STRATEGIES FOR MODELING REVERBERANT SPEECH IN THE FEATURE DOMAIN

Armin Sehr and *Walter Kellermann*

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

ABSTRACT

The length of the room impulse response characterizing the acoustic path between speaker and microphone is significantly larger than the length of the analysis window used for feature extraction in automatic speech recognition (ASR) systems. Therefore, reverberation caused by multi-path propagation of sound waves from the speaker to distant-talking microphones has a dispersive effect on speech feature sequences. This dispersive effect causes a mismatch between the input speech and the acoustic models of the recognizer, usually trained on clean speech, and leads to a significant reduction of recognition performance. In this contribution, different strategies for obtaining acoustic models capturing the dispersive effect of reverberation are investigated in terms of modeling accuracy, flexibility with respect to changing reverberation conditions, effort for obtaining the reverberation representation and decoding complexity.

Index Terms— Reverberation, acoustic modeling, distant-talking ASR, robust ASR

1. INTRODUCTION

Robust distant-talking 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-talking 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 [1, 2] 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.

This dispersive effect cannot be captured by traditional 'intra-frame' model adaptation techniques. Instead, information of the previous frames has to be exploited by the acoustic models. Different strategies have been proposed to obtain an acoustic model capable of capturing the dispersive effect. Possibly the most straightforward way is to use reverberant training data to train conventional Hidden Markov Models (HMMs). To reduce the effort for data collection, clean training data can be convolved with RIRs to obtain the reverberant data as suggested in [3, 4]. Instead of performing a complete training on reverberant 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 [5, 6]. These approaches are based on describing the reverberant feature sequence by a convolution of the clean-speech feature sequence with a feature-domain RIR representation in the melspectral (melspec) domain.

The disadvantage of all the aforementioned approaches based on adjusting the parameters of conventional HMMs is the conditional independence assumption underlying the HMMs. This assumption states that the current feature vector only depends on the current state and not on the previous feature vectors. Therefore, conventional HMMs cannot make use of the relationship between neighboring frames caused by reverberation. To overcome this limitation, adaptation of the mean vectors in each frame based on first-order linear prediction has been proposed in [7]. An acoustic model, consisting of a combination of conventional HMMs capturing the clean speech and a reverberation model capturing the effect of reverberation on the feature sequences has been proposed in [8]. A third possibility to account for the inter-frame dependencies due to reverberation is using HMMs with conditional densities, e.g. [9]. However, the authors are not aware of any approach using conditional-density HMMs explicitly for describing reverberant speech.

In this contribution, the aforementioned approaches for acoustic modeling of reverberant feature vector sequences are investigated in terms of modeling accuracy, flexibility with respect to changing reverberation conditions, effort for obtaining the reverberation representation and decoding complexity. The paper is structured as follows: The characteristics of reverberant speech in the feature domain are described in Sec. 2. The different strategies for modeling these characteristics are evaluated and compared in Sec. 3 and conclusions are presented in Sec. 4.

2. CHARACTERIZATION OF REVERBERANT SPEECH

This section describes the characteristics of reverberant speech in the feature domain. In the following, we will consider Mel Frequency Cepstral Coefficients (MFCCs). Since most approaches described in Sec. 3 use melspectral (melspec) or logarithmic melspectral (log-melspec) features to capture the dispersive effect of reverberation, we will also use these features which are intermediate stages in the MFCC computation.

Fig. 1 compares a) the clean and b) the reverberant feature vector sequences corresponding to the utterance "four, two, seven" in the melspec domain. The clean sequence a) exhibits a short period of silence before the plosive /t/ in "two" (around frame 52) and a region

This work was partly supported by the European Commission within the DICIT project under contract number FP6 IST–034624

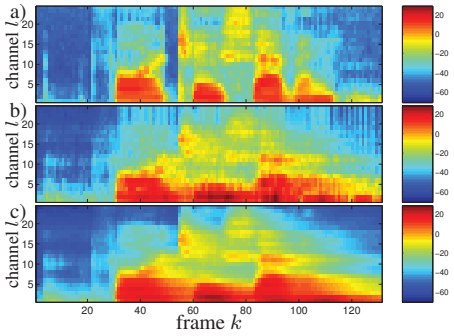

Fig. 1. Melspec feature sequences of the utterance "four, two, seven" in dB color scale a) clean utterance, recorded by a close-talking microphone, b) reverberant utterance, recorded by a microphone four meters away from the speaker, c) approximation of the reverberant utterance according to (2).

