Model based feature enhancement for automatic speech recognition in reverberant environments

Alexander Krueger, Reinhold Haeb-Umbach

Department of Communications Engineering, University of Paderborn, Germany
{krueger,haeb}@nt.uni-paderborn.de

Abstract

In this paper we present a new feature space dereverberation technique for automatic speech recognition. We derive an expression for the dependence of the reverberant speech features in the log-mel spectral domain on the non-reverberant speech features and the room impulse response. The obtained observation model is used for a model based speech enhancement based on Kalman filtering. The performance of the proposed enhancement technique is studied on the AURORA5 database. In our currently best configuration, which includes uncertainty decoding, the number of recognition errors is approximately halved compared to the recognition of unprocessed speech.
Index Terms: automatic speech recognition, dereverberation

1. Introduction

Automatic speech recognition is often considered a key technology for human machine communication. In certain applications distant-talking microphones are preferred to close-talking ones because of convenience or safety reasons. However, the increased speaker-microphone distance results in degraded signal quality at the microphones due to the pickup of acoustic environmental noise and reverberation. The latter is a convolutional distortion which is caused by reflections of the speech signals on walls and objects. The source signal is then superposed with its delayed and attenuated versions at the microphone.

Basically, there are two groups of approaches to cope with the convolutional distortion caused by reverberation. On the one hand, there are the signal based methods which try to dereverberate the incoming signal prior to recognition. In this category belong approaches like beamforming, inverse filtering or cepstral mean subtraction. An overview of these approaches can be found in [1]. On the other hand, the model based methods try to adapt the parameters of the recognizer to the effects of reverberation. These approaches comprise the adaptation of HMM means and covariances [2], [3] , state splitting [4] or adaptation of the likelihood evaluation within the Viterbi decoding [5]. All of these methods have shown to improve recognition results in reverberant environments. However, the adaptation of HMM parameters requires a large computational effort.

In this paper we propose to use a feature based method. Instead of trying to dereverberate the time signal itself, we concentrate on the dereverberation of the extracted features used for recognition. The motivation for this is that dereverberation of the features as opposed to the spectrum requires much less information about the room impulse response from the speaker to the microphones. Particularly, our approach only requires an estimate of the reverberation time.

This work is partially supported by the DFG RTG GK-693 of the Paderborn Institute for Scientific Computation (PaSCo).

The organization of the paper is as follows. In section 2 we present the basics of feature based dereverberation, i.e. the observation model, the model of the room impulse response, the posterior estimation method and the uncertainty decoding rule. In section 3 we list the results of recognition experiments which have been carried out on the AURORA5 database. The paper ends with some conclusions.

2. Feature space dereverberation

Commonly, in automatic speech recognition the incoming speech signal is processed by a a front-end to extract features which are successively used for recognition. We will consider the widely used Mel Frequency Cepstral Coefficients (MFCCs) obtained using the ETSI standard front end (SFE) [6]. However, we propose to perform the dereverberation on the log mel spectral coefficients (LMSCs) which are computed at an intermediate stage. The reason is that unlike MFCCs, the LMSCs take values in the same range.

2.1. Observation model

The single channel recorded discrete-time signal $\tilde{y}(l)$, where l denotes the time index, is assumed to result from the convolution of the desired speech source signal $\tilde{x}(l)$ with the room impulse response (RIR) $\tilde{h}(l)$ from the speaker to the microphone:

$$\tilde{y}(l) = \tilde{x}(l) * \tilde{h}(l) = \sum_{p=0}^{L_h} \tilde{h}(p)\tilde{x}(l-p). \tag{1}$$

Additive environmental noise is first ignored for the derivation and the RIR is approximated to be of finite length $L_h + 1$. The source signal can be expressed by its Gabor representation [7]

$$\tilde{x}(l) = \sum_{m} \sum_{k=0}^{N-1} \tilde{X}(m,k)\tilde{w}_s(l-mB)e^{j\frac{2\pi}{N}k(l-mB)} \tag{2}$$

with its short time discrete Fourier transform (STDFT)

$$\tilde{X}(m,k) = \sum_{l} \tilde{x}(l)\tilde{w}_a(l-mB)e^{-j\frac{2\pi}{N}k(l-mB)}. \tag{3}$$

Here, $\tilde{w}_s(l)$ and $\tilde{w}_a(l)$ denote the synthesis and analysis windows of finite length L_w which are biorthogonal to each other [7]. Further, m and k denote the block and frequency index, respectively, B denotes the frame advance and N is the frequency resolution.

