Reverberation Modeling for Robust Speech Recognition

Roland Maas, Armin Sehr, Walter Kellermann
Multimedia Communications and Signal Processing, University of Erlangen-Nuremberg, Erlangen, Germany
{*maas,sehr,wk*}*@LNT.de*

Abstract

The REMOS (REverberation MOdeling for Speech recognition) concept for reverberation-robust distant-talking speech recognition [1] is presented in this paper. REMOS extends a conventional hidden Markov model (HMM) trained on close-talking data with a reverberation model describing the acoustical environment. The combination of both models is performed during recognition to match the reverberant observation. Since varying acoustic conditions only require a reestimation of the reverberation model, REMOS is significantly more flexible than recognition systems trained on reverberant data.

Introduction

When moving from close- to distant-talking automatic speech recognition (ASR), various new problems arise. The ASR system will usually have to deal with background noise and interfering speakers. One of the major challenges in distant-talking ASR scenarios, however, is reverberation, which is captured by the microphones in addition to the desired signal. Reverberation significantly reduces the recognition performance if no countermeasures are taken. This reduction is caused by the dispersive effect of reverberation on the features [2].

There are different techniques to increase the robustness of ASR systems to reverberation. A popular and powerful approach is to train the recognizer's acoustic model on matched reverberant data [3], which has the obvious disadvantage that changing reverberation conditions necessitate a costly retraining. Other more flexible model adaptation techniques have been presented, for example, in [4] and [5]. All those HMM-based methods suffer from the assumption that the current observation vector is conditionally independent of the previous ones, which is clearly violated in the presence of reverberation.

REMOS is a generic framework especially designed for reverberation-robust distant-talking ASR to overcome the conditional independence assumption. The key idea of REMOS is to combine a conventional clean-speech HMM network and a statistical reverberation model (RVM) describing the acoustical environment in the feature domain. During recognition, the most likely contributions of both the HMM and the RVM to the current reverberant observation are determined. The main advantage of REMOS is therefore that changing reverberation conditions do not require an entire retraining of the recognizer. It suffices to reestimate the RVM to adapt REMOS to the new environment.

Effect of Reverberation

Due to its dispersive effect, reverberation strongly influences the time-frequency pattern of speech signals by increasing statistical inter-frame correlation. As can be seen in Fig. 1, the logarithmic melspectral (logmelspec) features of the reverberant utterance "four, two, seven" are smeared along the time-axis compared to the clean version. A given frame is therefore highly depending on its preceding frames, which contradicts the aforementioned HMMs' independence assumption.

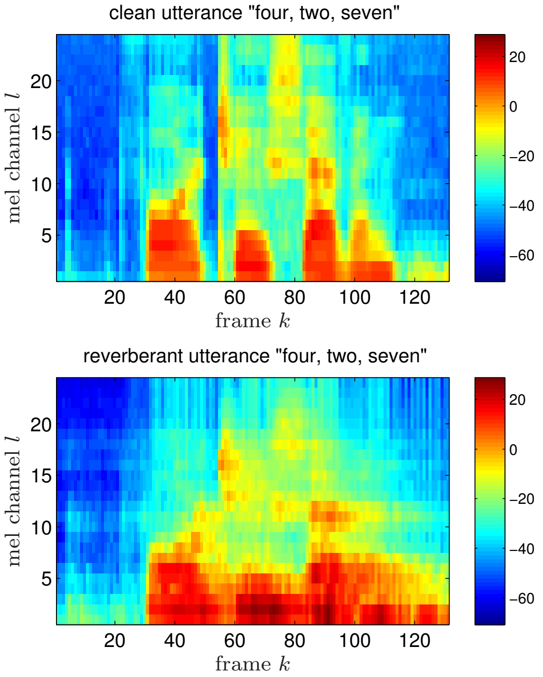

Figure 1: Illustration of the dispersive effect of reverberation [2].

The REMOS Framework

In order to cope with the changed properties of reverberant speech, a generic model describing the effect of reverberation is desired.

We consider a reverberant time-domain signal $x(t)$, which is given by the convolution of the room impulse response (RIR) $h(t)$ with the corresponding clean-speech signal $s(t)$:

$$x(t) = h(t) * s(t).$$

The main idea of the REMOS concept is to describe the corresponding reverberant feature vector sequence $\mathbf{x}(k)$

directly in the logmelspec domain by

$$\exp\big(\mathbf{x}(k)\big) = \exp\big(\mathbf{h}(0,k) + \mathbf{s}(k)\big) + \exp\big(\mathbf{a}(k) + \widehat{\mathbf{x}}_r(k)\big). \tag{1}$$

where

$$\widehat{\mathbf{x}}_r(k) = \log\left(\sum_{m=1}^{M-1} \mu_{\mathbf{h}_{\mathrm{mel}}(m)} \odot \mathbf{s}_{\mathrm{mel}}(k-m)\right) \tag{2}$$

is an approximation of the late reverberant component and the RVM consists of (for details see, e.g., [1])

- M melspec-feature vectors $\mu_{\mathbf{h}_{\mathrm{mel}}(0)}, \dots, \mu_{\mathbf{h}_{\mathrm{mel}}(M-1)}$ being a statistical description of the room impulse response partitioned into M frames,
- $\mathbf{h}(0,k)$ describing the early part of the room impulse response, modeled by a multivariate probability density function $f_{\mathbf{h}(0)}$ in the logmelspec domain,
- and $\mathbf{a}(k)$ capturing the uncertainty of the late reverberation estimation, modeled by a multivariate probability density function $f_{\mathbf{a}}$ in the logmelspec domain.