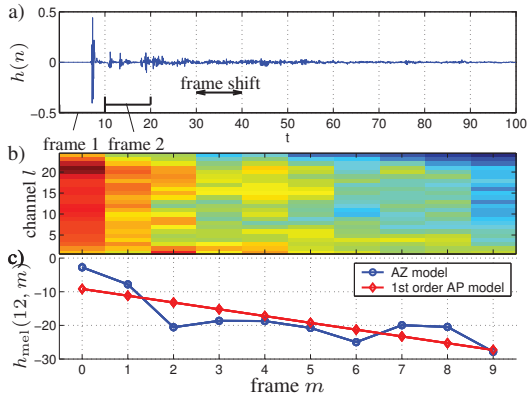

Fig. 2. RIR a) $h(n)$ in the time domain, b) $\mathbf{h}_{\mathrm{mel}}(m)$ in the melspec domain (dB color scale), c) $h_{\mathrm{mel}}(l, m)$ for $l = 12$ in the melspec domain (dB scale)

of low energy for the lower frequencies at the fricative /s/ in "seven" (around frame 78). These are filled with energy from the preceding frames in the reverberant case (subfigure b)). This illustrates that the reverberation has a dispersive effect on the feature sequences: the features are smeared along the time axis so that the current feature vector depends strongly on the previous feature vectors. We will evaluate in Sec. 3 whether the different strategies are able to capture these inter-frame dependencies.

The dispersive effect can be explained by the fact that the typical length of an RIR is much longer than the frame length used for feature extraction as discussed in Sec. 1 and illustrated in Fig. 2 for the initial part of an RIR. Therefore, the time-domain convolution

$$x(n) = h(n) * s(n) \tag{1}$$

of the clean-speech signal $s(n)$ and the RIR $h(n)$ cannot be expressed by a multiplication in the STFT domain but rather by a convolution in each frequency bin. This relationship can be approximated by a convolution in the melspec domain [10]

$$x_{\mathrm{mel}}(l, k) \approx \sum_{m=0}^{M-1} h_{\mathrm{mel}}(l, m)\; s_{\mathrm{mel}}(l, k-m) \tag{2}$$

as illustrated in Fig. 1 c), where l and k are the feature and frame indices, respectively. Due to the inherent approximations of (2), acoustic models based on (2) exhibit slightly more variability and thus

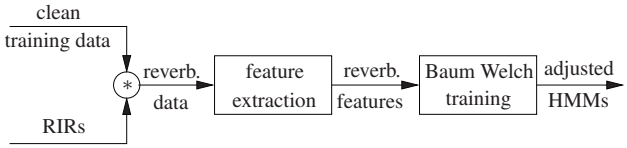

Fig. 3. Block diagram of HMM training with synthetically reverberated training data

slightly less discrimination capability than acoustic models based on the exact convolution (1) [10].

In general, there are two approaches how to capture reverberation in the feature domain. The first approach is based on all-zero (AZ) modeling according to (2), where the reverberant feature vector $x_{\mathrm{mel}}(l, k)$ is given as the weighted sum of the clean-speech feature vectors $s_{\mathrm{mel}}(l, k-m)$ in the melspec domain. Alternatively, all-pole (AP) modeling according to

$$x_{\mathrm{mel}}(l, k) \approx h_{\mathrm{mel}}(l, 0)\, s_{\mathrm{mel}}(l, k) + \sum_{i=1}^{I-1} a_{\mathrm{mel}}(l, i)\; x_{\mathrm{mel}}(l, k-i) \tag{3}$$

can be used where the reverberant part in $x_{\mathrm{mel}}(l, k)$ is given as a weighted sum of the previous reverberant observations $x_{\mathrm{mel}}(l, k-i)$. Since in geometrical acoustics reverberation is interpreted as multiple delayed and attenuated copies of the desired signal [11], AZ modeling captures exactly the physical mechanism of reverberation generation. Therefore, it can describe reverberation extremely accurately. Since the impulse response of an arbitrary minimum-phase filter (evaluation of several RIRs showed that RIR representations in the melspec domain are usually minimum-phase) can be approximated by an AP filter of order I [12], reverberation can in principle also be described with high accuracy by an AP model. However, if only first-order AP models are used as in [6, 7], the feature-domain RIR representation is approximated by a single exponential decay in each mel channel, providing only a basic approximation as depicted in Fig. 2 c).

3. EVALUATION OF MODELING STRATEGIES

In this section, different strategies for obtaining acoustic models capturing the dispersive effect of reverberation on speech feature sequences are compared in terms of modeling accuracy, modeling assumptions, effort for obtaining the reverberation representation and decoding complexity.

