Improving Reverberant VTS for Hands-free Robust Speech Recognition

Y.-Q. Wang, M. J. F. Gales

Cambridge University Engineering Department
Trumpington St., Cambridge CB2 1PZ, U.K.
{yw293, mjfg}@eng.cam.ac.uk

***Abstract*—Model-based approaches to handling additive background noise and channel distortion, such as Vector Taylor Series (VTS), have been intensively studied and extended in a number of ways. In previous work, VTS has been extended to handle both reverberant and background noise, yielding the Reverberant VTS (RVTS) scheme. In this work, rather than assuming the observation vector is generated by the reverberation of a sequence of background noise corrupted speech vectors, as in RVTS, the observation vector is modelled as a superposition of the background noise and the reverberation of clean speech. This yields a new compensation scheme RVTS Joint (RVTSJ), which allows an easy formulation for joint estimation of both additive and reverberation noise parameters. These two compensation schemes were evaluated and compared on a simulated reverberant noise corrupted AURORA4 task. Both yielded large gains over VTS baseline system, with RVTSJ outperforming the previous RVTS scheme.**

I. Introduction

Hands-free speech recognition using distant microphones is useful for many applications, e.g., voice control of consumer electronics, automatic meeting transcription and speech dialogue systems. Distant-talking automatic speech recognition (ASR) systems need to handle both the background noise and the reverberant noise. The background noise is caused by other interfering sources and is usually additive in the linear spectrum domain, while the reverberation is usually caused by multiple acoustic paths of sound waves from the source to the microphone. The reverberation effect can be described as a convolution of clean speech with Room Impulse Response(RIR). The RIR is usually characterised by the so called reverberant time, T_{60}, which is the time needed for reflections of a direct sound to decay by 60dB to the level of the direct sound. In a reverberant environment, the T_60 value is significant longer than the analysis window used for feature extraction in ASR. Thus the observed feature vector becomes a superposition of multiple delayed and attenuated copies of previous clean speech.

There are several approaches in the literature to handle reverberant noise. Signal processing methods like beamforming [1] and inverse filtering [2] can be used to clean the reverberant speech signals, or speech feature vectors can be enhanced [3], [4]. Recently, model-based approaches to robust speech recognition, e.g., Parallel Model Combination (PMC) [5] and Vector Taylor Series (VTS) [6], have been investigated to handle the additive and convolutional noise and were extended in a number of ways. Model-based approaches have also been extended to deal with the reverberant noise, e.g., in [7], [8] and [9], acoustic models are compensated using PMC. In [10], Reverberant VTS (RVTS) was proposed, where model-based VTS compensation was extended to handle the reverberant noise, as well as the background noise. RVTS allows the compensation formula for all the model parameters to be defined, and the parameters of the reverberant noise model can be estimated using maximum likelihood (ML) criterion. To handle background noise in a reverberant environment, a series of approximations were made such that the observation vector can be described as the reverberation of a sequence of background noise corrupted speech vectors. The background and reverberant noise were then estimated in a sequential way. In this work, an alternative assumption about the relationship of the additive and reverberant noise is explored, where the observation vector is assumed to be a combination of the reverberation of the clean speech and additive noise. Based on this assumption, an extension of RVTS, RVTS Joint (RVTSJ) compensation scheme is proposed. Compared with the RVTS scheme, the noise parameters for RVTSJ, including the additive and reverberant noise parameters, are estimated jointly, rather than sequentially.

The rest of the paper is organised as follows. The next section will discuss the mismatch functions that can be used to describe the impact of environment, as well as the extra statistics needed to model the dependency caused by reverberation. Section III describes the RVTS and its extension RVTSJ model compensation schemes. Noise estimation is presented in section IV. Experiments and results are discussed in section V with conclusion and future work in section VI.

II. Mismatch Functions

A. Additive Noise and Convolutional Distortion

In the time domain, the standard form used to describe the additive noise $n(\tau)$ and short-term convolutional noise $h(\tau)$ corrupting the clean speech $x(\tau)$ is

$$y(\tau) = h(\tau) * x(\tau) + n(\tau) \tag{1}$$

where the length of $h(\tau)$ is less than the length of analysis window in feature extraction. After a series of approximations, the mismatch function in the cepstral domain, relating the corrupted speech MFCCs $\boldsymbol{y}_t$ to the clean speech vector, $\boldsymbol{x}_t$, is

written as:

$$\begin{aligned}\boldsymbol{y}_t =& \mathbf{C}\log\left(\exp(\mathbf{C}^{-1}(\boldsymbol{x}_t+\boldsymbol{\mu}_{\mathtt{h}}))+\exp(\mathbf{C}^{-1}\boldsymbol{n}_t)\right)\\ =& \mathbf{f}(\boldsymbol{x}_t,\boldsymbol{\mu}_{\mathtt{h}},\boldsymbol{n}_t)\,,\end{aligned} \tag{2}$$

where $\boldsymbol{n}_t$ is the noise coefficient, $\boldsymbol{\mu}_{\mathtt{h}}$ the convolutional noise, and $\mathbf{C}$ the (truncated) DCT matrix. Note that this mismatch function assumes the noise and speech are linearly additive in the magnitude domain. Combining noise and speech in other domains is possible, which only requires a simple change: the DCT matrix $\mathbf{C}$ is replaced by $\gamma\mathbf{C}$, where $\gamma = 1$ represents the magnitude domain, and $\gamma = 2$ the power domain. Given this mismatch function, VTS can be used to yield a linear approximation. For a clean speech vector $\boldsymbol{x}_t$, generated from component m, its noise-corrupted observation $\boldsymbol{y}_t$ is