For recognition, an extended version of the Viterbi algorithm is employed to determine the most likely contributions of the HMM, i.e., $\mathbf{s}(k)$, as well as of the RVM, i.e., $\mathbf{h}(0,k)$ and $\mathbf{a}(k)$. At each step of the extended Viterbi algorithm, the Viterbi score is therefore weighted by the outcome of the following inner optimization problem:

$$\max_{\mathbf{s}(k),\mathbf{h}(0,k),\mathbf{a}(k)} \; f_{\mathbf{s}}\big(\mathbf{s}(k)\big) \cdot f_{\mathbf{h}(0)}\big(\mathbf{h}(0,k)\big) \cdot f_{\mathbf{a}}\big(\mathbf{a}(k)\big) \quad \text{subject to (1)}, \tag{3}$$

where $f_{\mathbf{s}}$ is the output density of the current HMM state and the late reverberation $\widehat{\mathbf{x}}_r(k)$ is calculated by using estimates of $\mathbf{s}_{\mathrm{mel}}(k-m)$, $m = 1, ..., M-1$, cf. (2), known from former Viterbi steps [1].

Experiments and Conclusions

We carried out connected-digit recognition experiments based on the TI-digit corpus to evaluate the performance of REMOS. For a detailed description of the test setup, we refer to [1]. The tests have been performed in three different rooms whose characteristics are summarized in Table 1. Table 2 compares the word accuracies of the REMOS concept to recognizers trained on clean-speech data and on matched reverberant data, respectively. All three system use 24 static logmelspec features with single-Gaussian output densities per HMM state. To obtain benchmark results, we employed a recognizer trained on matched reverberant data with 13 mel-frequency cepstral coefficients (MFCCs), delta coefficients and three Gaussian mixture densities per HMM state.

As can be seen, REMOS clearly outperforms both logmelspec-based recognizers. Although REMOS is based on less powerful features and HMMs with single-Gaussian densities, its performance comes close to that of a matched reverberantly trained state-of-the-art recognizer (3G+MFCC+Δ) in the most reverberant room R3. We recall that REMOS can efficiently be adapted to changing acoustic conditions by simply re-estimating the RVM, whereas a matched trained recognizer would have to be retrained, which is in general computationally very demanding and requires a considerable effort for data collection or generation.

Table 1: Summary of room characteristics: T_{60} is the reverberation time, d is the distance between speaker and microphone, and SRR denotes the Signal-to-Reverberation-Ratio.

Room	Type	T_{60}	d	SRR
R1	lab	300 ms	2.0 m	4 dB
R2	conf. room	780 ms	2.0 m	0.5 dB
R3	lecture room	900 ms	4.0 m	-4 dB

Table 2: Comparison of word accuracies in % for rooms R1 to R3 and different recognizers.

	R1	R2	R3
clean (1G+logmel)	76.3	46.7	32.7
matched rev. trained (1G+logmel)	89.8	81.9	74.5
REMOS (1G+logmel)	90.5	88.9	88.0
matched rev. trained (3G+MFCC+Δ)	98.2	95.0	91.7

We therefore consider REMOS to be a very flexible framework for reverberation-robust speech recognition and see numerous options for further improvements, e.g., by extension to multi-Gaussian densities and state-of-the-art features.

Acknowledgement

The authors would like to thank the Deutsche Forschungsgemeinschaft (DFG) for supporting this work (contract number KE 890/4-1).

References

[1] A. Sehr, R. Maas, and W. Kellermann, “Reverberation Model-Based Decoding in the Logmelspec Domain for Robust Distant-Talking Speech Recognition,” *IEEE Transactions on Audio, Speech, and Language Processing (TASLP)*, vol. 18, no. 7, pp. 1676–1691, 2010.

[2] A. Sehr and W. Kellermann, “On the Statistical Properties of Reverberant Speech Feature Vector Sequences,” in *Proc. Int. Workshop on Acoustic Echo and Noise Control (IWAENC)*, 2010.

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

[4] 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, Mar. 2008.

[5] C. K. Raut, T. Nishimoto, and S. Sagayama, “Model Adaptation by State Splitting of HMM for Long Reverberation,” in *Proc. Interspeech*, 2005, pp. 277–280.