3.1. Training conventional HMMs on reverberant data

By training conventional HMMs with data recorded in the target environment, the best acoustic models possible with conventional HMMs are obtained since all parameters of the HMMs are adjusted to the reverberation conditions in the target room. Therefore, reverberant training is usually used as a reference for the recognition rates achievable in reverberant environments. To reduce the enormous effort of collecting training data for each target environment, [3] proposes to convolve clean test data with RIRs measured in the target environment. Thus, slightly less accurate acoustic models are obtained with lower effort [4] compared to recorded training data. This approach is illustrated in Fig. 3. But still measurements of RIRs and a complete model training are required.

Due to the conditional independence assumption, HMMs are not able to model the inter-frame dependencies introduced by reverberation. Therefore, conventional HMMs cannot use the observed previous feature vectors to capture the effect of reverberation even if they are trained on reverberant speech. Instead, a certain reverberantly-trained HMM is only able to describe reverberation by averaging

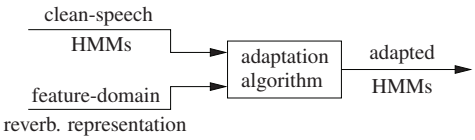

Fig. 4. Block diagram of HMM adaptation

over the reverberation corresponding to the previous feature vectors from all training utterances used for training the respective HMM. By limiting these training utterances to a certain left context, this average will become more specific. Therefore, reverberant training benefits from context-dependent HMMs like cross-word triphones. Whenever reverberant energy from preceding frames not covered by the left context has a major influence on the current feature vector, skewed or even multi-modal densities may result. Therefore, the modeling accuracy can be significantly increased by using several Gaussian mixtures for the HMM output densities. The inability of HMMs to model inter-frame dependencies can be partly remedied by adding dynamic features. However, the dynamic features can only capture dependencies across a few frames while reverberation causes dependencies across a large number of frames.

3.2. Adapting conventional HMMs to reverberation

The effort for obtaining reverberation-robust acoustic models can be reduced by adapting conventional clean-speech HMMs to reverberation. Based on a feature-domain reverberation representation and the melspec convolution (2), the HMM parameters are adjusted as illustrated in Fig. 4. Using (2) and assuming that neighboring feature vectors $\mathbf{s}_{\mathrm{mel}}(l,k)$ are statistically independent, the exact output densities of the adapted HMMs could be obtained by convolving the densities of the clean-speech HMMs. Since the calculation of the resulting densities is mathematically intractable [13], adaptation of only the mean values is proposed in [5] and [6] according to

$$m_{x,\mathrm{mel}}(l,q) = \sum_{p} \gamma_{\mathrm{mel}}(l,p)\; m_{s,\mathrm{mel}}(l,q-p)\;, \qquad (4)$$

where $m_{s,\mathrm{mel}}(l,q)$ and $m_{x,\mathrm{mel}}(l,q)$ are the means of the clean-speech HMM and the adapted HMM for state q and feature l in the melspec domain, respectively, and $\gamma_{\mathrm{mel}}(l,p)$ is a state-level representation of the reverberation. The state-level reverberation representation $\gamma_{\mathrm{mel}}(l,p)$ is obtained in [6] by estimating the reverberation time T_{60} during recognition. Modeling the melspec representation $h_{\mathrm{mel}}(l,m)$ of the RIR as a single exponential decay, $\gamma_{\mathrm{mel}}(l,p)$ is obtained by integrating $h_{\mathrm{mel}}(l,m)$ over the average duration of the preceding states. A Maximum Likelihood (ML) estimation approach based on a few calibration utterances with known transcription is used in [5] to determine $\gamma_{\mathrm{mel}}(l,p)$.

Like the approaches of Sec. 3.1, the adapted HMMs cannot make use of the inter-frame dependencies caused by reverberation. Therefore, the contribution of the preceding frames can only be captured in average over the preceding states of the current HMMs and the left-context HMMs. Since the sequence of the preceding states is not known during adaptation, the adaptation is only based on an average sequence of preceding states. Furthermore, the inherent approximations of the melspec convolution may lead to a slight mismatch in the adaptation. Thus, the reverberation capture is slightly less accurate than with reverberant training (Sec. 3.1), and the recognition rates reported in [5, 6] are slightly lower than for matched reverberant training.