$$\boldsymbol{y}_t|m \approx \mathbf{f}(\boldsymbol{\mu}_{\mathtt{sx}}^{(m)},\boldsymbol{\mu}_{\mathtt{sn}},\boldsymbol{\mu}_{\mathtt{h}}) + \mathbf{J}_{\mathtt{x}}^{(m)}(\boldsymbol{x}_t - \boldsymbol{\mu}_{\mathtt{sx}}^{(m)}) + \mathbf{J}_{\mathtt{n}}^{(m)}(\boldsymbol{n}_t - \boldsymbol{\mu}_{\mathtt{sn}}) \tag{3}$$

where subscript $\mathtt{s}$ indicates the static parameters, and

$$\mathbf{J}_{\mathtt{x}}^{(m)} = \frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}}\Big|_{\boldsymbol{\mu}_{\mathtt{sx}}^{(m)},\boldsymbol{\mu}_{\mathtt{sn}},\boldsymbol{\mu}_{\mathtt{h}}}, \quad \mathbf{J}_{\mathtt{n}}^{(m)} = \mathbf{I} - \mathbf{J}_{\mathtt{x}}^{(m)} \tag{4}$$

Using this approximation, the static model parameters can be compensated via

$$\boldsymbol{\mu}_{\mathtt{sy}}^{(m)} = \mathbf{f}(\boldsymbol{\mu}_{\mathtt{sx}}^{(m)},\boldsymbol{\mu}_{\mathtt{h}},\boldsymbol{\mu}_{\mathtt{n}}) \tag{5}$$

$$\boldsymbol{\Sigma}_{\mathtt{sy}}^{(m)} = \mathrm{diag}\left(\mathbf{J}_{\mathtt{x}}^{(m)}\boldsymbol{\Sigma}_{\mathtt{sx}}^{(m)}\mathbf{J}_{\mathtt{x}}^{(m)\mathsf{T}} + \mathbf{J}_{\mathtt{n}}^{(m)}\boldsymbol{\Sigma}_{\mathtt{sn}}\mathbf{J}_{\mathtt{n}}^{(m)\mathsf{T}}\right) \tag{6}$$

For delta model parameters, the continuous time approximation assumption[11] is used, yielding:

$$\boldsymbol{\mu}_{\Delta\mathtt{y}}^{(m)} = \mathbf{J}_{\mathtt{x}}^{(m)}\boldsymbol{\mu}_{\Delta\mathtt{x}}^{(m)} \tag{7}$$

$$\boldsymbol{\Sigma}_{\Delta\mathtt{y}}^{(m)} = \mathrm{diag}(\mathbf{J}_{\mathtt{x}}^{(m)}\boldsymbol{\Sigma}_{\Delta\mathtt{x}}^{(m)}\mathbf{J}_{\mathtt{x}}^{(m)\mathsf{T}} + \mathbf{J}_{\mathtt{n}}^{(m)}\boldsymbol{\Sigma}_{\Delta\mathtt{n}}\mathbf{J}_{\mathtt{n}}^{(m)\mathsf{T}}) \tag{8}$$

where Δ in the subscripts denotes static and delta parameters. The delta-delta parameters are also compensated in a similar form. It is necessary to estimate the noise model parameters $\boldsymbol{\mu}_{\mathtt{sn}}, \boldsymbol{\mu}_{\mathtt{h}}, \boldsymbol{\Sigma}_{\mathtt{n}}, \boldsymbol{\Sigma}_{\Delta\mathtt{n}}$ and $\boldsymbol{\Sigma}_{\Delta^2\mathtt{n}}$. These can be done via ML estimates, e.g., [12], [13].

B. Reverberant Noise

In a reverberant noise environment, if only a single multi-path term is considered, the signal $z(\tau)$, corrupted by the reverberant and background additive noise, may be expressed as

$$z(\tau) = h_r(\tau) * x(\tau) + n_r(\tau) \tag{9}$$

where $n_r(\tau)$ and $h_r(\tau)$ are the additive and reverberant noise term, including intra-frame distortion and inter-frame reverberation, i.e.,

$$h_r(\tau) = h(\tau) + h_1(\tau); \quad n_r(\tau) = n(\tau) + n_1(\tau) \tag{10}$$

$h_1(\tau)$ is usually caused by late-reflection of indirect acoustic path from the speaker to the microphone, whose length usually ranges from 200ms to 1s or more. Since this is much longer than the length of analysis window used for feature extraction (typically 25ms), the effect of reverberant noise cannot be described as a simple bias term in the cepstral domain.

In previous work [10], the following approximations were made to simplify the overall expressions:

$$h_1(\tau) \approx \tilde{h}_1(\tau) * h(\tau); \quad n_1(\tau) \approx \tilde{h}_1(\tau) * n(\tau) \tag{11}$$

It is then possible to write

$$z(\tau) = (1 + \tilde{h}_1(\tau)) * y(\tau) \tag{12}$$

By ignoring the cross-term correlation, the effect of reverberant distortion in the cepstral domain can be approximated as a combination of $n+1$ frame-level distortion terms, $\tilde{\boldsymbol{\mu}}_{\mathtt{l}} = [\tilde{\boldsymbol{\mu}}_{\mathtt{l}0}^{\mathsf{T}}, \cdots, \tilde{\boldsymbol{\mu}}_{\mathtt{l}n}^{\mathsf{T}}]^{\mathsf{T}}$, acting on a set of preceding noise-corrupted MFCC features, $\boldsymbol{y}_t, \cdots, \boldsymbol{y}_{t-n}$, i.e.,

$$\begin{aligned}\boldsymbol{z}_t &= \mathbf{C}\log\left(\sum_{\delta=0}^{n}\exp\left(\mathbf{C}^{-1}(\boldsymbol{y}_{t-\delta}+\tilde{\boldsymbol{\mu}}_{\mathtt{l}\delta})\right)\right)\\ &= \tilde{\mathbf{g}}(\boldsymbol{y}_t,\cdots,\boldsymbol{y}_{t-n},\tilde{\boldsymbol{\mu}}_{\mathtt{l}})\end{aligned} \tag{13}$$

