AN UNCERTAINTY DECODING APPROACH TO NOISE- AND REVERBERATION-ROBUST SPEECH RECOGNITION

Roland Maas[1], Akshaya Thippur[2‡], Armin Sehr[3‡], Walter Kellermann[1]

[1]Multimedia Communications and Signal Processing, University of Erlangen-Nuremberg, Erlangen, Germany
{maas,wk}@LNT.de

[2]Computer Vision and Active Perception Lab, School of Computer Science and Communication KTH Royal Institute of Technology, Stockholm, Sweden
akshayats@gmail.com

[3]Beuth University of Applied Sciences Berlin, Department VII, Berlin, Germany
sehr@beuth-hochschule.de

ABSTRACT

The generic REMOS (REverberation MOdeling for robust Speech recognition) concept is extended in this contribution to cope with additional noise components. REMOS originally embeds an explicit reverberation model into a hidden Markov model (HMM) leading to a relaxed conditional independence assumption for the observed feature vectors. During recognition, a nonlinear optimization problem is to be solved in order to adapt the HMMs' output probability density functions to the current reverberation conditions. The extension for additional noise components necessitates a modified numerical solver for the nonlinear optimization problem. We propose an approximation scheme based on continuous piecewise linear regression. Connected-digit recognition experiments demonstrate the potential of REMOS in reverberant and noisy environments. They furthermore reveal that the benefit of an explicit reverberation model, overcoming the conditional independence assumption, increases with increasing signal-to-noise-ratios.

Index Terms— automatic speech recognition, reverberation robustness, noise robustness, uncertainty decoding, piecewise linear regression

1. INTRODUCTION

When moving from close-talking to distant-talking automatic speech recognition (ASR), the ASR system will usually have to deal with additional background noise and reverberation, which significantly reduce the recognition performance if no countermeasures are taken. Such countermeasures are usually distinguished as to whether they act on the speech signals, the speech features, or the ASR models [1].

The REMOS concept [2] directly acts on the ASR models. Originally, REMOS has been designed to individually adapt clean-speech HMMs to a reverberant scenario in each step of the Viterbi decoder. To this end, the dispersive effect of reverberation on the feature vector sequence is explicitly modeled through a mapping function relating the reverberant observation to the underlying clean-speech sequence and a reverberation model. During the Viterbi decoding, the observation likelihood is evaluated by approximating the marginal integral by the maximum integrand resulting in an optimization problem, where the mapping function acts as nonlinear constraint. In [3], an efficient scheme is proposed allowing for a global solution of the constrained optimization problem.

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

‡At the time the work has been conducted, the authors were with the institute of Multimedia Communications and Signal Processing, Erlangen.

In this contribution, the REMOS concept is extended to capture additive noise components by incorporating a noise model into the mapping function. The noise model comprises the mean and variance of the additive distortion in the feature domain. As the extension of the mapping function directly affects the constrained optimization problem, the originally derived solver [3] is to be adapted. We, therefore, introduce a novel continuous piecewise linear regression technique similar to the K-means algorithm in order to approximate the mapping function and, hence, allow for an efficient decoding. The experimental results reveal that the REMOS concept outperforms the constrained maximum likelihood linear regression (CMLLR) technique [4] in the case of static features. Moreover, the benefit of an explicit reverberation model, overcoming the HMMs' conditional independence assumption, increases with increasing signal-to-noise-ratios (SNRs).

The paper is structured as follows: After clarifying notational conventions in Section 2, the REMOS concept is concisely reviewed in Section 3. The proposed extension for additive noise is presented in Section 4 together with the novel regression technique. Connected digit recognition experiments are discussed in Section 5 and Section 6 concludes the paper.