The influence of reverberation can be interpreted as the application of a Linear Time Invariant (LTI) system to the source signal. Hence, using (1) and (2), the STDFT of $\tilde{y}(l)$ can be

approximated by the convolution [8]

$$\tilde{Y}(m,k) \approx \sum_{m'=0}^{L_H} \tilde{X}(m-m',k) H_{m',k}. \quad (4)$$

where $L_H = \lfloor (L_h + L_w - 1)/B \rfloor$ and

$$H_{m,k} = e^{j\frac{2\pi}{N}k(L_w-1)} \sum_{p=0}^{2L_w-2} \tilde{h}(mB+p-L_w+1) \times \tilde{w}(L_w-1-p) e^{-j\frac{2\pi}{N}kp} \quad (5)$$

with $\tilde{w}(l)$ defined by $\tilde{w}(l) = \sum_{n=0}^{L_w-1} \tilde{w}_a(n)\tilde{w}_s(n+l)$.

The logarithmic mel-spectrum is obtained by applying a mel filter bank to the power spectrum and computing the logarithm:

$$y_{m,\kappa} = \log \left[\sum_{k=N^{(\mathrm{lo})}(\kappa)}^{N^{(\mathrm{up})}(\kappa)} |\tilde{Y}(m,k)|^2 \Lambda_\kappa(k) \right], \quad (6)$$

where $\Lambda_\kappa(k)$ is the weight function for the κ-th mel band and $N^{(\mathrm{lo})}(\kappa)$ and $N^{(\mathrm{up})}(\kappa)$ are the lower and upper mel band bounds, respectively. Substituting (4) into (6) yields

$$y_{m,\kappa} = \log \Big[\sum_{m'=0}^{L_H} e^{x_{m-m',\kappa} + \overline{h}_{m',\kappa}} + E^{(1)}_{m,\kappa} + E^{(2)}_{m,\kappa} \Big] \quad (7)$$

where

$$\overline{h}_{m,\kappa} = \log \left[\overline{H}^2_{m,\kappa} \right] \quad (8)$$

with

$$\overline{H}^2_{m,\kappa} = \frac{1}{N^{(\mathrm{up})}(\kappa) - N^{(\mathrm{lo})}(\kappa) + 1} \sum_{k=N^{(\mathrm{lo})}(\kappa)}^{N^{(\mathrm{up})}} |H_{m,k}|^2 \quad (9)$$

is an average representation of the RIR in the log mel spectral domain. Further, $x_{m,\kappa}$ is the corresponding log mel spectrum of the clean unreverberated signal. The terms

$$E^{(1)}_{m,\kappa} = \sum_{m'=0}^{L_H} \sum_{k=N_L(\kappa)}^{N_u(\kappa)} |\tilde{X}(m-m',k)|^2 \times \left(|H_{m',k}|^2 - \overline{H}^2_{m',\kappa} \right) \Lambda_\kappa(k) \quad (10)$$

$$E^{(2)}_{m,\kappa} = \sum_{m'=0}^{L_H} \sum_{m''=m'+1}^{L_H} \sum_{k=N_L(\kappa)}^{N_u(\kappa)} 2\Re \Big\{ \tilde{X}(m-m',k) \times \tilde{X}^*(m-m'',k) H_{m',k} H^*_{m'',k} \Big\} \Lambda_\kappa(k) \quad (11)$$

are interpreted as error terms. Here $\Re(\cdot)$ denotes the real part. In doing so, the observation model (7) depends only on the averaged representation $\overline{h}_{m,\kappa}$, i.e. only coarse knowledge about the RIR is required. The two error terms can be combined and approximately modelled as block and mel band dependent zero mean random variables. Defining the reverberant LMSC vector by