The Mismatch function in Eq. (13) allows reverberant noise compensation to be built on top of the VTS-compensated models. In this work, rather than using the approximation in Eq. (11), the mismatch function in Eq. (9) is used, which directly links the clean speech, additive noise and noise corrupted observation. Again, by ignoring the cross-term correlation, the corresponding mismatch function in the cepstral domain is written as

$$\begin{aligned}\boldsymbol{z}_t &= \mathbf{C}\log\left(\sum_{\delta=0}^{n}\exp\left(\mathbf{C}^{-1}(\boldsymbol{x}_{t-\delta}+\boldsymbol{\mu}_{\mathtt{l}\delta})\right)+\exp\left(\mathbf{C}^{-1}\boldsymbol{n}_t\right)\right)\\ &= \mathbf{g}(\boldsymbol{x}_t,\cdots,\boldsymbol{x}_{t-n},\boldsymbol{\mu}_{\mathtt{l}},\boldsymbol{n}_t)\end{aligned} \tag{14}$$

This mismatch function assumes the corrupted observation vector is a combination of additive noise and several delayed-and-attenuated copies of previous clean speech vectors $\boldsymbol{x}_t, \cdots, \boldsymbol{x}_{t-n}$, while the mismatch function in Eq. (13) assumes the observation is generated by the reverberation of additive noise corrupted speech vectors $\boldsymbol{y}_t, \ldots, \boldsymbol{y}_{t-n}$.

Similar to the VTS case, the mismatch functions in Eqs. (13-14) assume the linear combination in the magnitude domain. Power domain combination is also possible by setting $\gamma = 2$. The above mismatch functions describe the impact of reverberant and additive noise on the static features. It is also possible to derive the mismatch functions for delta and delta-delta features using continuous time approximation.

C. Model Statistics

For the above mismatch functions, the reverberant and additive noise corrupted static speech frame is a function of a window of $n+1$ clean speech frames $\boldsymbol{x}_t, \ldots, \boldsymbol{x}_{t-n}$ and additive noise $\boldsymbol{n}_t$ (or $\boldsymbol{n}_t, \ldots, \boldsymbol{n}_{t-n}$ to yield $\boldsymbol{y}_t, \ldots, \boldsymbol{y}_{t-n}$). Therefore, additional model statistics are needed to model this dependency.

Figure 1 shows the generating process of the reverberant observations (ignoring the dynamic parameters) according to the mismatch function in Eq. (14). Inference on this dynamic Bayesian network (DBN) is not practical as the number of states and components affecting the current state grows

exponentially. Approximations to this form are possible. For example, in [9] the Viterbi decoding algorithm is modified where model parameters are adapted based on the current best partial path. The model adaptation is done at each frame, which results in a large amount of computation. Alternatively, the model parameters can be adapted prior to recognition, based on the estimated preceding states, either the intra-phoneme preceding states or inferred from the context of biphone [8] or triphone [7] models. However, it is difficult to infer a long preceding state sequence in this way, especially when tied-state cross word triphone models are used.

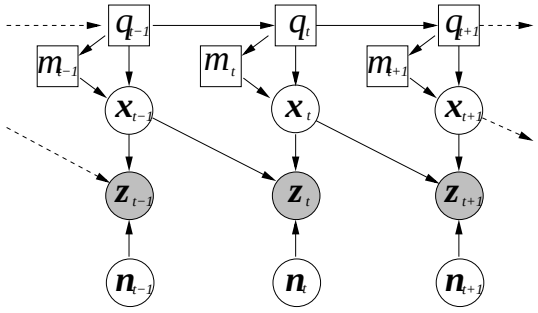

Fig. 1. Reverberant dynamic Bayesian network. q_t and m_t denote the state and component at time t,$n = 1$.

Alternatively, another form of approximation was proposed in [10] and is also used in this work. The DBN is shown in Figure 2. In this approximation, rather than an explicit dependence on the previous observation or states, the observation vector $\boldsymbol{z}_t$ is assumed to depend on an extended observation vector $\overline{\boldsymbol{x}}_t$. In this way, the standard Viterbi algorithm can be used for inference. This approximation results in two forms of smoothing. First statistics are smoothed over all possible previous states. This effect is moderated for the context dependent models as the left context automatically limits the range of possible states. The second impact is the smoothing over components for the previous state. It is worth noting that this is exactly the same form of approximation that is used in deriving the standard dynamic parameters.

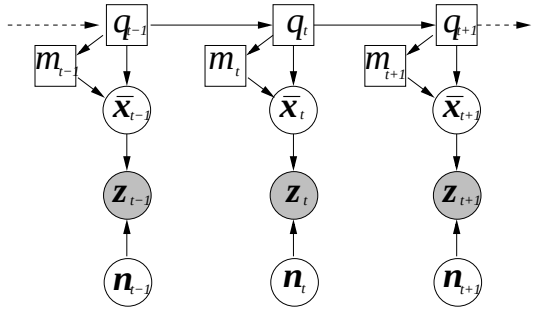

Fig. 2. Approximate reverberant environment dynamic Bayesian network.

