A MULTICHANNEL FEATURE COMPENSATION APPROACH FOR ROBUST ASR IN NOISY AND REVERBERANT ENVIRONMENTS

Ramón Fernandez Astudillo[1], *Sebastian Braun*[2], *Emanuël A. P. Habets*[2]

[1] Spoken Language Systems Laboratory, INESC-ID-Lisboa, Lisboa, Portugal
[2]International Audio Laboratories Erlangen†, Am Wolfsmantel 33, 91058 Erlangen, Germany

ABSTRACT

In this paper we propose a multichannel feature compensation approach for automatic speech recognition in reverberant and noisy environments. The proposed technique propagates the posterior of the clean signal estimated by a multichannel Wiener filter in short-time Fourier transform (STFT) domain into Mel-frequency cepstrum coefficients (MFCC) domain. The multichannel Wiener filter reduces both reverberation and additive noise. Furthermore, we approximate the propagation of the prior distributions of speech and interference through the inverse STFT and the STFT with different time-frequency resolutions. This allows us to derive a multichannel minimum mean square error MFCC estimator with an STFT resolution that is different from the resolution in the speech enhancement stage. The proposed approach is able to outperform a multichannel short-time spectral amplitude estimation approach on both the clean training and multi-condition training ASR tasks of the REVERB challenge.

Index Terms— Multichannel, dereverberation, automatic speech recognition, MMSE-STSA, MMSE-MFCC, observation uncertainty

1. INTRODUCTION

The application of automatic speech recognition (ASR) systems to challenging tasks has increased considerably in the last few years. One such task is the recognition of so-called distant speech in reverberant and noisy environments, for example required for hands-free home automation services. Unlike close-talking scenarios, where static adaptation using in-domain data already mitigates robustness problems, such a task can greatly benefit from multiple microphones. This turns ASR in such environments into a multi-disciplinary task, where the expertise of the speech enhancement (SE) and ASR fields have to be joined.

In this paper, we propose a robust ASR system for the REVERB challenge task which attains a tight integration of the SE and ASR stages by propagating the uncertainty associated to a minimum mean square error (MMSE) estimate. At the SE stage, the posterior distribution associated with the signal obtained using the multichannel MMSE (M-MMSE) estimator proposed in [1] is calculated. A multichannel MMSE short-time spectral amplitude (M-MMSE-SA) estimator is also used for an improved estimation of the parameters. The obtained posterior is then propagated through intermediate inverse short-time Fourier transform (ISTFT) and short-time Fourier transform (STFT) transformation to compute the posterior distribution in the domain of the ASR stage. Finally, this posterior is propagated into Mel-frequency cepstral coefficients (MFCC) domain to attain an M-MMSE estimate of the MFCC.

The proposed approach to integrate the SE and ASR stages improves upon previous approaches that make use of MMSE-MFCC estimators [2–4] and other similar integration approaches such as the one proposed in [5] in two ways. Firstly, it propagates the posterior associated with an M-MMSE estimate rather than a single-channel MMSE estimate. Secondly, it makes use of an approximation that allows the use of different time-frequency resolutions of the STFT in the SE and ASR stages.

The resulting multichannel MMSE Mel-frequency cepstral coefficients (M-MMSE-MFCC) estimator is able to exploit expertise in STFT domain SE while allowing both feature and model compensation in MFCC domain. Changes to the ASR architecture include only propagation through the feature extraction and model based enhancement, which results in a small increase in computational complexity. Results on the REVERB task show that by tightly integrating the SE and ASR systems, it is possible to improve upon a baseline ASR system that uses only the M-MMSE-SA estimator based on the system proposed in [1].

2. JOINT REVERBERATION AND NOISE REDUCTION BY INFORMED SPATIAL FILTERING

In this section, a method is presented to estimate the clean speech signal that is distorted by reverberation and additive

*This work was supported by the Portuguese Foundation for Science and Technology through grant number SFRH/BPD/68428/2010 and project PEst-OE/EEI/LA0021/2013 and the EU Project DIRHA, FP7-ICT-2011-7-288121.

†A joint institution of the University Erlangen-Nuremberg and Fraunhofer IIS, Germany.

noise. The method is derived using a parametric sound field model presented in [6]. In contrast to the method presented in [7], it does not require an estimate of the reverberation time.

2.1. Acoustic Signal Model

We consider an array of M microphones capturing the sound field pressure. In the STFT domain, the microphone signals are written into vectors of length M so that

$$\mathbf{y}(k,n) = [Y_1(k,n), \ldots, Y_M(k,n)]^T, \tag{1}$$

where k denotes the frequency inde, n the time frame index and $Y_m(k,n)$ is the observed signal at the m-th microphone. We employ a simple signal model with a single desired sound source in a reverberant room and additive stationary noise such as microphone self noise and ambient noise. The reverberant sound is modeled by a single plane-wave and a homogeneous and isotropic diffuse sound field. The microphone signals can be described by

$$\mathbf{y}(k,n) = \mathbf{d}(k,n)S(k,n) + \mathbf{r}(k,n) + \mathbf{v}(k,n), \tag{2}$$

where $S(k,n)$ denotes the direct sound of the desired signal as received by a reference microphone, $\mathbf{d}(k,n)$ is the complex-valued relative propagation vector of the direct sound from the reference microphone to all microphones. The signal vector $\mathbf{r}(k,n)$ is the diffuse sound and $\mathbf{v}(k,n)$ is the additive stationary noise.