$$\mathbf{y}_m = [y_{m,1}, ..., y_{m,K}]^T, \quad (12)$$

K being the number of mel bands, (7) can be written as

$$\mathbf{y}_m = \log \Big[\sum_{m'=0}^{L_H} e^{\mathbf{x}_{m-m'} + \overline{\mathbf{h}}_{m'}} \Big] + \mathbf{v}_m \quad (13)$$

$$= f\left(\mathbf{x}^m_{m-L_H}, \overline{\mathbf{h}}_0^{L_H} \right) + \mathbf{v}_m \quad (14)$$

where $\mathbf{v}_m$ is approximated to be an additive Gaussian noise vector with properly chosen second order characteristics. Here we used the notation

$$\mathbf{x}_m^{m+n} = [\mathbf{x}_m, ..., \mathbf{x}_{m+n}]. \quad (15)$$

to denote a sequence of feature vectors.

2.2. Coarse modelling of the room impulse response

In unknown environments the room impulse response (RIR) from the speaker to the microphone is obviously not available. It is well known that the accurate estimation of the RIR is a complicated task because of its time-variant character. For our purposes it is however sufficient to employ the following coarse model of the RIR

$$\tilde{h}(l) = \sigma_h \cdot \tilde{u}(l) \cdot \tilde{n}(l) e^{-\frac{l}{\tau}}. \quad (16)$$

Here $\tilde{n}(l)$ denotes a realization of a zero mean white Gaussian stochastic process with $\mathrm{E}[\tilde{n}^2(l)] = 1$ and $u(l)$ is given by

$$\tilde{u}(l) = \begin{cases} 1 & : \ 0 \le l \le L_h \\ 0 & : \ \text{else} \end{cases}. \quad (17)$$

Further, σ_h is a scalar which is set to $\sigma_h = \sqrt{\frac{e^{-2/\tau}-1}{e^{-2(L_h+1)/\tau}-1}}$ to normalize the RIR to unit energy:

$$\mathrm{E}\left[\sum_{l=0}^{L_h} \tilde{h}^2(l) \right] = 1. \quad (18)$$

The scalar parameter τ determines the exponential decrease of the envelope and is given by $\tau = T_{60}/(3\log(10) \cdot T_s)$ where T_s denotes the sampling duration and a frequency independent reverberation time T_{60} is assumed [9].

The application of the model requires only the estimation of T_{60} which can be e.g. carried out by a maximum likelihood approach [10]. The normalization (18) necessitates the incoming reverberant signal to be normalized to the same energy as the nonreverberant one.

Assuming the RIR model (16) as a basis it is convenient to assume a normal distribution for its log-mel spectral representation $\overline{h}_{m,\kappa}$. An estimate can be obtained by taking its mean

$$\hat{\overline{h}}_{m,\kappa} = \mathrm{E}[\overline{h}_{m,\kappa}] = \frac{1}{2} \log \left(\frac{\mu^{4^{(\overline{H}^2)}}_{m,\kappa}}{\sigma^{2^{(\overline{H}^2)}}_{m,\kappa} + \mu^{2^{(\overline{H}^2)}}_{m,\kappa}} \right) \quad (19)$$

where $\mu^{(\overline{H}^2)}_{m,\kappa}$ and $\sigma^{2^{(\overline{H}^2)}}_{m,\kappa}$ denote the mean and variance of the log normal distributed random variable $\overline{H}^2_{m,\kappa}$, respectively. Note that $\mu^{(\overline{H}^2)}_{m,\kappa}$ and $\sigma^{2^{(\overline{H}^2)}}_{m,\kappa}$ can be computed using (9) , (5), (16) and (17). The estimates (19) are finally used in (13) to replace the unknown $\overline{\mathbf{h}}_{m'}$.