It is also important to decide which form of the probability distribution, $p(\overline{\boldsymbol{x}}_t|m_t = m)$, to use. To ensure that if there is no reverberant noise, the compensated model becomes the original model, the following form is used:

$$\overline{\boldsymbol{x}}_t = \begin{bmatrix} \boldsymbol{x}_t \\ \Delta \boldsymbol{x}_t \\ \Delta^2 \boldsymbol{x}_t \\ \tilde{\boldsymbol{x}}_t \end{bmatrix} = \mathbf{W} \begin{bmatrix} \boldsymbol{x}_{t+w} \\ \cdots \\ \boldsymbol{x}_{t-n-w} \end{bmatrix} \sim \mathcal{N}(\overline{\boldsymbol{\mu}}_{\mathtt{x}}^{(m)}, \overline{\boldsymbol{\Sigma}}_{\mathtt{x}}^{(m)}) \quad (15)$$

where w is the window size to calculate the dynamic parameters, $\tilde{\boldsymbol{x}}_t$ can be any vector, provided $\mathbf{W}$ is square and invertible, and $\overline{\boldsymbol{\Sigma}}_{\mathtt{x}}^{(m)}$ is a diagonal matrix. Using this representation, it is simple to derive the clean speech statistics. For example, the mean and covariance of spliced frames $\boldsymbol{x}_{\mathtt{e}}$(ranging from t to $t-n$), $\boldsymbol{\mu}_{\mathtt{xe}}^{(m)}$ and $\boldsymbol{\Sigma}_{\mathtt{xe}}^{(m)}$, can be derived by

$$\boldsymbol{\mu}_{\mathtt{xe}}^{(m)} = \mathbf{P}\overline{\boldsymbol{\mu}}_{\mathtt{x}}^{(m)} \; ; \quad \boldsymbol{\Sigma}_{\mathtt{xe}}^{(m)} = \mathbf{P}\overline{\boldsymbol{\Sigma}}_{\mathtt{x}}^{(m)}\mathbf{P}^{\mathsf{T}} \quad (16)$$

where $\mathbf{P}$ is the matrix that maps $\overline{\boldsymbol{x}}_t$ to $\boldsymbol{x}_{\mathtt{e}}$. Since the deltas of $\boldsymbol{x}_{\mathtt{e}}$, $\Delta\boldsymbol{x}_{\mathtt{e}}$, are also linear combination of $\overline{\boldsymbol{x}}_t$, the mean and covariance of $\Delta\boldsymbol{x}_{\mathtt{e}}$, $\boldsymbol{\mu}_{\Delta\mathtt{xe}}^{(m)}$ and $\boldsymbol{\Sigma}_{\Delta\mathtt{xe}}^{(m)}$ can be obtained in a similar way.

The above expressions describe the derivation of the "clean" statistics required by the mismatch function in Eq. (14). Given the noise model parameters $\boldsymbol{\mu}_{\mathtt{sn}}, \boldsymbol{\mu}_{\mathtt{h}}, \boldsymbol{\Sigma}_{\mathtt{n}}$, it is possible to derive the "noisy" statistics required by the mismatch function in Eq. (13). To avoid computing a large number of Jacobian matrices, linear approximation is used: for example, the mean of $\boldsymbol{y}_{t-\delta}$, $\boldsymbol{\mu}_{\mathtt{y}\delta}^{(m)}$, is given by

$$\boldsymbol{\mu}_{\mathtt{y}\delta}^{(m)} = \mathbf{f}(\boldsymbol{\mu}_{\mathtt{x}}^{(m)}, \boldsymbol{\mu}_{\mathtt{h}}, \boldsymbol{\mu}_{\mathtt{sn}}) + \mathbf{J}_{\mathtt{x}}^{(m)}(\boldsymbol{\mu}_{\mathtt{x}\delta}^{(m)} - \boldsymbol{\mu}_{\mathtt{x}}^{(m)}) \quad (17)$$

where $\boldsymbol{\mu}_{\mathtt{x}\delta}^{(m)}$ is the mean of $\boldsymbol{x}_{t-\delta}$, conditioning on the component m. Once $\boldsymbol{\mu}_{\mathtt{y}\delta}^{(m)}(\delta = -w, \ldots n+w)$ are known, the noisy delta statistics, $\boldsymbol{\mu}_{\Delta\mathtt{ye}}^{(m)}$, can be obtained in the same way as $\boldsymbol{\mu}_{\Delta\mathtt{xe}}^{(m)}$.

III. Reverberant VTS Compensation

Given the mismatch functions in Eq. (13) and Eq. (14) and the statistics described in the previous section, it is possible to extend the use of VTS to handle reverberant noise. In previous work [10], the mismatch function in Eq. (13) is expanded about model parameters and current noise parameters, i.e.,

$$\boldsymbol{z}_t|m = \tilde{\mathbf{g}}(\boldsymbol{y}_{\mathtt{e}}, \tilde{\boldsymbol{\mu}}_{\mathtt{l}}) \approx \tilde{\mathbf{g}}(\boldsymbol{\mu}_{\mathtt{ye}}^{(m)}, \tilde{\boldsymbol{\mu}}_{\mathtt{l}}) + \mathbf{J}_{\mathtt{ye}}^{(m)}(\boldsymbol{y}_{\mathtt{e}} - \boldsymbol{\mu}_{\mathtt{ye}}^{(m)}) \quad (18)$$

where $\boldsymbol{y}_{\mathtt{e}}$ is the stacked noisy frames $\boldsymbol{y}_t, \ldots, \boldsymbol{y}_{t-n}$ and

$$\mathbf{J}_{\mathtt{ye}}^{(m)} = [\mathbf{J}_{\mathtt{y}0}^{(m)}, \cdots, \mathbf{J}_{\mathtt{y}n}^{(m)}] \; ; \quad \mathbf{J}_{\mathtt{y}\delta}^{(m)} = \frac{\partial \tilde{\mathbf{g}}}{\partial \boldsymbol{y}_{t-i}}\Big|_{\boldsymbol{\mu}_{\mathtt{ye}}^{(m)}, \tilde{\boldsymbol{\mu}}_{\mathtt{l}}} \quad (19)$$

With this expansion, as well as the continuous time approximation, the mean is compensated using

$$\boldsymbol{\mu}_{\mathtt{sz}}^{(m)} = \tilde{\mathbf{g}}(\boldsymbol{\mu}_{\mathtt{ye}}^{(m)}, \tilde{\boldsymbol{\mu}}_{\mathtt{l}}) \quad (20)$$

$$\boldsymbol{\mu}_{\Delta\mathtt{z}}^{(m)} = \mathbf{J}_{\mathtt{ye}}^{(m)}\boldsymbol{\mu}_{\Delta\mathtt{ye}}^{(m)} \quad (21)$$

This form of compensation is referred to as RVTS.