2. NOTATION

Throughout this paper, we stick to the following notational conventions: Feature vectors are D-dimensional and denoted by bold-face letters $\mathbf{v}[n] = (v_1[n], ..., v_D[n])$ with time index $n \in \{1, ..., N\}$. Feature vector sequences are written as $\mathbf{v}[1\!:\!N] = (\mathbf{v}[1], ..., \mathbf{v}[N])$. Every feature vector $\mathbf{v}[n]$ without the explicit subscript "m" is meant to be in the logarithmic melspectral (logmelspec) domain, whereas $\mathbf{v}_{\mathrm{m}}[n]$ denotes the melspectral (melspec) representation of $\mathbf{v}[n]$. The operators "exp" and "log" applied to vectors are meant to be applied component-wise. The operator "$\odot$" denotes the component-wise vector multiplication (Hadamard product). Without distinguishing a random variable from its realization, a probability density function (pdf) over a random variable z is denoted by $p(z)$. For a normally distributed random vector $\mathbf{z}$ with mean $\boldsymbol{\mu}_{\mathbf{z}} = (\mu_{z_1}, ..., \mu_{z_D})$ and diagonal covariance matrix $\mathbf{C}_{\mathbf{z}} = \mathrm{diag}(c_{z_1}, ..., c_{z_D})$, we write $\mathbf{z} \sim \mathcal{N}(\boldsymbol{\mu}_{\mathbf{z}}, \mathbf{C}_{\mathbf{z}})$ or $p(\mathbf{z}) \sim \mathcal{N}(\boldsymbol{\mu}_{\mathbf{z}}, \mathbf{C}_{\mathbf{z}})$.

3. REVIEW OF THE REMOS CONCEPT

As other uncertainty decoding techniques [5], REMOS is based on an observation model relating clean and corrupted feature vectors

while incorporating the environmental uncertainty as random variable. More precisely, the REMOS concept assumes an observed reverberant feature vector $\mathbf{y}_{\mathrm{m}}[n]$, the underlying hidden clean-speech vector $\mathbf{x}_{\mathrm{m}}[n]$ and a statistical reverberation model to be related by a discrete convolution in the melspec domain [2]:

$$\mathbf{y}_{\mathrm{m}}[n] = \mathbf{h}_{\mathrm{m}}[n] \odot \mathbf{x}_{\mathrm{m}}[n] + \mathbf{a}_{\mathrm{m}}[n] \odot \sum_{l=1}^{L} \boldsymbol{\alpha}_{\mathrm{m}}[l] \odot \widehat{\mathbf{x}}_{\mathrm{m}}[n-l],$$

which correspondingly reads in the logmelspec domain:

$$\mathbf{y}[n] = \log(\exp(\mathbf{h}[n] + \mathbf{x}[n]) + \exp(\mathbf{a}[n] + \mathbf{r}[n])) \tag{1}$$

with the late reverberant part

$$\mathbf{r}[n] = \log\left(\sum_{l=1}^{L} \exp(\boldsymbol{\alpha}[l] + \widehat{\mathbf{x}}[n-l])\right) \tag{2}$$

and the previous clean-speech estimates $\widehat{\mathbf{x}}[n-L:n-1]$. The reverberation model is to be estimated once per acoustic environment and consists of a random vector $\mathbf{h}[n] \sim \mathcal{N}(\boldsymbol{\mu}_{\mathbf{h}}, \mathbf{C}_{\mathbf{h}})$ describing the early part of the room impulse response (RIR), a random vector $\mathbf{a}[n] \sim \mathcal{N}(\boldsymbol{\mu}_{\mathbf{a}}, \mathbf{C}_{\mathbf{a}})$ describing the weighting of the late part of the RIR, and the parameters $\boldsymbol{\alpha}[1:L]$ providing a deterministic description of the late part of the RIR.