2.3. Enhancement

We employ a Bayesian framework for the enhancement of reverberant speech features. The a priori probabilistic model, i.e. the model of the dynamics of the clean non-reverberant log-mel spectral speech feature vector $\mathbf{x}_m$ is chosen to be a switching linear dynamical model (SLDM) [11], which has previously been successfully applied for noise robust speech recognition [12]:

$$p(\mathbf{x}_m | \mathbf{x}_{m-1}, \gamma_m = i) = \mathcal{N}(\mathbf{x}_m; \mathbf{A}_i \mathbf{x}_{m-1} + \mathbf{b}_i, \mathbf{C}_i) \quad (20)$$

$$p(\mathbf{x}_1 | \gamma_m = i) = \mathcal{N}(\mathbf{x}_1; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i). \quad (21)$$

Here, $\mathcal{N}(\cdot;\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i)$ denotes a Gaussian density with mean vector $\boldsymbol{\mu}_i$ and covariance matrix $\boldsymbol{\Sigma}_i$. The distribution (20) corresponds to a probabilistic auto regressive (AR) model of first order where the matrices $\mathbf{A}_i$ and offset vectors $\mathbf{b}_i$ describe the transition between successive feature vectors. Further, γ_m is the hidden state variable which indicates the valid state at time index m. The model parameters $\mathbf{A}_i, \mathbf{b}_i, \mathbf{C}_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i,\ i = 1,\ldots,M$, as well as the transition probabilities $a_{ij} = P(\gamma_{m+1} = j|\gamma_m = i)$ can be learnt from training data applying the well known EM algorithm [11]. Note that by assuming $\mathbf{A}_i = \mathbf{0}$ and $a_{ij} = 1/M^2$ a Gaussian mixture model (GMM) is obtained to model the clean speech distribution. However, we propose to employ a SLDM to take into account the high correlation between adjacent feature vectors.

We propose here to estimate the clean nonreverberant LMSC trajectory by approximation of the MMSE estimate $\mathbf{x}_{m|m} = \mathrm{E}[\mathbf{x}_m|\mathbf{y}_1^m]$. With the model conditioned expectations of the clean nonreverberant LMSCs and their error covariance matrices defined by

$$\mathbf{x}_{m|m}^{(i)} = \mathrm{E}[\mathbf{x}_m|\mathbf{y}_1^m, \gamma_m = i] \tag{22}$$

$$\boldsymbol{\Sigma}_{m|m}^{(i)} = \mathrm{E}\left[\left(\mathbf{x}_m - \hat{\mathbf{x}}_{m|m}^{(i)}\right)\left(\mathbf{x}_m - \hat{\mathbf{x}}_{m|m}^{(i)}\right)^T|\mathbf{y}_1^m, \gamma_m = i\right], \tag{23}$$

the estimates $\hat{\mathbf{x}}_{m|m}^{(i)}$ and $\hat{\boldsymbol{\Sigma}}_{m|m}^{(i)}$ can be iteratively computed by a set of Kalman filters and then combined to the final estimates

$$\hat{\mathbf{x}}_{m|m} = \sum_{i=1}^{M} \hat{\mathbf{x}}_{m|m}^{(i)} P(\gamma_m = i|\mathbf{y}_1^m) \tag{24}$$

$$\hat{\boldsymbol{\Sigma}}_{m|m} = \sum_{i=1}^{M} P(\gamma_m = i|\mathbf{y}_1^m) \times\left[\hat{\boldsymbol{\Sigma}}_{m|m}^{(i)} + \left(\hat{\mathbf{x}}_{m|m} - \hat{\mathbf{x}}_{m|m}^{(i)}\right)\left(\hat{\mathbf{x}}_{m|m} - \hat{\mathbf{x}}_{m|m}^{(i)}\right)^T\right]. \tag{25}$$

Because the observation model (13) is nonlinear we propose to use iterative Extended Kalman Filters using 3 iterations per state update. The Kalman Filters compute the predicted observation

$$\hat{\mathbf{y}}_{m|m-1}^{(i)} = f\left(\hat{\mathbf{x}}_{m|m-1}^{(i)}, \hat{\mathbf{x}}_{m-1|m-1}, ..., \hat{\mathbf{x}}_{m-L_H|m-L_H}, \hat{\overleftarrow{\mathbf{h}}}_0^{L_H}\right) \tag{26}$$

with $f(\cdot)$ defined in (13). As the number of possible model histories, i.e. sequences $\{\gamma_1, ..., \gamma_m\}$, increases exponentially with m, for the computation of (22) and (23) approximations have to be used to keep the approach practical, e.g. the Generalized Pseudo Bayesian approach of first (GPB1) or second order (GPB2) or the Interacting Multiple Model (IMM) approach [13].