A similar approximation can be carried out for the mismatch function in Eq. (14). Performing an expansion of the function $\mathbf{g}()$ around $\boldsymbol{\mu}_{\mathtt{x}}^{(m)}$, $\boldsymbol{\mu}_{\mathtt{l}}$ and $\boldsymbol{\mu}_{\mathtt{n}}$ yields

$$\boldsymbol{z}_t|m \approx \mathbf{g}(\boldsymbol{\mu}_{\mathtt{xe}}^{(m)}, \boldsymbol{\mu}_{\mathtt{l}}, \boldsymbol{\mu}_{\mathtt{n}}) + [\mathbf{J}_{\mathtt{xe}}^{(m)}, \mathbf{J}_{\mathtt{ne}}^{(m)}] \left[\begin{array}{c} \boldsymbol{x}_{\mathtt{e}} - \boldsymbol{\mu}_{\mathtt{xe}}^{(m)} \\ \boldsymbol{n}_t - \boldsymbol{\mu}_{\mathtt{n}} \end{array} \right] \quad (22)$$

where

$$\mathbf{J}_{\mathtt{xe}}^{(m)} = [\mathbf{J}_{\mathtt{x}0}^{(m)}, \cdots, \mathbf{J}_{\mathtt{x}n}^{(m)}]\,; \quad \mathbf{J}_{\mathtt{ne}}^{(m)} = \mathbf{I} - \sum_{\delta=0}^{n} \mathbf{J}_{\mathtt{x}\delta}^{(m)} \quad (23)$$

and

$$\mathbf{J}_{\mathtt{x}\delta}^{(m)} = \frac{\partial \mathbf{g}}{\partial \boldsymbol{x}_{t-\delta}}\Big|_{\boldsymbol{\mu}_{\mathtt{xe}}^{(m)}, \boldsymbol{\mu}_{\mathtt{l}}, \boldsymbol{\mu}_{\mathtt{n}}} \quad (24)$$

Thus the model parameters are compensated by

$$\boldsymbol{\mu}_{\mathtt{sz}}^{(m)} = \mathbf{g}(\boldsymbol{\mu}_{\mathtt{xe}}^{(m)}, \boldsymbol{\mu}_{\mathtt{l}}, \boldsymbol{\mu}_{\mathtt{n}}) \quad (25)$$

$$\boldsymbol{\mu}_{\Delta\mathtt{z}}^{(m)} = \mathbf{J}_{\mathtt{xe}}^{(m)} \boldsymbol{\mu}_{\Delta\mathtt{xe}}^{(m)} \quad (26)$$

The delta-delta parameters are compensated in a similar way as the delta parameters. This compensation form is referred to as RVTS Joint (RVTSJ).

It is possible to compensate the variances as well. However, in this initial investigation, it was found that variance compensation is quite sensitive and a good compensation is hard to obtain. Thus, in this work the variance compensation is done using standard VTS, i.e., $\boldsymbol{\Sigma}_{\mathtt{z}}^{(m)} = \boldsymbol{\Sigma}_{\mathtt{y}}^{(m)}$ (c.f. Eq. 5 - Eq. 8) , where $\boldsymbol{\mu}_{\mathtt{sn}}, \boldsymbol{\mu}_{\mathtt{h}}$ and $\boldsymbol{\Sigma}_{\mathtt{n}}$ are estimated via standard VTS noise estimation.

IV. Noise Estimation

In the previous section, two model compensation forms, RVTS and RVTSJ, are described. The noise parameters need to be determined. Though there exists a simple method to determine the frame-level distortion terms [7] based on the known reverberation time T_{60}, it is preferable to use ML estimate of noise parameters, as it yields consistent fit with the reverberant data. In [10], a sequential ML estimation of noise parameters for RVTS was presented, where the additive and convolutional noise parameters, $\boldsymbol{\mu}_{\mathtt{n}}, \boldsymbol{\mu}_{\mathtt{h}}$ and $\boldsymbol{\Sigma}_{\mathtt{n}}$, were first estimated using standard VTS noise estimation, then the noisy statistics $\boldsymbol{\mu}_{\mathtt{ye}}^{(m)}$ were obtained, followed by reverberant noise mean $\tilde{\boldsymbol{\mu}}_{\mathtt{l}}$ estimation. Though it is possible to jointly estimate $\tilde{\boldsymbol{\mu}}_{\mathtt{l}}, \boldsymbol{\mu}_{\mathtt{n}}$ and $\boldsymbol{\mu}_{\mathtt{h}}$ for RVTS, additional assumptions (e.g., $\mathbf{J}_{\mathtt{n}}^{(m)}$ is invariant of $\boldsymbol{\mu}_{\mathtt{h}}, \boldsymbol{\mu}_{\mathtt{n}}$) are needed. In this work, joint estimation of the reverberant and additive noise mean is presented for the RVTSJ compensation.

The estimation of reverberant and additive noise mean is done in the EM framework, similar to the convolutional and additive noise mean estimation using EM. The following auxiliary function is maximised:

$$\mathcal{Q}(\hat{\boldsymbol{\mu}}_{\mathtt{l}}, \hat{\boldsymbol{\mu}}_{\mathtt{n}}) = \sum_{t,m} \gamma_t^{(m)} \log p(\boldsymbol{z}_t; \boldsymbol{\mu}_{\mathtt{z}}^{(m)}, \boldsymbol{\Sigma}_{\mathtt{z}}^{(m)}) + \mathcal{R}(\hat{\boldsymbol{\mu}}_{\mathtt{l}}, \hat{\boldsymbol{\mu}}_{\mathtt{n}}) \quad (27)$$