We assume all components to be mutually uncorrelated such that the power spectral density (PSD) matrix of the microphone signals can be expressed as

$$\begin{aligned}\mathbf{\Phi}_{\mathbf{y}}(k,n) = {} & \phi_S(k,n)\,\mathbf{d}(k,n)\mathbf{d}^H(k,n) \\ & + \phi_R(k,n)\,\mathbf{\Gamma}_{\text{diff}}(k) + \mathbf{\Phi}_{\mathbf{v}}(k,n),\end{aligned} \tag{3}$$

where $\phi_S(k,n)$ is the PSD of the desired speech signal at the reference microphone, $\phi_R(k,n)$ is the PSD of the diffuse sound, $\mathbf{\Gamma}_{\text{diff}}(k)$ denotes the spatial coherence matrix of a purely diffuse sound field that is defined mainly by the array geometry, and $\mathbf{\Phi}_{\mathbf{v}}(k,n)$ is the noise PSD matrix. Our objective is to obtain the estimate $\hat{S}(k,n)$ of the desired signal $S(k,n)$ by applying the filter weights $\mathbf{h}(k,n)$ to the microphone signals, i. e.

$$\hat{S}(k,n) = \mathbf{h}^H(k,n)\mathbf{y}(k,n). \tag{4}$$

In the following section, two such spatial filters are derived.

2.2. Derivation of the Spatial Filter

A well known filter minimizing the mean square error (MSE) between the complex Fourier coefficients of the desired and estimated signal for the given signal model (2) is proposed in [1], i. e.

$$\mathbf{h}_{\text{M-MMSE}}(k,n) = \arg\min_{\mathbf{h}} E\left\{|S(k,n) - \hat{S}(k,n)|^2\right\}, \tag{5}$$

where $E\{\cdot\}$ denotes the expectation operator. This yields the M-MMSE estimator, also known as the multichannel Wiener filter (MWF). The filter can be decomposed into a minimum variance distortionless response (MVDR) beamformer $\mathbf{h}_{\text{MVDR}}(k,n)$ and a Wiener filter $H_{\text{MMSE}}(k,n)$ that reduces the residual interference $\phi_U(k,n)$ at the output of the MVDR beamformer. The filter is given by

$$\mathbf{h}_{\text{M-MMSE}}(k,n) = \underbrace{\frac{(\phi_R\,\mathbf{\Gamma}_{\text{diff}} + \mathbf{\Phi}_{\mathbf{v}})^{-1}\mathbf{d}}{\mathbf{d}^H(\phi_R\,\mathbf{\Gamma}_{\text{diff}} + \mathbf{\Phi}_{\mathbf{v}})^{-1}\mathbf{d}}}_{\mathbf{h}_{\text{MVDR}}(k,n)} \cdot \underbrace{\frac{\phi_S}{\phi_S + \phi_U}}_{H_{\text{MMSE}}(k,n)}, \tag{6}$$

where the residual interference of the MVDR is

$$U(k,n) = \mathbf{h}_{\text{MVDR}}^H(k,n)\,[\mathbf{r}(k,n) + \mathbf{v}(k,n)], \tag{7}$$

and its PSD can be obtained by

$$\phi_U(k,n) = \mathbf{h}_{\text{MVDR}}^H(\phi_R\,\mathbf{\Gamma}_{\text{diff}} + \mathbf{\Phi}_{\mathbf{v}})\mathbf{h}_{\text{MVDR}}. \tag{8}$$

The time and frequency indices are omitted in (6) and (8) for the sake of brevity.

Rather than estimating the clean speech spectral coefficient in the MSE sense as in (5), we can also estimate the short-time spectral amplitude (STSA) of the clean speech in the MSE sense. For a single microphone, the resulting STSA estimator was proposed in [8]. By decomposing the desired signal component

$$S(k,n) = A(k,n)\,\exp\left(j\alpha(k,n)\right), \tag{9}$$

into its magnitude $A(k,n)$ and phase $\alpha(k,n)$, we obtain the M-MMSE-SA estimator as

$$\hat{A}(k,n) = E\left\{A(k,n)|\mathbf{y}(k,n)\right\}. \tag{10}$$

In a similar manner as the M-MMSE filter, the M-MMSE-SA filter can be decomposed into an MVDR beamformer and an STSA filter reducing the residual interference $U(k,n)$, i. e.

$$\mathbf{h}_{\text{M-SA}}(k,n) = \mathbf{h}_{\text{MVDR}}(k,n) \cdot H_{\text{SA}}(k,n), \tag{11}$$

where $H_{\text{SA}}(k,n)$ can be found in, for example, [9, 10]. Note that (11) has the same structure as (6), where the Wiener filter $H_{\text{MMSE}}(k,n)$ is replaced here by the STSA filter $H_{\text{SA}}(k,n)$.