2.4. Uncertainty Decoding

The advantage of the proposed feature enhancement approach is that with the error covariance matrix $\hat{\boldsymbol{\Sigma}}_{m|m}$ a measure of uncertainty is computed in parallel to the estimation of the clean features. The utilization of this information for improving the recognition performance by changing the decoding rule is known as uncertainty decoding (UD) [14].

Typical recognizers employ Hidden Markov Models (HMMs) to model the production of the nonreverberant clean feature vectors. As features we propose here to use MFCCs $\mathbf{x}_m^{(c)} = \left(x_{m,1}^{(c)}, ..., x_{m,K_c}^{(c)}\right)^T$, K_c denoting the number of cepstral components, for recognition to benefit from a reduced dimension and nearly decorrelated feature vector components. The enhanced MFCCs $\hat{\mathbf{x}}_{\mathbf{m}}^{(\mathbf{c})}$ as well as their error variances $\boldsymbol{\sigma}_m^{2^{(c)}} = \left(\sigma_{m,1}^{2^{(c)}}, ..., \sigma_{m,K_c}^{2^{(c)}}\right)^T$ are computed from the enhanced LMSCs by

$$\hat{\mathbf{x}}_m^{(c)} = \mathbf{M}_{\mathrm{DCT}} \cdot \hat{\mathbf{x}}_{m|m} \tag{27}$$

$$\boldsymbol{\sigma}_m^{2^{(c)}} = \mathrm{diag}\left(\mathbf{M}_{\mathrm{DCT}} \cdot \hat{\boldsymbol{\Sigma}}_{m|m} \cdot \mathbf{M}_{\mathrm{DCT}}^T\right), \tag{28}$$

where $\mathbf{M}_{\mathrm{DCT}}$ is the Discrete Cosine Transform (DCT) matrix and $\mathrm{diag}(\cdot)$ denotes the operation of extracting the diagonal of a matrix. The observation probability for the recognizer which is dependent on the HMM state variable q_m is modelled as a GMM comprising J components

$$p\left(\mathbf{x}_m^{(c)}|q_m\right) = \sum_{j=1}^{J} c_{q_m,j} \cdot \mathcal{N}\left(\mathbf{x}_m^{(c)}; \boldsymbol{\mu}_{q_m,j}^{(\mathrm{GMM})}, \boldsymbol{\sigma}_{q_m,j}^{2^{(\mathrm{GMM})}}\right). \tag{29}$$

where $\boldsymbol{\mu}_{q,j}^{(\mathrm{GMM})}$ and $\boldsymbol{\sigma}_{q,j}^{2^{(\mathrm{GMM})}}$ are the mean and variance vectors of the GMM corresponding to the j-th mixture component and state q. The variance information obtained by (28) can be employed to compute the density of the reverberant MFCCs $\mathbf{y}_m^{(c)}$ by

$$p\left(\mathbf{y}_m^{(c)}|q_m\right) \propto \sum_{j=1}^{J} c'_{q_m,j}\left(\boldsymbol{\mu}_m^{(\mathrm{eq})}; \boldsymbol{\mu}_{q_m,j}^{(\mathrm{GMM})}, \boldsymbol{\sigma}_{q_m,j}^{2^{(\mathrm{GMM})}} + \boldsymbol{\sigma}_m^{2^{(\mathrm{eq})}}\right). \tag{30}$$

where the i-th component of the equivalent means $\boldsymbol{\mu}_m^{(\mathrm{eq})}$ and variances $\boldsymbol{\sigma}_m^{2^{(\mathrm{eq})}}$ and the modified mixture weights $c'_{q,j}$ are given by the following equations [14]

$$\sigma_{m,i}^{2^{(\mathrm{eq})}} = \left[\left(\sigma_{m,i}^{2^{(c)}}\right)^{-1} - \left(\sigma_i^{2^{(c,x)}}\right)^{-1}\right]^{-1} \tag{31}$$

