ADAPTING HMMS OF DISTANT-TALKING ASR SYSTEMS USING FEATURE-DOMAIN REVERBERATION MODELS

Armin Sehr, Markus Gardill, and Walter Kellermann

Multimedia Communications and Signal Processing,
University of Erlangen-Nuremberg
Cauerstr. 7, 91058 Erlangen, Germany
Email: {sehr,gardill,wk}@lnt.de

ABSTRACT

To capture the dispersive effect of reverberation by Hidden Markov Model (HMM)-based distant-talking speech recognition systems, adapting the means of the current HMM state based on the means of the preceding states has been suggested in [1]. In this contribution, we propose to incorporate the reverberation models of [2] into the adaptation approach to describe the effect of reverberation with higher accuracy. Connected-digit recognition experiments in three different rooms confirm that the suggested more accurate reverberation representation leads to a significant performance increase in all investigated environments.

1. INTRODUCTION

Robust distant-talking Automatic Speech Recognition (ASR) is desirable for many applications, like seamless human/machine interfaces, speech dialogue systems, and automatic meeting transcription. However, the reverberation caused by multi-path propagation of sound waves from the source to the distant microphone leads to a mismatch between the input utterances and the acoustic model of the recognizer, usually trained on close-talking speech. Therefore, the performance of ASR systems is significantly reduced [3] if no countermeasures are taken.

Reverberant speech can be described by a convolution of clean speech with the Room Impulse Response (RIR) characterizing the acoustic path from the speaker to the microphone. The length of the RIR, typically ranging from 200 ms to 1000 ms, significantly exceeds the length of the analysis window used for feature extraction in ASR systems, typically ranging from 10 ms to 40 ms. Therefore, the time-domain convolution is not transformed into a simple multiplication in the short-time frequency transform (STFT) domain. Instead, reverberation still has a dispersive effect in the STFT domain and also in STFT-based feature domains so that the current feature vector contains a superposition of multiple delayed and attenuated versions of the previous feature vectors. The effect of reverberation on speech feature sequences can be captured approximately by a convolution in the melspectral (melspec) (see Figure 1) domain [4] as given by

$$x_{\mathrm{mel}}(l,k) \approx \sum_{m=0}^{M-1} h_{\mathrm{mel}}(l,m)\, s_{\mathrm{mel}}(l,k-m)\,, \qquad (1)$$

where $x_{\mathrm{mel}}(l,k)$, $h_{\mathrm{mel}}(l,m)$, and $s_{\mathrm{mel}}(l,k-m)$ denote the melspec representations of channel l and frame k for the reverberant speech, the RIR, and the clean speech, respectively.

The dispersion of feature vectors is illustrated for the utterance "four, two, seven" in Figure 2. While the clean sequence a) exhibits a short period of silence before the plosive /t/ in "two" (around frame 52) and a region of low energy for the lower frequencies at the fricative /s/ in "seven" (around frame 78), these low-energy regions are filled with energy from the preceding frames in the reverberant case shown in subfigure b). The melspec convolution according to (1) captures the envelope of the reverberant feature sequence very well as shown in subfigure c). The smearing of the features along the time axis causes the current feature vector to depend strongly on the previous feature vectors. Therefore, the previous feature vectors have to be taken into account for HMM adaptation so that the performance gain of traditional 'intra-frame' model adaptation techniques is limited.

Different approaches have been proposed to obtain acoustic models capturing the dispersive effect. Possibly the most straightforward way is to use reverberant training data to train HMMs. To reduce the effort for data collection, clean training data can be convolved with RIRs to obtain the reverberated data as suggested in [5]. Instead of performing a complete training on reverberated data, the mean vectors of clean HMMs can be adapted to the reverberation conditions of a certain room by taking the means of the preceding states into account [1, 6].

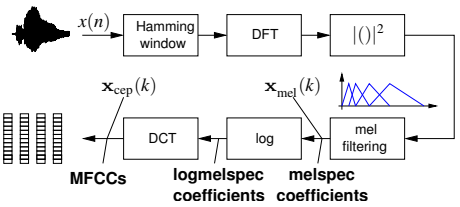

Figure 1: Processing scheme for the calculation of MFCCs.