For the Viterbi decoding, the emission likelihood of the observation $\mathbf{y}[n]$ given the current HMM state $q[n]$ is accordingly evaluated as

$$\begin{aligned} p(\mathbf{y}[n]|q[n]) \quad &= \int\int\int p(\mathbf{y}[n]|\mathbf{x}[n], \mathbf{h}[n], \mathbf{a}[n]) \cdot \\ &\cdot p(\mathbf{x}[n]|q[n]) \cdot p(\mathbf{h}[n]) \cdot p(\mathbf{a}[n])\, d\mathbf{x}[n]\, d\mathbf{h}[n]\, d\mathbf{a}[n], \end{aligned}$$

where the integrals are approximated by the maximum value of the integrand [1]:

$$\max_{\mathbf{x}[n],\mathbf{h}[n],\mathbf{a}[n]} \left\{ p(\mathbf{x}[n]|q[n]) \cdot p(\mathbf{h}[n]) \cdot p(\mathbf{a}[n]) \right\} \text{ s. t.: (1)} \tag{3}$$

and the previous clean-speech estimates $\widehat{\mathbf{x}}[n-L:n-1]$ in (2) are obtained during the optimization process for the L previous frames along the most likely Viterbi path [2].

Assuming HMMs with a single Gaussian output pdf per state and diagonal covariance matrices, i.e.,

$$p(\mathbf{x}[n]|q[n]) \sim \mathcal{N}(\boldsymbol{\mu}_{\mathbf{x}|q}[n], \mathbf{C}_{\mathbf{x}|q}[n]),$$

the optimization problem (3) reduces to the following subproblem for each logmelspec channel $i \in \{1, ..., D\}$:

$$\begin{aligned} \min_{u_i[n], v_i[n]} \quad & \left\{ \frac{(u_i[n] - \mu_{u_i}[n])^2}{2c_{u_i}[n]} + \frac{(v_i[n] - \mu_{v_i}[n])^2}{2c_{v_i}[n]} \right\} \\ \text{s. t.:} \quad & \exp(u_i[n]) + \exp(v_i[n]) = 1, \end{aligned} \tag{4}$$

where we define

$$\begin{aligned} \mathbf{u}[n] &= \mathbf{h}[n] + \mathbf{x}[n] - \mathbf{y}[n], \\ \mathbf{v}[n] &= \mathbf{a}[n] + \mathbf{r}[n] - \mathbf{y}[n], \\ \boldsymbol{\mu}_{\mathbf{u}}[n] &= \boldsymbol{\mu}_{\mathbf{h}} + \boldsymbol{\mu}_{\mathbf{x}|q}[n] - \mathbf{y}[n], \\ \boldsymbol{\mu}_{\mathbf{v}}[n] &= \boldsymbol{\mu}_{\mathbf{a}} + \mathbf{r}[n] - \mathbf{y}[n], \\ \mathbf{C}_{\mathbf{u}}[n] &= \mathbf{C}_{\mathbf{h}} + \mathbf{C}_{\mathbf{x}|q}[n], \\ \mathbf{C}_{\mathbf{v}}[n] &= \mathbf{C}_{\mathbf{a}}. \end{aligned} \tag{5}$$

Once the optimum $\mathbf{u}[n], \mathbf{v}[n]$ are determined, the corresponding optimum $\mathbf{x}[n], \mathbf{h}[n], \mathbf{a}[n]$ can be derived in a straightforward manner [2].

4. EXTENSION FOR ADDITIVE NOISE

This section details the methodology for extending the REMOS concept to additive noise. After incorporating an explicit noise component into the observation model, the modified solution scheme for the extended optimization problem is formulated, necessitating a continuous piecewise planar approximation (CPPA) of the observation model. For the sake of readability, we omit the explicit time dependency n throughout this section.

4.1. Extended observation model

We extend the REMOS observation model (1) by assuming the noise components to be additive in the melspec domain and, hence, log-additive in the logmelspec domain:

$$\mathbf{y} \quad = \quad \log(\exp(\mathbf{h} + \mathbf{x}) + \exp(\mathbf{a} + \mathbf{r}) + \exp(\mathbf{b})), \tag{6}$$

where the additive noise component $\mathbf{b}$ is again modeled as normally distributed random vector $\mathbf{b} \sim \mathcal{N}(\boldsymbol{\mu}_{\mathbf{b}}, \mathbf{C}_{\mathbf{b}})$. Proceeding along the same lines as before, (3) now reads

$$\max_{\mathbf{x},\mathbf{h},\mathbf{a}} \left\{ p(\mathbf{x}|q) \cdot p(\mathbf{h}) \cdot p(\mathbf{a}) \cdot p(\mathbf{b}) \right\} \text{ s. t.: (6)} \tag{7}$$