The complex filter coefficients $\mathbf{h}_{\text{M-SA}}(k,n)$ are applied to the complex signal $\mathbf{y}(k,n)$. This yields an optimal estimate of the spectral amplitude $A(k,n)$ while the phase spectrum is equal to the phase spectrum of the signal at the output of the MVDR beamformer.

The filter $\mathbf{h}_{\text{MVDR}}(k,n)$, inherent in both the M-MMSE and M-MMSE-SA filters, depends on the propagation vector $\mathbf{d}(k,n)$ that is determined by the direction-of-arrival (DOA) of the desired source, and on the diffuse plus noise PSD matrix $\phi_R(k,n)\,\mathbf{\Gamma}_{\text{diff}}(k) + \mathbf{\Phi}_{\mathbf{v}}(k,n)$. These three parameters can be time varying and therefore have to be estimated for each time and frequency instant.

2.3. Parameter Estimation

The data in the REVERB challenge was obtained using uniform circular arrays. In the current implementation, we used beamspace root-MUSIC [11] to estimate the narrowband DOAs. The estimation of the diffuse sound PSD is a critical task for the dereverberation performance and the quality of the output signal of the SE stage as well as the feature compensation. For the estimation of the diffuse sound PSD, we used the maximum likelihood estimator utilizing multiple reference signals as proposed in [1]. The noise PSD matrix is estimated with the speech presence probability-based estimator proposed in [12] that updates the noise PSD matrix mainly during speech pauses.

As the used estimators are able to track changes in the sound scene rather quickly, the resulting filter is highly time varying. This results in a fast adaption and superior interference reduction performance compared to signal-independent spatial filters.

3. MODEL-BASED FEATURE COMPENSATION

This section presents a method for integration of the SE and ASR stages by using uncertainty propagation. It also details the configuration of the final system used for the REVERB challenge.

3.1. Integration through Propagation of Uncertainty

When using STFT domain SE techniques to attain robust ASR, the straightforward approach is to treat the SE and ASR stages independently, as shown in Fig. 1. The SE stage delivers a time domain estimate of the direct signal that is passed to the feature extraction process of the ASR stage, in this case the computation of the MFCC. One way to better integrate the SE and ASR stages is to calculate the MMSE estimates in the MFCC domain using the STFT domain signal model (2). This can be achieved by using MMSE-MFCC estimators [2–4], instead of STFT domain estimators like the M-MMSE-SA proposed in Section 2.

Minimizing the MSE directly in the ASR domain yields more accurate estimates. Furthermore, when the MSE in feature domain is available, as in [4], ASR model compensation can be used to further improve the ASR performance. MMSE-MFCC estimators can be seen as a particular case of uncertainty propagation of the posterior distribution associated with the MMSE estimate of the desired signal in the STFT domain, which is also known as a Wiener filter [4]. The same approach can be therefore applied to other feature extractions for which propagation formulas exist such as RASTA-LPCCs or MLP-based features.

Under the uncertainty propagation principle, it is only necessary to derive the posterior distribution of the Wiener filter for a given STFT domain speech enhancement approach. The appropriate formulas for propagation and decoding can be then applied to obtain the compensated features and optionally perform model compensation for a given ASR system. This is however not straightforward in complex speech processing schemes like the one used in the proposed system. In the following, we develop solutions to apply uncertainty propagation to the multichannel dereverberation and noise reduction approach presented in Section 2.

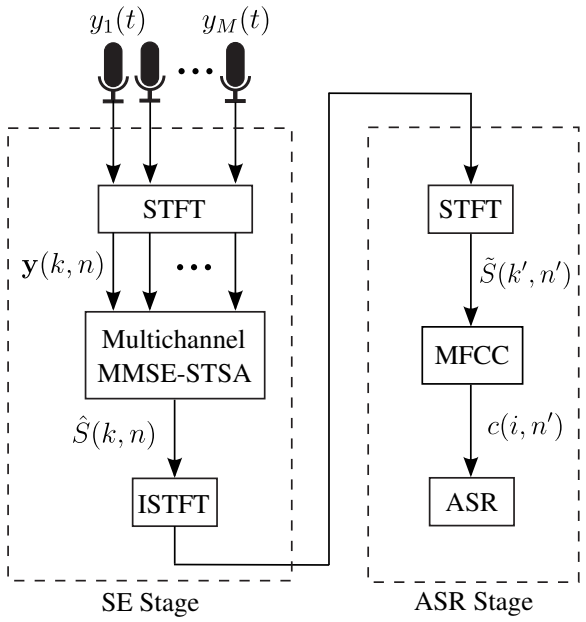

Fig. 1. Robust ASR system with decoupled SE and ASR stages. The estimated time-domain speech signal is passed from one stage to the other. Note that the MFCC acts directly on the provided STFT coefficients.