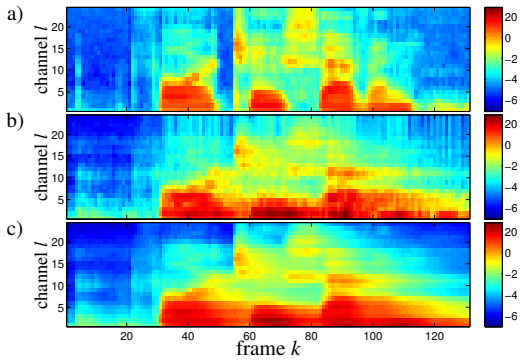

Figure 2: Melspec feature sequences of utterance "four, two, seven" in dB color scale a) clean utterance (close-talking recording), b) reverberant utterance (distant-talking recording), c) approximation of reverberant utterance by (1).

Since the feature-domain RIR representation used in [1] is based on a frequency-independent strictly exponential decay, only the model reverberation time T_{60M} has to be estimated for the adaptation. Therefore, the adaptation can be performed during recognition as described in [1]. However, the relatively simple RIR representation captures the true reverberation characteristics only with relatively low accuracy.

To increase the accuracy of the reverberation capture, using the reverberation models according to [2] as RIR representation for the adaptation approach of [1] is proposed in this contribution. The remainder of the paper is structured as follows: The adaptation algorithm proposed in [1] and the reverberation models of [2] are concisely reviewed in Section 2. The proposed adaptation algorithm is introduced in Section 3 and its performance is compared to reverberant training and to [1] using a connected digit recognition task in Section 4. In Section 5, the paper is summarized and conclusions are drawn.

2. REVIEW OF UNDERLYING ALGORITHMS

2.1 Adaptation Algorithm

The adaptation algorithm according to [1] adjusts the parameters of an HMM λ_S trained on clean speech to obtain an adapted HMM λ_X capturing the characteristics of reverberant feature vector sequences in a certain room. An HMM is defined by the matrix of state transition probabilities, the vector of initial state occupation probabilities, and the output densities for each state [7]. Usually Gaussian mixture densities, completely described by a set of mean vectors, a set of diagonal covariance matrices, and a set of mixture weights, are used to model the output densities in the MFCC domain. Since, according to [8], the adaptation of the covariance matrices has only a minor effect on the recognition performance, only the mean vectors of the HMMs are adapted in [1].

Since the adaptation is performed in the melspec domain, the means of the HMMs are transformed from the MFCC to the melspec domain by a multiplication with an inverse Discrete Cosine Transform (DCT) matrix and an element-wise exponential function. To simplify the adaptation for Gaussian mixture densities, cepstral averages $\bar{\mu}_{\mathbf{S}_{\text{cep}}}(i)$ across the means $\mu_{\mathbf{S}_{\text{cep}}}(i,r)$ of all R mixtures in the MFCC domain are obtained according to

$$\bar{\mu}_{\mathbf{S}_{\text{cep}}}(i) = \sum_{r=1}^{R} w(i,r)\mu_{\mathbf{S}_{\text{cep}}}(i,r)\,, \tag{2}$$

where i and r are the state and mixture indices, respectively, the subscript cep denotes cepstral (MFCC) domain, and $w(i,r)$ is the weight of mixture r in state i. The adapted mean vector $\mu_{\mathbf{X}_{\text{mel}}}(j,r)$ of the static features for state j and mixture r in the melspec domain is obtained by a weighted sum over the melspec representation $\bar{\mu}_{\mathbf{S}_{\text{mel}}}(i)$ of the cepstral averages from (2) for the preceding states i according to

$$\mu_{\mathbf{X}_{\text{mel}}}(j,r) = \alpha(j,j)\,\mu_{\mathbf{S}_{\text{mel}}}(j,r) + \sum_{i=1}^{j-1} \alpha(j,i)\bar{\mu}_{\mathbf{S}_{\text{mel}}}(i)\,, \tag{3}$$

where the subscript mel denotes melspec domain. The adapted mean vector $\mu_{\mathbf{X}_{\text{cep}}}(j,r)$ in the MFCC domain is obtained by applying an element-wise logarithm and a DCT to $\mu_{\mathbf{X}_{\text{mel}}}(j,r)$.