$$\mu_{m,i}^{(\mathrm{eq})} = \sigma_{m,i}^{2^{(\mathrm{eq})}}\left(\frac{\hat{x}_{m,i}^{(c)}}{\sigma_{m,i}^2} - \frac{\mu_{x,i}^{(c)}}{\sigma_i^{2^{(c,x)}}}\right) \tag{32}$$

$$c'_{q,j} = c_{q,j}\frac{\mathcal{N}\left(0; \hat{x}_{m,i}^{(c)}, \sigma_{m,i}^{2^{(c)}}\right)}{\mathcal{N}\left(0; \mu_i^{(c,x)}, \sigma_i^{2^{(c,x)}}\right)\mathcal{N}\left(0; \mu_{m,i}^{(\mathrm{eq})}, \sigma_{m,i}^{2^{(\mathrm{eq})}}\right)}. \tag{33}$$

Here, $\mu_i^{(cx)}$ and $\sigma_i^{2^{(c,x)}}$ denote the mean and variance of the i-th cepstral component of a Gaussian prior $p(\mathbf{x}_m^{(c)})$ obtained from the whole training data. It has been observed that the a posteriori variances $\sigma_{m,i}^2$ can become too large as consequence of poor estimation. Its value is therefore limited to $0.5\sigma_{x,i}^{2^{(c)}}$. This UD approach was also applied for the MFCC Δ- and $\Delta\Delta$-components whose estimates and error variances were computed from (27) and (28).

3. Simulation results

The performance of the proposed feature enhancement algorithm was evaluated on the noisefree reverberant utterances of the AURORA 5 connected digits recognition task [15].

3.1. Model training

For the EM training of the SLDMs we used 8623 clean unreverberant utterances. The training procedure is well described in [11]. The acoustic models of the recognizer consisted of speaker independent word based HMMs with 16 states per word and 4 Gaussian mixture components per state. Simple left-to-right models without skips over states were used. The HTK software was employed for the training of the HMM parameters and Viterbi decoding for the recognition.

3.2. Baseline results

Apart from clean non-reverberant test utterances, the AURORA5 database contains reverberant test utterances originating from the convolution of clean test utterances with randomly generated RIRs for two conditions, office room and living room. The reverberation time T_{60} is varied in the range from 0.3s to 0.4s for the office room and in the range from 0.4s to 0.5s for the living room. For each condition there were 8700 test utterances. The baseline recognition results using either the ETSI SFE or advanced front end (AFE) are depicted in table 1. It can be clearly seen that the recognition performance severely degrades when the reverberation increases.

Table 1: Baseline word accuracies [%].

Front end	Non-reverberant	Office	Living room
SFE	99.33	92.64	82.02
AFE	99.36	92.12	81.14

3.3. Results with proposed algorithm

The feature enhancement, i.e. dereverberation, was performed in the log mel spectral domain as described in section 2.3. We employed the SFE for the extraction of log mel spectral feature vectors. To compute the log-mel spectral representation of the RIR (8) we used a fixed randomly generated impulse response according to (16) with a reverberation time of 0.35s for the office and 0.45s for the living room condition. The enhanced log mel spectral feature vetors were converted to MFCCs ($K_c = 13$), which were successively used for recognition together with their Δ- and $\Delta\Delta$-components. We evaluated the recognition performance depending on the number M of SLDM models, the chosen inference algorithm (GPB1/2 or IMM) and the ASR decoding algorithm (with or without uncertainty decoding). The results for office and living room are given in tables 2 and 3, respectively.

Table 2: Word accuracies [%] for office room (T_{60}=0.3s-0.4s).

