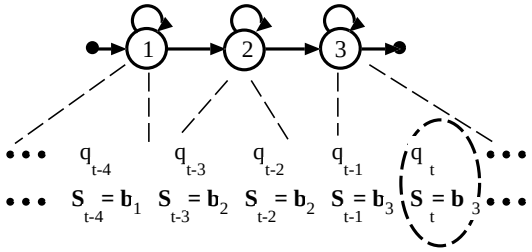

Figure 1: Given HMM for adaptation

Equations 1∼5 show that the spectral parameter of corrupted speech, at frame t, is convolution of spectral parameter of clean speech and channel parameter, and thus does not depend only upon frame at t, but also upon the preceding frames of speech at $t-1$, $t-2$ and so on.

3 Model Convolution by State Splitting

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, and be convolved with channel parameters. However, with such the conventional HMM used in most of speech recognition system, nothing can be inferred deterministically about the observations, and not even the state sequence preceding to a given state (as adaptation being considered is before observations become available). Therefore, the philosophy adopted in the method is to use expected number of occupancy of preceding states to estimate the preceding frame sequence and use observation density function of the states instead of observations for the sequence.

Secondly, compensation required for the same state j in such HMM will be different at time $t, t+1, t+2$ etc., as self-transition loop is executed repetitively and state sequence changes. Therefore, to accommodate these different compensated values for the single state, it is split into optimum (regarding efficiency of decoder) number of substates (proportional to duration of state) as shown in Fig. 2, such that repeated occupancy of a state is reduced. The transition probability from a substate to another substate of its own parent state i or to itself 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}. The output distribution of each substate is initialized to be equal to that of its parent state.

Once such the split-state HMM is obtained, preceding state sequence and corresponding densities are obtained for each state, for convolution with

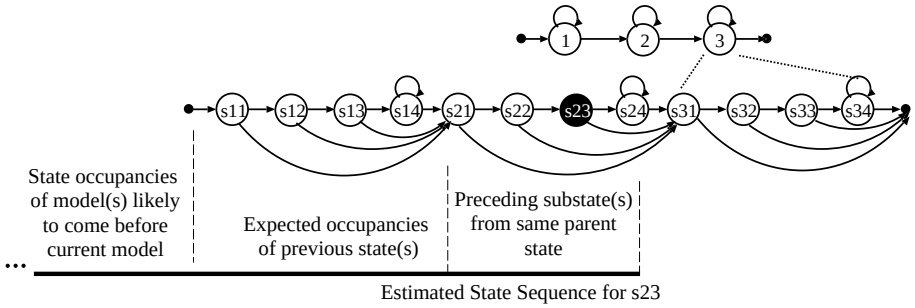

Figure 2: Model Convolution by State Splitting

spectral parameters of channel. For example, for substate 23 in Fig 2, the last two states must be $s22$ and $s21$. Beyond that occupancy of state 1 and other states from preceding models (e.g., known by bigram models) are taken for the sequence.

For the convolution, however, as Gaussian modeled cepstral parameters transformed to mel-domain are no more Gaussian (rather log-normal), the convolution involving a number of additions of densities is approximated with the assumption that sum of two or more log-normally distributed variables is still log-normal (as done in log-normal approximation of Parallel Model Combination [1]). Alternatively, instead of using mixtures, only (composite) mean vectors of the estimated sequence of mixtures can be used to estimate the distributions for reverberant speech, retaining the covariance matrix same as clean speech.

4 Evaluation

For the evaluation of state splitting approach, it was tested on a speaker-dependent isolated word recognition task. The clean speech HMMs were trained with 2620 words of the same speaker taken from ATR speech database A-Set, and comprised of 41 context-independent phoneme models, each with three emitting states single mixture Gaussian model initially. The single channel speech signal with sampling frequency of 16 kHz 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 filterbanks. The test set consisted of 655 words of the same speaker taken exclusively from the ATR speech database A-set. The word accuracy for clean speech with the clean model was 93.1%, when recognized with Julian3.3p3 Multipath version [4] decoder.

The reverberant speech for test set was prepared by a linear convolution of clean speech and impulse response taken from RWCP Sound Scene Database in Real Environment.

To evaluate state splitting approach, each emitting state of models was split into 20 substates and transition probabilities were updated as described in Section 3. Implicit duration density was used for estimating average state durations. Same impulse response used for simulating reverberant speech was used to compute averaged spectral parameters $\bar{H}_k(t)$. For state sequence, frames coming from other preceding models were not considered, but only intra-model frames were taken into account; and only mean vectors were adapted by this approach.

The experimental results with clean model, CMS [2] and state splitting approach, as listed in Table 1 for different impulse responses, show better performance of state splitting approach. Comparatively lower improvement in case of E1B is due to the fact that it has longer reverberation time, and requires longer frame sequence, including frames from preceding models, to be considered for compensation, whereas only intra-model frames were considered in the experiment.

Table 1: Experimental Results (Word Recognition Rate %)

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

5 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 applying the state splitting approach to explicitly modeled duration density HMM, and estimating frames contributed by preceding models using context-dependent models. The decoding-time implementation of the approach will be also considered.

References

[1] 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.

[2] 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.

[3] C. K. Raut, T. Nishimoto and S. Sagayama, "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. (to appear.)

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