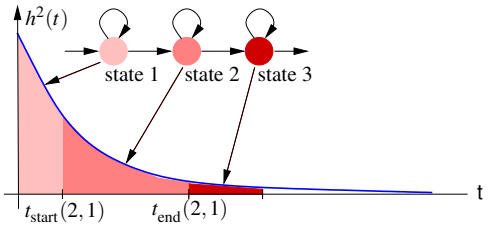

Figure 3: Strictly exponential energy decay used for the calculation of $\alpha(j,i=1)$ according to [1].

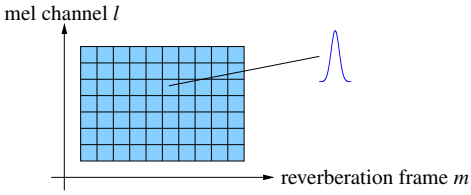

Figure 4: Reverberation model η for observation frame k.

The calculation of the state-level reverberation representation $\alpha(j,i)$, describing the energy dispersion of state i to state j, is based on the assumption of a continuous-time RIR $h(t)$ with strictly exponentially decaying power $h^2(t)$ and unit energy $\int_0^\infty h^2(t)dt = 1$ as depicted in Figure 3. In this case, the melspec RIR representation $h_{\text{mel}}(l,m)$ is independent of the channel l and also exponentially decaying. Therefore, the state-level reverberation representation $\alpha(j,i)$ is also channel-independent and can be calculated by integrating over the squared RIR according to

$$\alpha(j,i) = \int_{t_{\text{start}}(j,i)}^{t_{\text{end}}(j,i)} h^2(t)\,\mathrm{d}t\,. \tag{4}$$

The average start time $t_{\text{start}}(j,i)$ and end time $t_{\text{end}}(j,i)$ of state j for determining the energy dispersion from state i are calculated based on the average duration of state j and its preceding states as illustrated in Figure 3. (See [8] for details on the start and end time calculation as well as on the adaptation procedure for dynamic features.)

2.2 Reverberation Model

A statistical ReVerberation Model (RVM) η is used in [2] for robust distant-talking ASR. This RVM can be considered as a feature-domain representation of all possible RIRs for arbitrary speaker and microphone positions in a certain room. The RVM exhibits a matrix structure where each row corresponds to a certain mel channel and each column to a certain frame as shown in Figure 4. Each matrix element is modeled by a Gaussian Independent Identically Distributed (IID) random process. For simplicity, the elements are assumed to be mutually statistically independent [2]. Thus, the RVM is completely described by its mean matrix $\mu_{\mathbf{H}_{\text{mel}}}$ and its variance matrix $\sigma^2_{\mathbf{H}_{\text{mel}}}$.

3. THE PROPOSED ADAPTATION ALGORITHM

Since the mean adaptation according to [1] is based on a very simple reverberation representation with the model reverberation time T_{60M} as the only parameter, the reverberation representation can be estimated during recognition. However, the strictly exponentially decaying RIR captures the effect of reverberation relatively inaccurately because of two reasons:

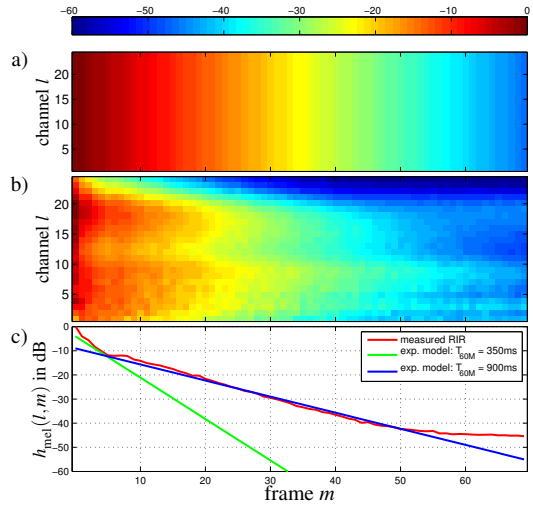

Figure 5: a) Exponential model in the melspec domain (dB color scale) b) melspec representation of measured RIR (dB color scale) c) comparison between the decays of a measured RIR and the exponential model for different model reverberation times T_{60M} for mel channel $l = 18$.

1) The frequency-dependence of the reverberation is not taken into account. Due to room resonances and frequency-dependent absorption coefficients, the acoustic path between speaker and microphone varies strongly with frequency. Therefore, the melspec RIR representation $h_{\text{mel}}(l,m)$ strongly depends on the mel channel l as depicted in Figure 5 b). The exponential model depicted in subfigure a) does not capture this channel dependency.
2) Since real-world RIRs typically exhibit a two-sloped decay with a rapid initial and a slow late decay [9] as depicted in Figure 5 c), the strictly exponential decay can either capture the initial or the late decay with high accuracy. But it is not able to capture the two-sloped behavior.

To tackle these modeling inaccuracies, a combination of the adaptation algorithm [1] and the reverberation model according to [2] is proposed in this contribution. The means $\mu_{H_{\text{mel}}}(l,m)$ of the reverberation model η are used directly as feature domain representation $h_{\text{mel}}(l,m)$ of the RIR. By linear interpolation, a discrete-time version $h_{\text{mel}}(l,t)$ of the melspec RIR representation is obtained and is used for the calculation of the state-level reverberation representation $\alpha(j,i,l)$ according to

$$\alpha(j,i,l) = \int_{t_{\text{start}}(j,i)}^{t_{\text{end}}(j,i)} h_{\text{mel}}(l,t)\,\mathrm{d}t\,. \tag{5}$$

capturing the energy dispersion from state i to state j in mel channel l. Since $h_{\text{mel}}(l,t)$ captures the frequency dependence of the reverberation, the resulting weights $\alpha(j,i,l)$ are channel-dependent.

To adjust the HMM to reverberation, the means in the MFCC domain are transformed to the melspec domain. Based on the state-level reverberation representation $\alpha(j,i,l)$ the mean adaptation is performed according to

$$\mu_{X_{\text{mel}}}(j,r,l) = \alpha(j,j,l)\,\mu_{S_{\text{mel}}}(j,r,l) + \sum_{i=1}^{j-1} \alpha(j,i,l)\bar{\mu}_{S_{\text{mel}}}(i,l) \tag{6}$$

for all mel channels $l = 1 \ldots L$. Finally $\mu_{X_{\text{cep}}}(j,r,c)$ is obtained by transforming $\mu_{X_{\text{mel}}}(j,r,l)$ to the cepstral domain.

	Room A	Room B	Room C
Type	lab	studio	lecture room
T_{60}	300 ms	700 ms	900 ms
d	2.0 m	4.1 m	4.0 m

Table 1: Summary of room characteristics: T_{60} is the reverberation time measured according to [9], d is the distance between speaker and microphone.

The adaptation of the dynamic features is performed according to [1]. Since the adaptation of the Δ-coefficients in [1] is based on the adapted output pdfs for the static features, the adaptation of the Δ-coefficients indirectly benefits from the more accurate reverberation capture of the RVM.

4. EXPERIMENTS

Connected-digit recognition experiments are carried out to compare the performance of the proposed approach to the original adaptation algorithm of [1] and to HMMs trained on reverberant data.