where $\gamma_t^{(m)}$ is the posterior of component m at time t, given the current hypothesis and current noise estimates $\boldsymbol{\mu}_{\mathtt{l}}, \boldsymbol{\mu}_{\mathtt{n}}$, $\mathcal{R}(\hat{\boldsymbol{\mu}}_{\mathtt{l}}, \hat{\boldsymbol{\mu}}_{\mathtt{n}})$ is a regularisation term to improve the stability of noise estimation. In this work, the following form of regularisation was used:

$$\mathcal{R}(\hat{\boldsymbol{\mu}}_{\mathtt{l}}, \hat{\boldsymbol{\mu}}_{\mathtt{n}}) = \alpha \left((\hat{\boldsymbol{\mu}}_{\mathtt{l}} - \boldsymbol{\mu}_{\mathtt{l}})^{\mathsf{T}} (\hat{\boldsymbol{\mu}}_{\mathtt{l}} - \boldsymbol{\mu}_{\mathtt{l}}) + (\hat{\boldsymbol{\mu}}_{\mathtt{n}} - \boldsymbol{\mu}_{\mathtt{n}})^{\mathsf{T}} (\hat{\boldsymbol{\mu}}_{\mathtt{n}} - \boldsymbol{\mu}_{\mathtt{n}}) \right)$$

Performing a first-order expansion of $\boldsymbol{\mu}_{\mathtt{z}}^{(m)}$ using the current estimates, $\boldsymbol{\mu}_{\mathtt{l}}, \boldsymbol{\mu}_{\mathtt{n}}$, yields:

$$\hat{\boldsymbol{\mu}}_{\mathtt{sz}}^{(m)} \approx \boldsymbol{\mu}_{\mathtt{sz}}^{(m)} + \left[\mathbf{J}_{\mathtt{le}}^{(m)} \mathbf{J}_{\mathtt{ne}}^{(m)} \right] \left[\begin{array}{c} \hat{\boldsymbol{\mu}}_{\mathtt{l}} - \boldsymbol{\mu}_{\mathtt{l}} \\ \hat{\boldsymbol{\mu}}_{\mathtt{n}} - \boldsymbol{\mu}_{\mathtt{n}} \end{array} \right] \quad (28)$$

where

$$\mathbf{J}_{\mathtt{le}}^{(m)} = [\mathbf{J}_{\mathtt{l}0}^{(m)}, \cdots, \mathbf{J}_{\mathtt{l}n}^{(m)}] \quad \mathbf{J}_{\mathtt{l}\delta}^{(m)} = \frac{\partial \mathbf{g}}{\partial \boldsymbol{\mu}_{\mathtt{l}\delta}}\Big|_{\boldsymbol{\mu}_{\mathtt{xe}}^{(m)}, \boldsymbol{\mu}_{\mathtt{l}}, \boldsymbol{\mu}_{\mathtt{n}}}$$

Differentiating the auxiliary function and equating to zero gives the following update:

$$\left[\begin{array}{c} \hat{\boldsymbol{\mu}}_{\mathtt{l}} \\ \hat{\boldsymbol{\mu}}_{\mathtt{n}} \end{array} \right] = \left(\sum_{t,m} \gamma_t^{(m)} \mathbf{J}^{(m)\mathsf{T}} \boldsymbol{\Sigma}_{\mathtt{sz}}^{(m)-1} \mathbf{J}^{(m)} + \alpha \mathbf{I} \right)^{-1} \times \quad (29)$$

$$\left(\sum_{t,m} \gamma_t^{(m)} \mathbf{J}^{(m)\mathsf{T}} \boldsymbol{\Sigma}_{\mathtt{sz}}^{(m)-1} \left(\boldsymbol{\mu}_{\mathtt{sz}}^{(m)} - \mathbf{J}^{(m)} \left[\begin{array}{c} \boldsymbol{\mu}_{\mathtt{l}} \\ \boldsymbol{\mu}_{\mathtt{n}} \end{array} \right] \right) \right)$$

where $\mathbf{J}^{(m)} = [\mathbf{J}_{\mathtt{le}}^{(m)} \mathbf{J}_{\mathtt{ne}}^{(m)}]$. Note in noise estimation for RVTS, similar expression were used, except that only $\tilde{\boldsymbol{\mu}}_{\mathtt{l}}$ were updated while $\boldsymbol{\mu}_{\mathtt{n}}$ were fixed at the value estimated in VTS.

The above updating formula only consider the static parameters in the auxiliary function. To yield best performance, all the compensated parameters should be included in the auxiliary function. The updating formula is slightly modified to reflect that compensated delta and delta-delta parameters are also functions of $\hat{\boldsymbol{\mu}}_{\mathtt{l}}$ and $\hat{\boldsymbol{\mu}}_{\mathtt{n}}$. Moreover, due to the linear approximation in Eq. (28), the auxiliary function needs to be checked after every updating iteration to ensure the auxiliary is non-decreasing. More details of the noise estimation can be found in [14].

Since the auxiliary function is highly nonlinear, it is crucial to have a good initialisation. The initialisation scheme in this work uses an initial (rough) estimate of T_{60} value, similar to the one used in [7]. The initialisation scheme is slightly modified so that the initial compensated mean vectors are approximately the same as the VTS compensated means. For RVTS, the initial frame-level distortion terms are given by

$$\tilde{\boldsymbol{\mu}}_{\mathtt{l}\delta} = \mathbf{C}[\, \delta\eta + \beta \;\; \ldots \;\; \delta\eta + \beta]^{\mathsf{T}} \quad (30)$$

where

$$\eta = -3\log(10)\frac{\Delta}{T_{60}}\,; \quad \beta = -\log\left(\frac{1 - e^{(n+1)\eta}}{1 - e^{\eta}}\right) \quad (31)$$

and Δ is the shift of analysis window (10ms in this work). Note that here the cepstral coefficients are extracted from the magnitude spectrum rather than power spectrum, therefore Eq. (31) is slightly different from the one in [7]. For RVTSJ, $\boldsymbol{\mu}_{\mathtt{n}}$ is initialised as the additive noise mean estimated by standard VTS noise estimation, while for $\boldsymbol{\mu}_{\mathtt{l}\delta}$

$$\boldsymbol{\mu}_{\mathtt{l}\delta} = \boldsymbol{\mu}_{\mathtt{h}} + \mathbf{C}[\, \delta\eta + \beta \;\; \ldots \;\; \delta\eta + \beta]^{\mathsf{T}} \quad (32)$$

and $\boldsymbol{\mu}_{\mathrm{h}}$ is the convolutional noise estimated in VTS noise estimation.