3.2. Derivation of the Posterior associated with the Multichannel MMSE Estimator

As a linear Bayesian estimator, the M-MMSE estimator has also an associated posterior distribution with mean equal to the estimated signal and variance equal to the minimum MSE of the estimated signal, see e.g. [13, Sec. 10.6]. Despite having dimension equal to the number of channels M, only a single hidden variable, the clean Fourier coefficient $S(k,n)$, is estimated per time-frequency bin. The posterior associated with the M-MMSE estimate can be thus defined for each time-frequency bin as

$$p(S(k,n)|\mathbf{y}(k,n)) \sim \mathcal{N}_{\mathbb{C}}\left(\hat{S}_{\text{M-MMSE}}(k,n), \lambda(k,n)\right). \quad (12)$$

The mean of the posterior is equal to the M-MMSE estimate

$$\hat{S}_{\text{M-MMSE}}(k,n) = \mathbf{h}^H_{\text{M-MMSE}}(k,n)\mathbf{y}(k,n), \quad (13)$$

where the M-MMSE filter is given by (6). The variance of the posterior $\lambda(k,n)$ is equal to the minimum MSE, which for the

M-MMSE is given by [13, Eqs. 10.28, 10.29]

$$\begin{aligned}\lambda(k,n) &= E\left\{|S(k,n)-\hat{S}_{\text{M-MMSE}}(k,n)|^2\right\}\\ &= \phi_S(k,n)-\mathbf{h}^H_{\text{M-MMSE}}(k,n)\mathbf{d}(k)\phi_S(k,n)\\ &= \frac{\phi_S(k,n)\phi_U(k,n)}{\phi_S(k,n)+\phi_U(k,n)}.\end{aligned} \tag{14}$$

Following [4], any non-linear single-channel MMSE estimator can be seen as the result of propagating the single-channel Wiener posterior through the corresponding non-linearity. Given the analogy between the single- and multichannel MMSE linear estimators, it is clear that the same procedure can be used to derive non-linear multichannel MMSE estimators.

Therefore, the M-MMSE-SA estimator discussed in Section 2 can be seen as a special case of propagation of the posterior of the M-MMSE through the amplitude non-linearity. Furthermore, by applying the propagation formulas in [4] to this posterior, an M-MMSE-MFCC estimator can be derived.

3.3. Propagating through Different STFT Configurations

Although possible, deriving an M-MMSE-MFCC estimator from the posterior of the M-MMSE requires the SE and ASR stages to use the same STFT. This is often undesired as the SE stages require larger processing frames or different frame rates. In the conventional M-MMSE or M-MMSE-SA cases, intermediate ISTFT+STFT separate SE and ASR stages, see Fig. 1. This allows for different STFT configurations and also helps to mitigate the artifacts created by the spatial and spectral filtering. The direct use of a M-MMSE-MFCC estimate derived from the M-MMSE in the REVERB test-set led to poor results, which might be explained by this fact.

In this paper, we propose an approach that attains uncertainty propagation through the intermediate ISTFT and STFT. This makes it possible to use different STFT resolutions for SE and ASR and to mitigate artifacts.

The idea behind the approach, here termed prior uncertainty propagation (Prior-UP), is to transform the prior distributions associated with the M-MMSE through the ISTFT+STFT and derive the posterior at the ASR side.

The M-MMSE can be expressed as single-channel Wiener filter acting on the output signal of the MVDR beamformer that is given by

$$\begin{aligned}Z(k,n) &= \mathbf{h}^H_{\text{MVDR}}(k,n)\mathbf{y}(k,n)\\ &= S(k,n)+U(k,n).\end{aligned} \tag{15}$$

Consequently, the posterior of the M-MMSE estimate (12) can be interpreted as arising from observing $Z(k,n)$ with the a priori clean speech distribution

$$p(S(k,n)) \sim \mathcal{N}_{\mathbb{C}}\left(0,\phi_S(k,n)\right), \tag{16}$$

and the a priori distribution of the residual interference

$$p(U(k,n)) \sim \mathcal{N}_{\mathbb{C}}\left(0,\phi_U(k,n)\right). \tag{17}$$

Since the ISTFT and STFT are both linear transformations, we can directly transform the priors and the observable through them. The result is each Fourier coefficient in the STFT domain of ASR of the beamformed signal along with two new a priori distributions. From these new a priori distributions and the observable, a new posterior can be derived.

One problem of this approach is that the propagation of (16) and (17) through the ISTFT+STFT induces correlations between the Fourier coefficients. To simplify computations, these correlations are ignored to maintain the same distributions on the ASR side as on the SE side but with transformed variances $\tilde{\phi}_S(k',n')$ and $\tilde{\phi}_U(k',n')$. It has to be taken into account that similar correlations, for example those induced by windowing operations, are already ignored in the conventional speech enhancement model for additive noise [8].