4.1 Experimental Setup

To calculate the MFCC features used for recognition, a DFT length of 512, a window length of 25 ms and a frame shift of 10 ms are used. The 12 MFCC coefficients, including the 0-th coefficient, are augmented by their first derivate calculated according to the HTK [10] defaults. 16-state word-level HMMs with mixtures of three Gaussians serve as clean speech models.

To get the reverberated test data (and the reverberated training data for the training of reverberant HMMs used for comparison), the clean speech TI digits [11] data are convolved with different RIRs measured at different loudspeaker and microphone positions in three rooms with the characteristics given in Table 1. Each test utterance is convolved with an RIR selected randomly from a number of measured RIRs in order to simulate changes of the RIR during the test.

For the adaptation of the HMMs according to the proposed approach, RVMs trained by averaging over RIRs measured at different positions in the target room – different from each other and different form the positions used for the generation of the test data – according to [2] are used to adapt both the static and the dynamic features. Thus, a strict separation of test and training data is maintained.

4.2 Experimental Results

For a deeper understanding of the original algorithm according to [1], a first experiment, investigating the recognition rate as a function of the model reverberation time T_{60M}, is performed in room C. Figure 6 shows the resulting word accuracy. With increasing T_{60M}, the recognition rate increases rapidly until it reaches its maximum at $T_{60M} \approx 325$ ms. Further increasing T_{60M} leads to a slow decrease of the word accuracy. Note that the model reverberation time $T_{60M} \approx 325$ ms achieving the best recognition rate is much lower than the actual reverberation time $T_{60} = 900$ ms of room C measured according to [9]. The difference between T_{60} and T_{60M} arises because the strictly exponential RIR cannot model the two-sloped decay of real-world RIRs with high accuracy. The exponential model can only find a compromise between capturing the rapid initial decay or the slower late decay as illustrated in Figure 5 c).

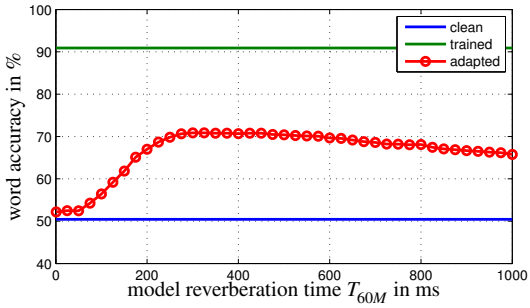

Figure 6: Word accuracy of the original adaptation algorithm according to [1] as a function of the model reverberation time T_{60M} in room C.

For the investigated scenario, a value of $T_{60M} \approx 325\,\text{ms}$, which is fairly close to the rapid initial decay, yields the best recognition results. Since the above example indicates that T_{60M} achieving the best recognition results and T_{60} can be very different, conventional methods for estimating the reverberation time T_{60} cannot be used for estimating T_{60M}. Therefore, an online estimation method for T_{60M} is used in [1]. The advantage of the online method is that it can estimate and track T_{60M} without the need of calibration utterances or RIR measurements in the target room. However, the online estimation also increases the decoding complexity.

In a second experiment, the performance of the adaptation algorithms according to [1] and of the proposed approach according to Section 3 is compared to HMMs trained on clean and matched reverberant data in rooms A, B, and C. As shown in Figure 7, the adaptation according to [1] yields a significant improvement over the clean HMMs in all rooms. A further significant improvement over [1] is achieved in all rooms by the proposed adaptation algorithm. This improvement confirms that using the RVM according to [2], the reverberant feature sequences can be described much more accurately. In room B, the proposed adaptation algorithm is even approaching the accuracy of the HMMs trained on matched reverberant data. Since training on reverberant data allows to adjust all HMM parameters (not just the means) to capture the reverberant feature sequence as closely as possible, the reverberantly trained HMMs can be considered as an upper bound for mean adaptation algorithms.

Furthermore, the RVMs used in the proposed approach can be estimated prior to the recognition either by measuring RIRs in the target environment [2] or by using a few calibration utterances [12]. Since the RVMs can be estimated completely independently of the HMMs and the complexity of the adaptation is very low, the proposed approach is extremely flexible. Moreover, the offline estimation of the RVMs ensures that the low decoding complexity of conventional HMMs is maintained, making the proposed approach very attractive for real-time applications.