Algorithm	$M = 2$	$M = 4$	$M = 8$	$M = 16$
GPB1	93.36	93.94	94.08	93.97
GPB1 + UD	94.99	95.50	95.46	95.02
IMM	94.17	94.63	94.85	94.56
IMM + UD	95.01	95.18	95.08	94.49
GPB2	94.28	94.84	95.19	-
GPB2 + UD	94.99	95.72	95.72	-

First of all, it can be observed that the recognition performance is improved with the proposed approach. As expected, the IMM approach gives better results for all considered numbers of models compared to GPB1, especially for the living room condition. The GPB2 approach delivers even slightly better results for all considered cases, however at the cost of significantly increased computational effort. It can also be seen that increasing the number of models only moderately improves recognition results. By the application of the UD the recognition results can be further improved For the best configuration the number of recognition errors is approximately reduced by a factor of two.

Table 3: Word accuracies [%] for living room (T_{60}=0.4s-0.5s).

Algorithm	$M = 2$	$M = 4$	$M = 8$	$M = 16$
GPB1	84.63	83.14	82.67	83.79
GPB1 + UD	89.57	88.61	88.09	88.98
IMM	87.03	87.53	88.43	88.63
IMM + UD	90.57	90.04	90.36	90.07
GPB2	87.70	88.10	88.99	-
GPB2 + UD	91.00	90.67	91.11	-

4. Conclusions

In this paper we have presented a new observation model for reverberant feature vectors in the log mel spectral domain and applied it to feature enhancement. Simulation results showed that the proposed approach delivers significantly improved recognition results compared to recognition of unprocessed reverberant features for utterances of connected digits emitted in reverberant environments.

5. References

[1] E. Habets, "Single- and multi-microphone speech dereverberation using spectral enhancement", *Ph.D. diss., TU Eindhoven*, 2007.

[2] 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.

[3] M. Delcroix and T. Nakatani and S. Watanabe, "Static and dynamic variance compensation for recognition of reverberant speech with dereverberation preprocessing", *IEEE Trans. Audio Speech Lang. Process.*, vol. 17, no. 2, pp. 324-334, Feb. 2009.

[4] C. K. Raut and T. Nishimoto and S. Sagayama, "Model adaptation by state splitting of hmm for long reverberation", *Proc. Interspeech*, Sep. 2005.

[5] A. Sehr and W. Kellermann, "A new concept for feature-domain dereverberation for robust distant-talking asr", *Proc. IEEE ICASSP*, vol. 4, pp. IV-369-IV-372, Apr. 2007.

[6] "ETSI standard document, 2000." *Speech Processing, Transmission and Quality aspects (STQ); Distributed speech recognition; Front-end feature extraction algorithm; Compression algorithms.* ETSI ES 201 108 v1.1.2 (2000-04).

[7] S. Qian and D. Chen, "Discrete gabor transform", *IEEE Trans. Signal Process.*, vol. 41, no. 7, pp. 2429-2438, Jul. 1993.

[8] Y. Avargel and I. Cohen, "System identification in the short-time fourier transform domain with crossband filtering", *IEEE Trans. on ASLP*, vol. 15, no. 4, pp. 1305-1319, 2007.

[9] H. Kuttruff, "Room acoustics", *Spon Press, London, UK*, 4th edition, 2000.

[10] R. Ratnam and D. L. Jones and B. C. Wheeler and W. D. O'Brien and C. R. Lansing and A. S. Feng, "Blind estimation of reverberation time", *J Acoust Soc Am*, vol. 114, no. 5, Nov. 2003.

[11] K. P. Murphy, "Switching kalman filters", *U.C. Berkeley, Tech. Rep.*, 1998.

[12] J. Droppo and A. Acero, "Noise robust speech recognition with a switching linear dynamic model", *Proc. IEEE ICASSP*, vol. 1, pp. 953-6, May 2004.

[13] Y. Bar-Shalom and X. R. Li and T. Kirubarajan, "Estimation with applications to tracking and navigation: theory, algorithms, and software", *Wiley, New York*, 2001.

[14] V. Ion and R. Haeb-Umbach, "A novel uncertainty decoding rule with applications to transmission error robust speech recognition", *IEEE Trans. ASLP*, vol. 16, no. 5, pp.1047-1060, 2008.

[15] H. G. Hirsch, "Aurora-5 experimental framework for the performance evaluation of speech recognition in case of a hands-free speech input in noisy environments", *Niederrhein University of Applied Sciences, Tech. Rep.*, 2007.