V. Experiments and Results

The above RVTS and RVTSJ were evaluated and contrasted on a reverberant version of the AURORA4 task [15]. The original AURORA4 task is derived from Wall Street Journal (WSJ0) 5k-word closed vocabulary task, with 330 utterances from 8 speakers in the test set. Test set A (`test01`) was recorded with a close-talking microphone; set B(`test02-07`) had 6 different types of noise added, with randomly selected SNRs, ranging from 15dB to 5dB; Set C (`test08`) was recorded with secondary microphones; noise was also added to set C to form set D(`test09-14`). Three of these 14 sets, `01,04` and `08` were selected and passed through a simulation tool [16] to simulate the effect of reverberant noise, with the reverberation time T_{60} set to 400ms. These additive and reverberant noise corrupted sets form the reverberant AURORA4 task[1].

The HTK frontend was used to derive a 39-dimensional feature vector, consisting of 12 MFCCs, extracted from magnitude spectrum, appended with zeroth cepstrum, delta and delta-delta coefficients. Cross-word triphone models with 3140 distinct states and 16 component per state were trained on the "clean" data (7138 utterances/83 speakers). For the extended model statistics, the feature vector was appended with high-order DCT elements of an appropriate window width. $n = 10, w = 4$ were used as the length of history frames and the window length used for calculating the dynamic parameters, respectively. The standard bi-gram LM for the AURORA4 task was used in decoding.

Adaptation in the experiments were performed in an unsupervised mode. All noise parameters were estimated at the utterance level. Initially, acoustic models were compensated based on an initial estimation of the additive noise, using the first and last 20 frames of each utterance. These compensated models were used to generate an initial hypothesis. With this initial hypothesis, the noise models were re-estimated. New hypotheses were then generated. This process was optionally repeated several times. Table I shows the how the VTS systems performed on the *original* AURORA4 task. Compared with the unadapted system, VTS adaptation greatly reduces the error rates on this additive and convolutional noise corrupted corpus.

Est.	set A	set B	set C	set D	Avg.
–	7.1	55.9	47.1	71.7	58.5
Init	8.0	24.0	44.4	49.7	35.3
ML	6.9	15.1	11.8	23.3	17.8

TABLE I
Estimation of additive and convolutional noise on the original AURORA4 task.

The same VTS compensation scheme was also run on the reverberant AURORA4 task. Results are shown in the first line of Table II. Due to the reverberation effect, the performance were seriously degraded: the average WERs were 49.5% while the WERs on the original three sets were only 12.7% (`01`, 6.9%, `04` 19.5%, `08` 11.8%). This demonstrates the challenge of this task.

Schemes	Est.	test			Avg.
		01	04	08	
VTS	ML	43.8	48.9	55.7	49.5
RVTS	Init	29.7	48.7	43.4	40.6
	ML	26.7	43.9	40.4	37.0
RVTSJ	Init	29.4	46.9	41.9	39.4
	ML	24.1	40.4	35.2	33.2

TABLE II
WER% of RVTS and RVTSJ using initial and ML estimated noise.

RVTS and RVTSJ model compensation experiments were run using both initial and ML estimates of noise parameters. For initialisation, the T_{60} value was set as 400ms, matched with the simulator's setting. Results were shown in the second and fourth rows of Table II. It is observed that the model compensation using initial noise parameters already yielded large gains, especially on `01` and `08`, with RVTSJ slightly better. As a comparison, reference [7] reports a WER of 39.8% on the same `01` set, using the compensation scheme therein.

As demonstrated in Table I, using ML estimated noise yields large gains over simple initial estimates of noise. Therefore, it is also preferable to use ML estimated noise for RVTS and RVTSJ compensation. The VTS hypothesis was taken as the initial supervision, noise parameters were re-estimated while the model variance was locked as the VTS compensated variance. 4 EM iterations were used. The supervision hypothesis was also updated (1 in the experiments) to yield better noise estimation before final decoding. Results are shown in the third and fifth rows of Table II. As expected, ML estimation of noise yields consistent gains over initial noise estimation. RVTSJ outperforms RVTS in all three sets. This is due to the sequential approach to noise estimation in RVTS, where the additive noise was estimated using the VTS-style mismatch function, then the frame-level distortion terms were estimated given the additive noise. Because of this sequential approach, the additive noise was used to model some attributes of reverberation, yielding inaccurate noise estimates. Joint estimation of both additive and reverberant noise alleviates this issue by taking the effect of both reverberant and additive noise into account.

Experiments in Table II assumed the reverberation time, T_{60}, was known. In practice, it is only possible to know the reverberation time to some extent. Another set of experiments were run with different T_{60} values ranging from 200ms to 800ms. Though using initial estimated noise based on mismatched T_{60} value do have an impact on performance, the ML estimate was relatively insensitive to the initialisation. For example, using the initial noise estimate, RVTSJ performance varied from 27.5% to 31.7%, while the performance of ML estimated noise only varied from 24.3% to 25.0%. This again

[1]To keep the experiments repeatable, the reverberant noise is added after background noise distortion, which matches the assumption of RVTS. Experiments on the real data are on-going and will be reported elsewhere.

Schemes	domain	test			Avg.
		01	04	08	
VTS	power	46.7	61.1	61.7	56.5
	magnitude	43.8	48.9	55.7	49.5
RVTSJ	power	30.5	53.1	46.9	43.5
	magnitude	24.1	40.4	35.2	33.2

TABLE III
VTS AND RVTSJ ADAPTATION USING MISMATCH FUNCTIONS IN POWER AND MAGNITUDE DOMAIN.