leading to the following normalized subproblem for each logmelspec channel $i \in \{1, ..., D\}$:

$$\begin{aligned} \min_{u_i, v_i} \quad & \left\{ \frac{(u_i - \mu_{u_i})^2}{2c_{u_i}} + \frac{(v_i - \mu_{v_i})^2}{2c_{v_i}} + \frac{(w_i - \mu_{w_i})^2}{2c_{w_i}} \right\} \\ \text{s. t.:} \quad & \exp(u_i) + \exp(v_i) + \exp(w_i) = 1, \end{aligned} \tag{8}$$

where we introduced analogously to (5)

$$\mathbf{w} = \mathbf{b} - \mathbf{y}, \quad \boldsymbol{\mu}_{\mathbf{w}} = \boldsymbol{\mu}_{\mathbf{b}} - \mathbf{y}, \quad \mathbf{C}_{\mathbf{w}} = \mathbf{C}_{\mathbf{b}}. \tag{9}$$

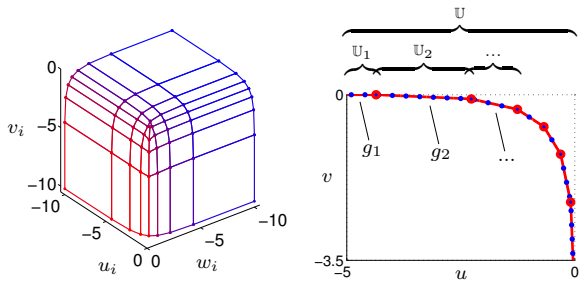
4.2. Solution scheme for the optimization problem

As the Lagrange system of (8) is analytically intractable, an approximative numerical scheme is proposed. Aiming for a global solution, [3] showed the benefit of deriving an approximative solver tailored to the optimization problem (4) compared to a generic numerical solver, such as an inner point optimizer. We therefore construct a novel numerical solver for (8) based on a CPPA of the nonlinear constraint. To this end, we consider the 3D curve in Fig. 1. (a) as a family of 2D curves. This corresponds to considering, e.g., u_i and v_i as variables in (8) while regarding w_i as parameter. Hence, the CPPA of the 3D curve is broken down to a continuous piecewise linear regression problem of a family of 2D curves for various values of w_i, as described in Section 4.3. Once one decided for a certain number of sub-planes, the CPPA is fixed. During recognition, solely the normalization steps (5) and (9) are to be adapted and the optimization problem is to be solved on each sub-plane of the CPPA in order to determine a global solution.

4.3. Continuous piecewise linear regression

In this section, the channel index i is omitted for brevity. We focus on the problem of approximating the exponential curve defined by (8) for a given value of w and therefore rewrite (8) as function of u with parameter w:

$$v = \log(1 - \exp(u) + \exp(w)) =: f(u). \tag{10}$$



(a) (b)

Fig. 1. Constraint equation of the optimization problem (8) in its (a) 3D version and (b) sampled cross-section including its continuous piecewise linear approximation.

Since the approximation will be based on a least-square error (LSE) measure and the integral over $f(u)$ is analytically intractable, we define a set of regression points $u^{(m)} \in \mathbb{R}$, for which $f(u)$ is evaluated: $\mathbb{U} = \left\{u^{(1)}, ..., u^{(M)}\right\}$ with $u^{(1)} < ... < u^{(M)}$. The approximation of (10) by a continuous piecewise linear function consists of two steps:

1. At first, the set of points $\mathbb{U}$ needs to be divided into K subsets $\mathbb{U}_1, ..., \mathbb{U}_K$, where any two subsets $\mathbb{U}_{k-1}, \mathbb{U}_k$ ($2 \leq k \leq K$) corresponding to two adjacent regression line segments overlap in exactly one point as depicted in Fig. 1. (b).
2. Secondly, the parameters α_k, β_k of the regression line segments g_k,

$$g_k(u) = \alpha_k u + \beta_k \text{ for } u \in \mathbb{U}_k,$$

have to be determined on each subset $\mathbb{U}_k$.