5. SUMMARY AND CONCLUSIONS

A combination of the reverberation models according to [2] with the adaptation approach of [1] has been proposed in this contribution. By taking the frequency-dependence and the exact energy decay into account, the reverberation models capture the effect of reverberation more accurately than the strictly exponentially decaying RIR used in [1]. Connected digit recognition experiments confirm that the more accurate reverberation description leads to a significant improvement

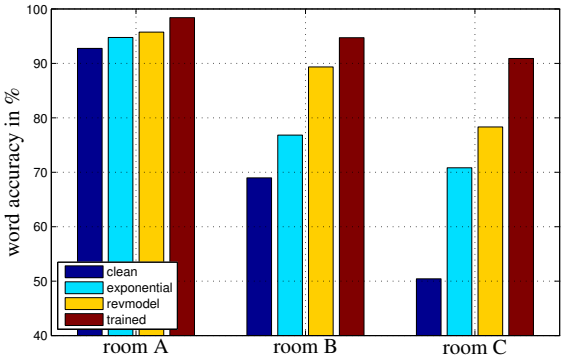

Figure 7: ASR accuracy of clean HMMs, HMMs adapted according to [1] using the model reverberation time T_{60M} achieving the best performance, HMMs adapted according to Section 3, and HMMs trained on matched reverberant data.

in word accuracy in all investigated scenarios. Thus, in some environments, the proposed adaptation concept approaches the performance of HMMs trained on matched reverberant data. Since the effort for estimating the reverberation model and adapting the HMMs is significantly lower than reverberant training, the proposed approach can be used much more flexibly. Future work will include tests with large-vocabulary tasks and a combination with adaptation schemes for additive distortions.

REFERENCES

[1] H.-G. Hirsch and H. Finster, "A new HMM adaptation approach for the case of a hands-free speech input in reverberant rooms," *Proc. INTERSPEECH*, pp. 781–783, September 2006.

[2] A. Sehr, M. Zeller, and W. Kellermann, "Distant-talking continuous speech recognition based on a novel reverberation model in the feature domain," *Proc. INTERSPEECH*, pp. 769–772, 2006.

[3] B. E. D. Kingsbury and N. Morgan, "Recognizing reverberant speech with RASTA-PLP," *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 1259–1262, 1997.

[4] A. Sehr and W. Kellermann, "Towards robust distant-talking automatic speech recognition in reverberant environments," in *Topics in Speech and Audio Processing in Adverse Environments*, E. Hänsler and G. Schmidt, Eds., pp. 679–728. Springer, Berlin, 2008.

[5] D. Giuliani, M. Matassoni, M. Omologo, and P. Svaizer, "Training of HMM with filtered speech material for hands-free recognition," *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1, pp. 449–452, March 1999.

[6] C. K. Raut, T. Nishimoto, and S. Sagayama, "Model adaptation for long convolutional distortion by maximum likelihood based state filtering approach," *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. I–1133 – I–1136, May 2006.

[7] L. R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proc. of IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[8] H.-G. Hirsch and H. Finster, "A new approach for the adaptation of HMMs to reverberation and background noise," *Speech Communication*, vol. 50, no. 3, pp. 244–263, March 2008.

[9] M.R. Schroeder, "New method of measuring reverberation time," *The Journal of the Acoustic Society of America (JASA)*, vol. 37, no. 3, pp. 409–412, 1965.

[10] "HTK webpage," http://htk.eng.cam.ac.uk/.

[11] R. G. Leonard, "A database for speaker-independent digit recognition," *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 42.11.1–42.11.4, 1984.

[12] A. Sehr, J.Y.C. Wen, W. Kellermann, and P.A. Naylor, "A combined approach for estimating a feature-domain reverberation model in non-diffuse environments," *Proc. Int. Workshop on Acoustic Echo and Noise Control (IWAENC)*, 2008.