The only remaining problem is to compute the transformed a priori variances $\tilde{\phi}_S(k',n')$ and $\tilde{\phi}_U(k',n')$. For arbitrary STFT configurations the n'-th frame at the ASR side is going to overlap with various frames at the SE side. Consequently the frequency vector of variances $\boldsymbol{\phi}(n')$ can be computed as

$$\tilde{\boldsymbol{\phi}}(n') = \sum_{n\in \text{Ov}(n')} |\mathbf{R}_{n'-n}|^2\boldsymbol{\phi}(n), \tag{18}$$

where $\text{Ov}(n')$ are the indices of the SE frames overlapping with the n'-th ASR frame and $|\cdot|^2$ denotes the element-wise absolute square operation. The matrix $\mathbf{R}_{n'-n}$ is built by multiplying the inverse Fourier and Fourier matrices truncated to the corresponding overlap, which only depends on the time frame shift $n'-n$.

From the a priori variances $\tilde{\phi}_S(k',n')$, $\tilde{\phi}_U(k',n')$ and the MVDR beamformer output at the ASR side $\tilde{Z}(k',n')$, we can construct the posterior $p(\tilde{S}(k',n')|\mathbf{y}(k,n))$ associated with the corresponding Wiener filter as in [4].

3.4. Uncertainty Propagation and Modified Imputation

Once the a posteriori distribution of the clean STFT has been obtained at the ASR side, it can be propagated through the MFCC by applying the approach proposed in [4]. This approach only requires the assumption of log-normality for the Mel-filterbank uncertain features and has a low computational cost. The result of the propagation is a Gaussian posterior distribution in the MFCC domain

$$p(c(i,n')|\mathbf{y}(k,n)) \sim \mathcal{N}\left(c^{\text{M-MMSE}}(i,n'),\lambda^c(i,n')\right). \tag{19}$$

The mean of this distribution is the M-MMSE-MFCC estimate $c(i,n')^{\text{M-MMSE}}$ and the variance $\lambda^c(i,n')$ is the minimum MSE in the MFCC domain.

To further enhance the performance, observation uncertainty techniques [14] like uncertainty decoding (UD) [15] or modified imputation (MI) [16] can be used. In the context of the REVERB challenge, MI showed superior performance and was therefore used. This is also generally the case when using UD and MI together with uncertainty propagation, see [4] for a discussion on the topic. MI can be described as a model-based feature compensation scheme where the features are re-estimated for each Gaussian mixture of the ASR model as

$$\hat{c}_q^{\text{MI}}(i,n') = \frac{\Sigma_q(i)}{\Sigma_q(i)+\lambda^c(i,n')}\, c^{\text{M-MMSE}}(i,n') + \frac{\lambda^c(i,n')}{\Sigma_q(i)+\lambda^c(i,n')}\, \mu_q(i), \quad (20)$$

where $\mu_q(i)$ and $\Sigma_q(i)$ are the mean and variance for each Gaussian mixture of the ASR model.

It should be noted that the posterior distribution in (19) neglects the correlations between the different Mel-cepstral coefficients, which are induced by the Mel-filterbank transformation. Although MI can also be computed taking into consideration these correlations, they are ignored to reduce the computational load.

3.5. The System Submitted to the REVERB Challenge

Figure 2 depicts the structure of the final system used in the REVERB task. The system propagates the posterior associated with the M-MMSE estimate through the ISTFT, STFT and the MFCC transformation to attain a M-MMSE-MFCC estimate as described in this section. For this purpose, the MVDR beamformer is applied at the SE side to determine the variances $\phi_S(k,n)$, $\phi_U(k,n)$ and the observable $Z(k,l)$. These parameters are then propagated to the ASR side where a new posterior is computed. This posterior is then propagated into the MFCC domain to compute the M-MMSE-MFCC.

In order to improve the performance, the estimated a priori speech variance at the SE side $\hat{\phi}_S(k,n)$ is approximated directly using the M-MMSE-SA estimate computed i.e.

$$\hat{\phi}_S(k,n) \approx \hat{A}^2(k,n), \quad (21)$$

where the amplitude estimate is obtained with (10).

The resulting M-MMSE-MFCC estimate is applied to both the training and test data of the REVERB task. When processing test data, MI is optionally applied to reduce the acoustic mismatch furthermore.

All methods are real-time capable and introduce no delay, with the exception of a small delay for the computation of the deltas and accelerations, and the propagation through the ISTFT and STFT stages. The computational cost of the M-MMSE-SA estimator is manageable for real-time application. The DOA estimator is computationally complex due to the required polynomial rooting, but also less complex estimators can be used.