3.3. Frame-wise adaptation of conventional HMMs

One possibility to model inter-frame dependencies is to adapt conventional clean-speech HMMs at each frame (see Fig. 5 b)) using

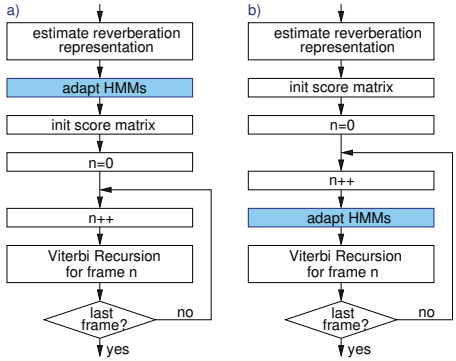

Fig. 5. Flow chart for HMM adaptation a) according to [5, 6] (Sec. 3.2) and b) according to [7] (Sec. 3.3).

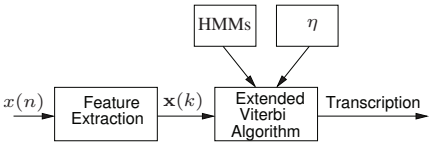

Fig. 6. Structure of a REMOS-based recognizer according to [8].

a first-order AP model as suggested by [7]. Here, the means of all relevant HMM states are adapted according to

$$m_{x,\mathrm{mel}}(l,q) = h_{\mathrm{mel}}(l,0)\; m_{s,\mathrm{mel}}(l,q) + a_{\mathrm{mel}}(l)\; x_{\mathrm{mel}}(l,k-1)\;, \quad (5)$$

where $h_{\mathrm{mel}}(l,0)$ is the melspec RIR representation for feature l and frame 0, $a_{\mathrm{mel}}(l)$ is the AP coefficient, and $x_{\mathrm{mel}}(l,k-1)$ is the observation of the previous frame in the melspec domain. Thus, the information provided by the previous observed feature vector $\mathbf{x}_{\mathrm{mel}}(k-1)$ is utilized and the reverberation is modeled by a strictly exponential decay. Since the exponential decay provides only a basic approximation of the reverberation as illustrated in Fig. 2 c), the recognition rates reported in [7] are slightly lower than that of matched reverberant training according to Sec. 3.1. The parameters $h_{\mathrm{mel}}(l,0)$ and $a_{\mathrm{mel}}(l)$ are readily obtained by ML estimation based on a few calibration utterances. The drawback of the approach is the enormous computational complexity for adapting the HMMs in each frame (see Table 1).

3.4. Combined acoustic model – REMOS

A combined acoustic model consisting of a clean-speech HMM network and a statistical ReVerberation Model (RVM) η as illustrated in Fig. 6 is used in the REMOS (REverberation Modeling for Speech recognition) concept [8] for capturing the inter-frame dependencies due to reverberation. In the melspec domain, the clean-speech HMM output sequence $\mathbf{s}_{\mathrm{mel}}(k)$ and the output sequence $\mathbf{h}_{\mathrm{mel}}(m,k)$ of the reverberation model are combined by the melspec convolution (2) in order to describe the reverberant feature vector sequence $\mathbf{x}_{\mathrm{mel}}(k)$ [8]. The statistical RVM η can be considered as a feature-domain representation of all possible RIRs for arbitrary speaker and microphone positions in the target 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 Fig. 7. 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 [8]. The combined acoustic model used in REMOS can be considered as an AZ model with a stochastic time-variant system function and an independent non-stationary excitation. Thus, the inter-frame dependencies can be modeled very accurately. For recognition, an extended version of the Viterbi algo-

	convolution	reverberation representation	inter-frame dependencies	mean	var.	estimation complexity	decoding complexity	
							qual.	quantitative
3.1	TD	AZ	-	+	+	- -	++	$O(QN(P+L))$
3.2 [5]	MD	AZ	-	+	-	+	++	$O(QN(P+L)$
3.2 [6]	MD	AP, 1st order	-	+	-	++	++	$O(QN(P+L))$
3.3	MD	AP, 1st order	AP, 1st order	+	-	+	-	$O(QN(P+23\cdot L)$
3.4	MD	AZ	AZ	+	+	+	- -	$O(QN(P+(31+4M)L))$

Table 1. Comparison of different modeling strategies. TD: time domain, MD: melspec domain, AZ: all-zero model, AP all-pole model, var.: variance, $O()$: order of, $Q = 176$: number of states in HMM network, $P = 2.56$: average number of predecessor states, $L = 24$: number of features, $M = 50$: number of frames in the reverberation model, $N = 300$: number of frames of the utterance, the given numbers are (typical) values for the digit recognition task used in [8], for large-vocabulary continuous speech recognition, Q and P will be significantly higher.