While [3] used a brute force approach to determine the subsets $\mathbb{U}_k$, we propose a novel regression technique named *gradient cluster regression* that iteratively fulfills the two above-mentioned steps in a similar way to the K-means algorithm [6]. In analogy to K-means, we start by defining a cost function to be minimized, which we choose as the conventional LSE of the exact function and its approximation subject to a continuity constraint:

$$J := \sum_{m=1}^{M} \sum_{k=1}^{K} r_{mk} \left| f(u^{(m)}) - g_k(u^{(m)}) \right|^2$$

$$\text{with } g_{k-1} = g_k \text{ on } \mathbb{U}_{k-1} \cap \mathbb{U}_k \text{ for } 2 \leq k \leq K.$$

The binary variable r_{mk} indicates which subset $\mathbb{U}_k$ a given point $u^{(m)}$ is assigned to, i.e., which regression line segment g_k it contributes to form:

$$r_{mk} = 1 \iff u^{(m)} \in \mathbb{U}_k,$$
$$r_{mk} = 0 \iff u^{(m)} \notin \mathbb{U}_k.$$

The determination of r_{mk} is denoted as the expectation (E) step and aims at identifying the most significant outliers for each regression line segment in order to reassign them to a "more suitable" segment. This is achieved by "clustering" all points according to their gradient value relative to the regression line gradients:

E: For fixed regression line parameters α_k, β_k, a given point $u^{(m)}$ is assigned to the subset $\mathbb{U}_k$ with the regression line segment g_k having the closest gradient value, i.e.,

$$r_{mk} = \begin{cases} 1 & \text{if } k = \arg\min_j \left| f'(u^{(m)}) - \beta_j \right|, \\ 0 & \text{otherwise.} \end{cases}$$

Room	Type	T_{60}	d	DRR
R1	conf. room	600 ms	2.0 m	+ 0.5 dB
R2	conf. room	700 ms	2.0 m	- 0.5 dB
R3	lecture room	900 ms	4.0 m	- 4 dB

Table 1. Summary of room characteristics: T_{60} is the reverberation time, d is the distance between speaker and microphone, and DRR denotes the direct-to-reverberation ratio.

To ensure the continuity of the overall approximation, the smallest point $u^{(m)}$ of a subset $\mathbb{U}_k$ is also assigned to $\mathbb{U}_{k-1}$, i.e., to the adjacent regression line segment g_{k-1}.

The E step is followed by the update of the parameters of each regression line segment g_k in the maximization (M) step:

M: For fixed assignments r_{mk}, the minimization of J with respect to the regression line parameters results in a constrained optimization problem with analytically solvable Lagrange system.

The E and M steps are alternately repeated until a desired convergence property is fulfilled.

Without giving the detailed proof of convergence, we briefly summarize its main steps: For a given iteration, let $u^{(m)}$ be the intersection point of the line segments g_{k-1} and g_k, i.e., $\{u^{(m)}\} = \mathbb{U}_{k-1} \cap \mathbb{U}_k$, and furthermore

$$\left| f'(u^{(m-1)}) - \beta_k \right| < \left| f'(u^{(m-1)}) - \beta_{k-1} \right|.$$

Consider adding $u^{(m-1)}$ to $\mathbb{U}_k$ and removing $u^{(m)}$ from $\mathbb{U}_{k-1}$ in the E step. The update of the regression line parameters in the subsequent M step then leads to an LSE reduction by δ_{k-1} on the subset $\mathbb{U}_{k-1}$ while increasing the LSE on $\mathbb{U}_k$ by δ_k. Exploiting the strict monotonicity of f and f', it can be shown that $|\delta_k| < |\delta_{k-1}|$.