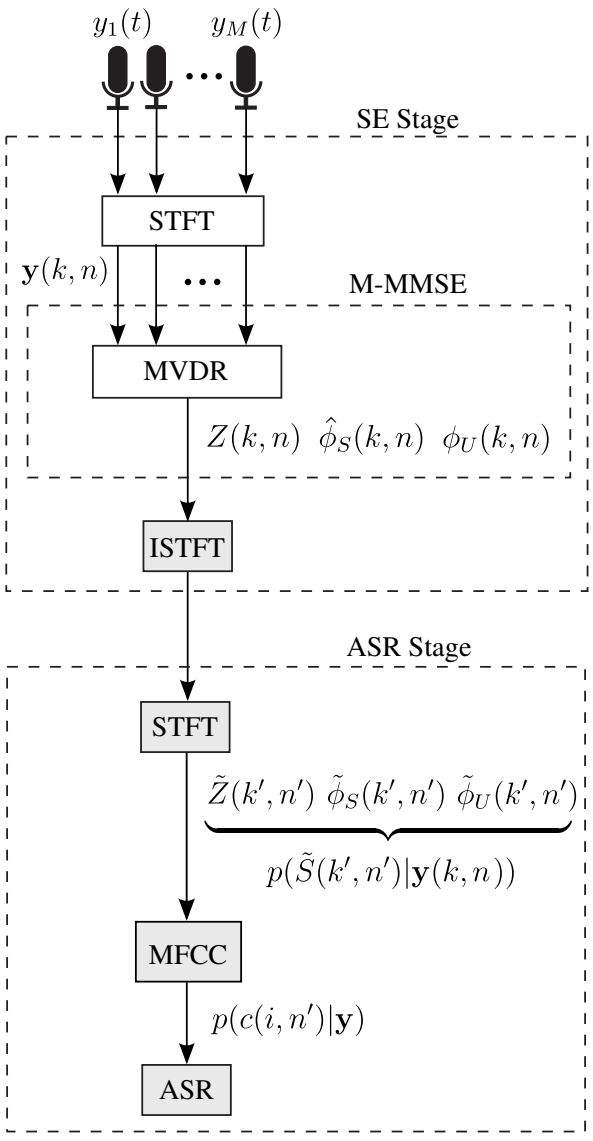

Fig. 2. Robust ASR system with tight integration of SE and ASR stages. The M-MMSE is propagated through ISTFT, STFT and MFCC transformations. Note that the MFCC acts directly on the provided STFT. Grey blocks denote stages through which there is uncertainty propagation.

Computing a MMSE-MFCC estimate represents a small increase in cost compared to STFT domain MMSE estimates, see [4]. The introduction of prior propagation through the ISTFT and STFT stages increases this cost as it implies a variable number of matrix multiplications per frame. Nevertheless, the cost of prior propagation can be reduced when the whole utterance is available. In this case each matrix $\mathbf{R}_{n'-n}$ only needs to be computed once. It should be noted that, in this case, the system is not any more real-time capable as the

whole speech utterance has to be processed before the priors can be propagated.

4. EXPERIMENTAL RESULTS

4.1. Experimental Setup

To test the proposed system, the REVERB ASR task was used [17]. The REVERB ASR task provides two training sets based on the WSJCAM0 corpus [18], a medium vocabulary size read news database in British English. The first set consists of the WSJCAM0 clean-training set. The second is a 8 microphone multi-condition version of the same set obtained by convolving the clean training set with recorded room impulse responses and adding background noise. Regarding development and evaluation sets, both simulated data, based on the WSJCAM0 and recorded data from the MC-WSJ-AV corpus [19], are provided. The former includes three different rooms (small, medium, large) and two different recording distances (near, far) whereas the latter includes one single room and two possible distances. All development and evaluation sets were recorded using a circular microphone array with $M = 8$ and a radius of 10 cm. Train and test scripts were provided for the widely used HTK toolbox [20].

Regarding the SE configuration, the STFT was implemented with a square-root Hann window of length of 32 ms, 50% overlap and 512 samples FFT length for the given sampling frequency of 16 kHz. The input and reverberant PSD matrices were estimated recursively with a time constant of 70 ms. For the noise PSD matrix, a time constant of 150 ms was used and a fixed a priori speech presence probability of 0.99 was assumed for the noise estimation procedure [21] to prevent leakage of speech components into the noise PSD estimate.

Regarding ASR side configuration, the defaults provided in the REVERB challenge were used. The STFT configuration corresponded thus to a window of 25 ms with 15 ms overlap and 512 samples FFT length with zero padding. To perform uncertainty propagation through the MFCC transformation, a Matlab implementation compatible with the HTK configuration was used[1]. Full covariance propagation was used as described in [4]. Note that despite the fact that the original challenge MFCC configuration was used, the Matlab implementation led to small differences in the baseline results. To perform MI on HTK, modified binaries of the recognition function HVite were compiled using the available patches[2].

4.2. Comparison Tests

The final system presented to the REVERB challenge was decoupled into three main categories and each tested individually in all development set scenarios.

[1] See https://github.com/ramon-astudillo/stft_up_tools

[2] See http://www.astudillo.com/ramon/research/stft-up/

The first category corresponded to STFT domain techniques with no integration with the ASR side. In this category the M-MMSE-SA estimator as described in Section 2 was used. The second category corresponded to feature compensation techniques: The M-MMSE-MFCC estimator described in Section 3.5, including prior uncertainty propagation (Prior-UP) and the re-estimation of $\phi_S(k, n)$, was used. The last category corresponded to feature compensation methods using model-side information. Here, the previous M-MMSE-MFCC was combined with MI.

It should be noted that the eight microphones were always used when available. This included all test and evaluation sets and the multi-condition training set.

4.3. Analysis of the Results

Tables 1 and 2 contain the results for the development set on clean and multi-condition training sets respectively. Table 3 contains the results on the evaluation set that were submitted to the REVERB challenge.