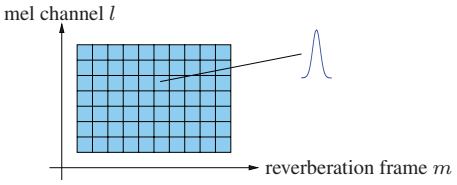

Fig. 7. Reverberation model η for observation frame k.

rithm [8] is employed to find the most likely path through the network of HMMs. The reverberation model η is taken into account by an inner optimization operation determining the most likely contribution of the current HMM state and the reverberation model to the current reverberant observation vector $\mathbf{x}_{\mathrm{mel}}(k)$. Thus, recognition rates significantly higher than those of matched reverberant training according to Sec. 3.1 are achieved in [8]. However, the approach is so far only implemented for melspec features, and the inner optimization significantly increases the decoding complexity.

3.5. Comparison

Table 1 qualitatively compares the most important properties of the above mentioned approaches. All approaches except the reverberant training method (Sec. 3.1) capture the dispersive effect of reverberation by performing a melspec convolution of a clean-speech HMM with a feature-domain reverberation representation. While 3.2 [6] and 3.3 use a first-order AP model for describing the reverberation, the other approaches are based on an AZ model which is closer to the physical reverberation mechanism. The approaches which are solely based on conventional HMMs (3.1 and 3.2) cannot capture the inter-frame dependencies caused by reverberation. 3.3 captures these dependencies by first-order AP models, and 3.4 uses an AZ model. While the approaches 3.2 and 3.3 only adjust the means, the approaches 3.1 and 3.4 also adjust the variances of the output densities. The complexity for estimating the reverberation representation and the decoding complexity are compared in the last two columns of Table 1. Note that the estimation complexity indicates how flexibly the approach can be used in changing acoustic environments.

4. SUMMARY AND CONCLUSIONS

Different approaches for capturing the dispersive effect of reverberation on the speech feature sequences used for ASR have been compared in this contribution in terms of modeling accuracy, effort for the determination of the reverberation representation, and decoding complexity. While reverberant training of HMMs (3.1) can provide very accurate acoustic models, the high effort for obtaining suitable training data for the target environment makes this approach relatively inflexible. Adaptation of conventional HMMs (3.2) requires less effort for the adjustment to reverberation but is also less accurate as only the means of the HMMs are adapted. Furthermore, all approaches based on conventional HMMs (3.1, 3.2) cannot make use of the inter-frame dependencies caused by reverberation. These dependencies can be captured by frame-wise adaptation of HMMs (3.3), or a combined acoustic model consisting of an HMM and a reverberation model (3.4). Since the reverberation representation can be estimated with very little effort for both 3.3 and 3.4, these approaches can be used very flexibly in changing acoustic environments and achieve very promising recognition rates. The combined acoustic model (3.4) even outperforms matched reverberant training. However, both approaches (3.3, 3.4) exhibit an increased decoding complexity. Therefore, it is dependent on the application, which of the evaluated strategies is most suitable.

5. REFERENCES

[1] 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.

[2] R. Petrick, K. Lohde, M. Wolff, and R. Hoffmann, "The harming part of room acoustics in automatic speech recognition," *Proc. INTERSPEECH*, pp. 1094–1097, August 2007.

[3] 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.

[4] V. Stahl, A. Fischer, and R. Bippus, "Acoustic synthesis of training data for speech recognition in living-room environments," *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1, pp. 285–288, May 2001.

[5] 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.

[6] 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.

[7] T. Takiguchi, M. Nishimura, and Y. Ariki, "Acoustic model adaptation using first-order linear prediction for reverberant speech," *IEICE Transactions on Information and Systems*, vol. E89-D, no. 3, pp. 908–914, March 2006.

[8] 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.

[9] J. Ming and F. J. Smith, "Modelling of the interframe dependence in an HMM using conditional Gaussian mixtures," *Computer Speech and Language*, vol. 10, pp. 229–247, 1996.

[10] A. Sehr and W. Kellermann, "New results for feature-domain reverberation modeling," *Proc. Joint Workshop on Hands-Free Speech Communication and Microphone Arrays (HSCMA)*, pp. 168–171, 2008.

[11] H. Kuttruff, *Room Acoustics*, Spon Press, London, UK, 4th edition, 2000.

[12] S. L. Marple, Jr., *Digital spectral analysis: With applications*, Prentice-Hall, Englewood Cliffs, NJ, USA, 1987.

[13] N. C. Beaulieu and F. Rajwani, "Highly accurate simple closed-form approximations to lognormal sum distributions and densities," *IEEE Communication Letters*, vol. 8, no. 12, pp. 709–711, December 2004.