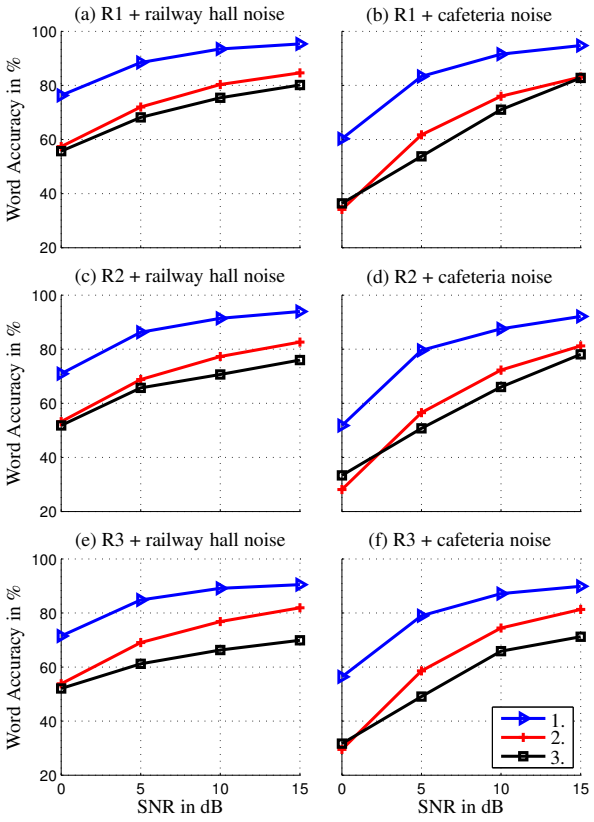
5. EXPERIMENTS

Experiments with the TI digit corpus [7] are carried out to analyze the performance of REMOS with the added noise model. This task is chosen for evaluation since the probability of the current digit can be assumed to be independent of the preceding digits so that the recognition rate is entirely determined by the quality of the acoustic model.

5.1. Experimental setup

The training data comprises solely clean-speech TI digits. The test data are artificially reverberated with measured RIRs and mixed with continuous streams of real noise recordings from the ANITA (Audio eNhancement In Telecom Applications) database [8]. Two different kinds of babble noise conditions are considered: cafeteria and railway hall noise, both scaled relative to the TI digits recordings to achieve SNRs from 0 to 15 dB. The RIRs are 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. While the RIRs for the rooms R1 and R2 are taken from the RWCP sound scene database [9], the RIRs for the most reverberant room R3 are recorded in a lecture hall of the University of Erlangen-Nuremberg [2].

The *REMOS recognizer* is implemented by extending the decoding routines of HTK [10]. REMOS is realized for 24 static logmelspec features with a single Gaussian (1G) output pdf per HMM state.



1.: Full CMLLR (MFCC+Δ+ΔΔ+CMN, 3G) 2.: REMOS, (logmel, 1G) 3.: Static CMLLR (MFCC, 3G)

Fig. 2. Word accuracies for the different test conditions.

For the derivation of the noise model, the first and the last 5 frames of each utterance are assumed to contain only noise and used to estimate the mean $\boldsymbol{\mu}_{\mathbf{b}}$ and variance $\mathbf{C}_{\mathbf{b}}$ in the logmelspec domain. The reverberation model is estimated using RIRs measured at slightly different loudspeaker-microphone positions [2].

To obtain baseline results for reference, we considered two clean-speech recognizers based on Gaussian mixture models with 3 Gaussian components (3G) per HMM state. Both recognizers are adapted to the test data by performing supervised CMLLR [11] based on 100 adaptation utterances. They solely differ in the employed feature set: While the *static CMLLR recognizer* uses 13 static mel frequency cepstral coefficients (MFCCs), the *full CMLLR recognizer* employs 13 MFCCs, 13 delta (Δ) and 13 acceleration ($\Delta\Delta$) coefficients together with cepstral mean normalization (CMN) [12].

5.2. Experimental results

The word accuracies for the three recognizers are depicted in Fig. 2. It can be seen that the proposed extension of REMOS consistently outperforms the static CMLLR recognizer for SNRs greater than $0\,\mathrm{dB}$. In the case of an SNR of $0\,\mathrm{dB}$, the performance of the REMOS and the static CMLLR recognizer differ only slightly. We furthermore note that the highest gains – relative to the static CMLLR recognizer – are achieved in the case of strong reverberation and high SNRs (Fig. 2 (e) and (f)), where the word accuracies of REMOS are closest to ones achieved by the full CMLLR recognizer.