All the proposed methods provide substantive improvements over the baseline without processing across all conditions with the exception of the smaller Room 1 and clean-training conditions. The reason for this might be artifacts caused by overestimation of the diffuse and noise PSDs in high SNR conditions. This can be mitigated by more advanced implementations of the estimators, but was was not implemented due to time constraints. For larger rooms with stronger reverberation, the processing provides large reductions of WER, particularly for clean-training. WER reduction against the baseline for the multi-condition training set, displayed in Table 2, are smaller but consistent.

Regarding the comparison across the different categories considered, the use of the M-MMSE-MFCC feature compensation approach consistently outperformed the M-MMSE-SA estimator in STFT domain. The only exception was again the near condition for Room 1 on the multi-condition scenario. Average scores are however favourable for the M-MMSE-MFCC approach both in real and simulated scenarios. It has to be taken into account that the M-MMSE-MFCC uses indirectly also the M-MMSE-SA for a refined estimation of $\phi_S(k, n)$, as explained in Section 3.5.

Regarding the third category considered, corresponding to feature compensation methods using model-side information, results are mixed. Although the results when using MI are better on average in all test scenarios, differences are rather small in the simulated data sets. Furthermore, for some room configurations, in particular larger ones, MI fails to outperform the M-MMSE-MFCC estimator alone. Interestingly, MI provides large improvements in situations where a large mismatch can be expected. This includes the Room 1 and the recorded data, where signal processing artifacts can lead to a mismatch between the clean signal and its estimate. In those scenarios MI provides WER reductions comparable to those

Table 1. WER scores for the REVERB development sets and the clean-training HTK baseline, *no* CMLLR used. Integration modes are: Uncertainty propagation into MFCC domain including prior propagation through resynthesis (Prior-UP) and modified imputation (MI). The best results are highlighted in bold.

		Simulated Data							Recorded Data		
		Room 1		Room 2		Room 2		Avg.	Room 1		Avg.
	Integration	Near	Far	Near	Far	Near	Far		Near	Far	
No Processing	None	**14.43**	25.15	43.46	86.64	52.20	88.40	51.67	88.33	87.56	87.94
M-MMSE-SA	None	19.25	27.65	18.68	36.55	24.60	47.16	28.97	58.27	61.18	59.71
M-MMSE-MFCC	Prior-UP	16.94	23.57	17.20	**33.47**	**20.80**	**44.29**	26.03	54.15	54.41	54.27
M-MMSE-MFCC	Prior-UP, MI	15.34	**21.85**	**16.96**	33.67	20.99	45.03	**25.64**	**51.72**	**50.31**	**51.02**

Table 2. WER scores for the REVERB development sets and the multi-condition-training HTK baseline, *no* CMLLR used. Integration modes are: Uncertainty propagation into MFCC domain including prior propagation through resynthesis (Prior-UP) and modified imputation (MI). The best results are highlighted in bold.

		Simulated Data							Recorded Data		
		Room 1		Room 2		Room 2		Avg.	Room 1		Avg.
	Integration	Near	Far	Near	Far	Near	Far		Near	Far	
No Processing	None	16.54	18.88	23.37	43.18	27.40	46.79	29.34	52.90	50.79	51.85
M-MMSE-SA	None	15.46	17.75	17.23	26.13	18.40	30.91	20.97	42.48	41.49	41.98
M-MMSE-MFCC	Prior-UP	15.73	16.79	14.81	**21.99**	18.05	**27.35**	19.11	40.61	39.23	39.92
M-MMSE-MFCC	Prior-UP, MI	**14.70**	**16.74**	**14.30**	23.05	**17.80**	27.42	**19.00**	**39.74**	**37.18**	**38.46**

Table 3. Submitted WER scores for the REVERB evaluation set and the multi-condition HTK baseline.

		Simulated Data							Recorded Data		
		Room 1		Room 2		Room 2		Avg.	Room 1		Avg.
	Integration	Near	Far	Near	Far	Near	Far		Near	Far	
M-MMSE-MFCC	Prior-UP, MI	17.18	18.18	15.68	23.91	17.8	28.08	20.14	41.14	42.3	41.72
+CMLLR	Prior-UP	14.59	17.28	15.26	20.53	16.12	24.11	17.97	35.64	37.54	36.59

attained when switching from the MSTSA to the M-MMSE-MFCC system proposed for the challenge.

Given the obtained results for the development set, it was decided to submit the system based on the M-MMSE-MFCC+MI to the challenge. Since Cepstral Mean Subtraction is used in the default HTK configuration the results were submitted under the per-file batch category. It should be noted, however, that the methods are real-time capable. In addition to this system, a M-MMSE-MFCC estimator using CMLLR was submitted to the full-batch category. MI was not included in this last submission due to time and implementation constraints, although it is likely to bring no improvement in a full-batch adaptation scenario. Results are summarized in Table 3.