Schemes		test			Avg.
		01	04	08	
VTS	ML	43.8	48.9	55.7	49.5
	+CMLLR	32.0	45.2	44.6	40.6
RVTS	ML	27.3	44.0	40.8	37.4
	+CMLLR	22.8	41.0	32.8	32.2
RVTSJ	ML	24.1	40.4	35.2	33.2
	+CMLLR	20.2	36.5	29.0	28.6

TABLE IV
VTS, RVTS AND RVTSJ AND THEIR COMBINATION WITH CMLLR ON REVERBERANT AURORA4 TASK.

demonstrates the advantage of the ML noise estimation.

The above experiments assume the noise and speech are additive in the magnitude domain, as it was empirically found magnitude domain combination yielded better results [17] for additive noise corrupted data. It is also interesting to examine this conclusion for reverberant noise corrupted data. VTS and RVTSJ adaptation experiments were re-run using the power domain mismatch functions ($\gamma = 2$) with the same setup. Results are shown in Table III. Consistent with the finding in [17], magnitude domain combination performs better.

To further improve the performance, a linear transform, CMLLR transform [18], was combined with previous model compensation schemes. A global CMLLR transform was estimated for each speaker. Results are shown in Table IV. As expected, adding linear transforms to further reduce the mismatch yielded large gains. The best performance was achieved by RVTSJ combined with CMLLR adaptation, which was a 42.2% relative error reduction, compared with VTS adaptation alone. The combination of RVTSJ and CMLLR transform also outperforms the Direct CMLLR approach proposed in [10], in which a linear transform was employed to project several neighbouring frames. This demonstrates that the use of nonlinear mismatch functions is helpful for the reverberant noise distortion.

VI. CONCLUSION

This paper investigates Reverberant VTS model compensation for hands-free speech recognition. In [10], the VTS model compensation was extended to handle reverberant noise, where it was assumed that the observation vector is generated by the reverberation of a sequence of additive noise corrupted noisy speech vectors. An alternative form of RVTS model compensation, RVTSJ, was examined, where another form of mismatch is explored, in which the corrupted observation is generated by the combination of additive and the reverberant of previous clean speech vectors. This form of mismatch function allows an easy formulation of estimating background and reverberant noise jointly. These two model compensation schemes were evaluated on the Reverberant AURORA4 task. Both RVTS and RVTSJ yielded large gains over VTS baseline system, with RVTSJ being consistently better.

ACKNOWLEDGEMENT

The authors would like to thank Dr. F. Flego for making VTS code available. This work was partially supported by Google research award and DARPA under the GALE and RATS program. The paper does not necessarily reflect the position or the policy of US Government and no official endorsement should be inferred.

REFERENCES

[1] M. Omologo, M. Matassoni, P. Svaizer, and P. Giuliani, "Microphone array based speech recognition with different talker-array positions," in *Proc. of ICASSP*, 1997, pp. 227–230.

[2] M. Miyoshi and Y. Kaneda, "Inverse filtering of room acoustics," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 36, no. 2, pp. 145 –152, Feb 1988.

[3] 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.

[4] A. Krueger and R. Haeb-Umbach, "Model based feature enhancement for automatic speech recognition in reverberant environments," *IEEE Transactions on Speech, Audio and Language Processing*, vol. 18, no. 7, pp. 1692–1707, 2010.

[5] M. J. F. Gales, "Model-based techniques for noise robust speech recognition," Ph.D. dissertation, Cambridge University, 1995.

[6] A. Acero, L. Deng, T. Kristjansson, and J. Zhang, "HMM adaptation using vector Taylor series for noisy speech recognition," in *Proc. ICSLP-2000*.

[7] 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.

[8] C. K. Raut, T. Nishimoto, and S. Sagayama, "Model adaptation for long convolutional distortion by maximum likelihood state filtering approach," in *Proc. ICASSP-2006*, pp. 277–280.

[9] A. Sehr, R. Maas, and W. Kellermann, " Frame-wise HMM adaptation using state-dependent reverberation estimates," in *Proc. ICASSP-2011*.

[10] M. J. F. Gales and Y.-Q. Wang, "Model-based approaches to handling additive noise in reverberant environments," in *Workshop on Hands-free Speech Communication and Microphone Arrays*, 2011.

[11] R. A. Gopinath *et al.*, "Robust speech recognition in noise – performance of the IBM continuous speech recogniser on the ARPA noise spoke task," in *Proc. APRA Workshop on Spoken Language System Technology*, 1995, pp. 127–130.

[12] H. Liao and M. J. F. Gales, "Joint uncertainty decoding for robust large vocabulary speech recognition," University of Cambridge, Tech. Rep. CUED/F-INFENG/TR552, 2006.

[13] J. Li, D. Yu, Y. Gong, and A. Acero, "High-performance HMM adaptation with joint compensation of additive and convolutive distortions via vector Taylor series," in *Proc. ASRU-2007*.

[14] Y.-Q. Wang and M. J. F. Gales, "Model based approaches to handling additive and reverberant noise," University of Cambridge, Tech. Rep. CUED/F-INFENG/TR666, 2011.

[15] N. Parihar and J. Picone, "Aurora working group: DSR front end LVCSR evaluation AU/384/02," Inst. for Signal and Information Process, Mississippi State University, Tech. Rep.

[16] H. G. Hirsch and H. Finster, "The simulation of realistic acoustic input scenarios for speech recognition systems," in *Proc. ICSLP*, 2005.

[17] M. J. F. Gales and F. Flego, "Discriminative classifiers with adaptive kernels for noise robust speech recognition," *Computer speech and language*, 2009.

[18] M. J. F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer speech and language*, vol. 12, pp. 75–98, 1998.