For the interpretation of these results, we need to consider the different structure of the observation models of REMOS and CMLLR: While CMLLR is based on an affine model relating only the current clean and current corrupted feature vector, REMOS explicitly models the dispersive character of reverberation by considering not only the current but also the previous feature vectors (2). Thus, REMOS conceptually extends the conventional HMM structure in order to overcome the conditional independence assumption, that is well-known to be strongly violated in the presence of reverberation [1].

We can therefore conclude from the experimental results that in the case of low SNRs the harming reverberation tail, violating the conditional independence assumption, is considerably masked leading to a lower inter-frame correlation compared to high SNRs. As a consequence, the convolutive reverberation model of REMOS gets all the more beneficial – relative to the affine model of CMLLR – with increasing SNR. The promising results for the considered logmel features and single Gaussian densities indicate that an extension of REMOS to the MFCC domain and Gaussian mixture densities will allow for further improvement in ASR performance. We finally like to point out that the noise and reverberation model of REMOS comprises $L \times D + 6 \times D$ parameters to be estimated, whereas CMLLR necessitates the estimation of $D \times D/3 + D$ adaptation parameters (for MFCC+Δ+$\Delta\Delta$) per regression class [13]. For typical values of $L = 50$, $D = 39$, and, e.g., 32 regression classes [11], REMOS would thus condense the environmental influence into 2,184 parameters at the cost of an increased computational load, while CMLLR would comprise 17,472 parameters requiring a higher amount of adaptation data.

6. RELATION TO PRIOR WORK

As mentioned in the introduction, the concepts for robustifying an ASR system are typically distinguished as to whether they act on the speech signals, the speech features or the ASR models. While a variety of signal enhancement algorithms has been proposed to jointly tackle the problems of additive and long convolutive distortions [14–22], only very few approaches simultaneously address both aspects in the feature or model domain.

To the authors' best knowledge, there are mainly two such approaches operating in the feature domain: While [23] combines multistep linear prediction with particle filters, [24, 25] propose a Kalman filter-based framework incorporating a joint speech, noise, and RIR model.

Turning to the model-based approaches, [26] exploits the energy components of previous frames to adapt the HMM while also stipulating an additive noise term. Based on the parallel model combination procedure, [27] proposes to adapt the HMM parameters using an exponentially decaying RIR model together with a noise spectral estimate. Recently, [28] introduced both an extension to the vector Taylor series and to the CMLLR approach in order to jointly compensate for additive noise and reverberation.

The major difference of the proposed REMOS concept to the existing model-based approaches is two-fold: First of all, REMOS estimates the reverberation tail from the most likely Viterbi path. Secondly, the influence of both the room acoustics and the additive noise sources are modeled by random variables, i.e., from an uncertainty decoding perspective. Hence, REMOS allows for an individual adaptation of the HMMs' output pdfs in each step of the Viterbi decoder while coping at the same time for the uncertainty of the time-varying reverberation and noise components.

7. REFERENCES

[1] T. Yoshioka, A. Sehr, M. Delcroix, K. Kinoshita, R. Maas, T. Nakatani, and W. Kellermann, "Making machines understand us in reverberant rooms: robustness against reverberation for automatic speech recognition," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 114–126, 2012.

[2] 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*, vol. 18, no. 7, pp. 1676–1691, 2010.

[3] R. Maas, A. Sehr, M. Gugat, and W. Kellermann, "A highly efficient optimization scheme for REMOS-based distant-talking speech recognition," in *Proceedings European Signal Processing Conference (EUSIPCO)*, 2010, pp. 1983–1987.

[4] M. J. F. Gales and P. C. Woodland, "Mean and variance adaptation within the MLLR framework," *Computer Speech and Language*, vol. 10, no. 4, pp. 249–264, 1996.