5. CONCLUSIONS

We proposed a robust ASR system for the REVERB challenge based on a multichannel feature compensation approach. The system integrated an M-MMSE estimator in the ASR side through uncertainty propagation. For this purpose the posterior distribution associated with the M-MMSE estimate was used. A solution was also proposed to allow different STFT parameters for the SE and ASR when using propagation. The resulting system outperforms a recent STFT domain multichannel STSA estimation method while remaining real-time capable. The use of model-based feature compensation in situations of high mismatch is also shown to be advantageous.

6. REFERENCES

[1] S. Braun and E. A. P. Habets, "Dereverberation in noisy environments using reference signals and a maximum likelihood estimator," in *Proc. European Signal Processing Conf. (EUSIPCO)*, Marrakesh, Morocco, Sept. 2013.

[2] D. Yu, L. Deng, J. Droppo, J. Wu, Y. Gong, and A. Acero, "A Minimum-Mean-Square-Error Noise Re-

duction Algorithm on Mel-Frequency Cepstra for Robust Speech Recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP).*, 2008, pp. 4041–4044.

[3] K. Paliwal A. Stark, "MMSE estimation of log-filterbank energies for robust speech recognition," *Speech Communication*, vol. 53 (3), pp. 403–416, 2011.

[4] R. F. Astudillo and R. Orglmeister, "Computing MMSE estimates and residual uncertainty directly in the feature domain of ASR using STFT domain speech distortion models," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 21, no. 5, pp. 1023 – 1034, May 2013.

[5] R. F. Astudillo, D. Kolossa, A. Abad, S. Zeiler, R. Saeidi, P. Mowlaee, J. P. Neto, and R. Martin, "Integration of beamforming and uncertainty-of-observation techniques for robust ASR in multi-source environments," *Computer Speech & Language*, vol. 27, no. 3, pp. 837–850, May 2013.

[6] O. Thiergart, M. Taseska, and E. A. P. Habets, "An informed MMSE filter based on multiple instantaneous direction-of-arrival estimates," in *Proc. European Signal Processing Conf. (EUSIPCO)*, Marrakesh, Morocco, Sept. 2013, IEEE.

[7] R. Maas, E. A P Habets, A. Sehr, and W. Kellermann, "On the application of reverberation suppression to robust speech recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP).*, 2012, pp. 297–300.

[8] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 32, no. 6, pp. 1109–1121, Dec. 1984.

[9] R. Balan and J. Rosca, "Microphone array speech enhancement by Bayesian estimation of spectral amplitude and phase," in *Proceedings of the Sensor Array and Multichannel Signal Processing Workshop*, Aug. 2002, pp. 209–213.

[10] S. Lefkimmiatis and P. Maragos, "A generalized estimation approach for linear and nonlinear microphone array post-filters," *Speech Communication*, vol. 49, no. 7-8, pp. 657–666, July 2007.

[11] M. D. Zoltowski and C. P. Mathews, "Direction finding with uniform circular arrays via phase mode excitation and beamspace root-MUSIC," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP).*, 1992, vol. 5, pp. 245–248.

[12] M. Souden, J. Chen, J. Benesty, and S. Affes, "An integrated solution for online multichannel noise tracking and reduction," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 7, pp. 2159–2169, 2011.

[13] Steven M. Kay, *Fundamentals of Statistical Signal Processing*, Prentice Hall signal processing series, 1993.

[14] D. Kolossa and R. Haeb-Umbach, Eds., *Robust Speech Recognition of Uncertain or Missing Data - Theory and Applications*, Springer, 2011.

[15] J. Droppo, A. Acero, and Li Deng, "Uncertainty decoding with SPLICE for noise robust speech recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP).*, 2002, vol. 1, pp. I–57–I–60 vol.1.

[16] D. Kolossa, A. Klimas, and R. Orglmeister, "Separation and robust recognition of noisy, convolutive speech mixtures using time-frequency masking and missing data techniques," in *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, Oct. 2005, pp. 82–85.

[17] K. Kinoshita, M. Delcroix, T. Yoshioka, T. Nakatani, E. Habets, R. Haeb-Umbach, V. Leutnant, A. Sehr, W Kellermann, R. Maas, S. Gannot, and B. Raj, "The REVERB Challenge: A Common Evaluation Framework for Dereverberation and Recognition of Reverberant Speech," in *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA-13)*, 2013.

[18] T. Robinson, J. Fransen, D. Pye, J. Foote, and S. Renals, "WSJCAMO: a British English speech corpus for large vocabulary continuous speech recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP).*, 1995, vol. 1, pp. 81–84.

[19] M. Lincoln, I. McCowan, J. Vepa, and H.K. Maganti, "The multi-channel Wall Street Journal audio visual corpus (MC-WSJ-AV): specification and initial experiments," in *Proceedings of the 2005 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU-05)*, 2005, pp. 357–362.

[20] S. Young, *The HTK Book (for HTK Version 3.4)*, Cambridge University Engineering Department., 2006.

[21] T. Gerkmann and R. Hendriks, "Unbiased MMSE-based noise power estimation with low complexity and low tracking delay," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 4, pp. 1383 –1393, May 2012.