[5] D. Kolossa and R. Haeb-Umbach, *Robust speech recognition of uncertain or missing data*, Springer, 2011.

[6] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*, Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[7] R. Leonard, "A database for speaker-independent digit recognition," in *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 1984, vol. 9, pp. 328–331.

[8] EADS Telecom, *ANITA (Audio eNhancement In Telecom Applications) database*, 2004, http://catalog.elra.info/product_info.php?products_id=12.

[9] S. Nakamura, K. Hiyane, F. Asano, and T. Endo, "Sound scene data collection in real acoustical environments," *Journal of the Acoustical Society of Japan*, vol. 20, no. 3, pp. 225–231, 1999.

[10] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK book*, Cambridge University Engineering Department, 2002.

[11] S. Young, G. Evermann, M. J. F. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK book*, Cambridge University Engineering Department, 2009.

[12] B. S. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," *Journal of the Acoustical Society of America*, vol. 55, no. 6, pp. 1304–1312, 1974.

[13] M. J. F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer Speech and Language*, vol. 12, pp. 75–98, 1997.

[14] J. Gonzalez-Rodriguez, J. L. Sanchez-Bote, and J. Ortega-Garcia, "Speech dereverberation and noise reduction with a combined microphone array approach," in *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2000, vol. 2, pp. II1037–II1040.

[15] K. Kinoshita, M. Delcroix, T. Nakatani, and M. Miyoshi, "Multi-step linear prediction based speech dereverberation in noisy reverberant environment," in *Proceedings Interspeech*, 2007, pp. 854–857.

[16] T. Yoshioka, T. Nakatani, and M. Miyoshi, "Integrated speech enhancement method using noise suppression and dereverberation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 2, pp. 231–246, 2009.

[17] H. W. Löllmann and P. Vary, "A blind speech enhancement algorithm for the suppression of late reverberation and noise," in *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2009, pp. 3989–3992.

[18] R. Talmon, I. Cohen, and S. Gannot, "Multichannel speech enhancement using convolutive transfer function approximation in reverberant environments," in *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2009, pp. 3885–3888.

[19] J. S. Erkelens and R. Heusdens, "Correlation-based and model-based blind single-channel late-reverberation suppression in noisy time-varying acoustical environments," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1746–1765, 2010.

[20] E. A. P. Habets and J. Benesty, "Joint dereverberation and noise reduction using a two-stage beamforming approach," in *Proceedings IEEE Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA)*, 2011, pp. 191–195.

[21] K. Reindl, Y. Zheng, A. Schwarz, S. Meier, R. Maas, A. Sehr, and W. Kellermann, "A stereophonic acoustic signal extraction scheme for noisy and reverberant environments," *Computer Speech and Language*, vol. 27, no. 3, pp. 726–745, 2012.

[22] D. Schmid, S. Malik, and G. Enzner, "A maximum a posteriori approach to multichannel speech dereverberation and denoising," in *Proceedings International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2012, pp. 1–4.

[23] M. Wölfel, "Enhanced speech features by single-channel joint compensation of noise and reverberation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 2, pp. 312–323, 2009.

[24] A. Krüger and R. Haeb-Umbach, "A model-based approach to joint compensation of noise and reverberation for speech recognition," in *Robust speech recognition of uncertain or missing data*, D. Kolossa and R. Haeb-Umbach, Eds., pp. 257–290. Springer, 2011.

[25] V. Leutnant, A. Krüger, and R. Haeb-Umbach, "A statistical observation model for noisy reverberant speech features and its application to robust ASR," in *Proceedings IEEE International Conference on Signal Processing, Communication and Computing (ICSPCC)*, 2012, pp. 142–147.

[26] C. K. Raut, T. Nishimoto, and S. Sagayama, "Maximum likelihood based HMM state filtering approach to model adaptation for long reverberation," in *Proceedings IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2005, pp. 353–356.

[27] 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, 2008.

[28] M. J. F. Gales and Y. Q. Wang, "Model-based approaches to handling additive noise in reverberant environments," in *Proceedings IEEE Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA)*, 2011, pp. 121–126.