A Trajectory-based Parallel Model Combination with a Unified Static and Dynamic Parameter Compensation For Noisy Speech Recognition

Khe Chai SIM #1, Minh-Thang LUONG #2

Department of Computer Science, School of Computing, National University of Singapore
13 Computing Drive, Singapore 117417
[1] simkc@comp.nus.edu.sg
[2] luongmin@comp.nus.edu.sg

***Abstract*—Parallel Model Combination (PMC) is widely used as a technique to compensate Gaussian parameters of a clean speech model for noisy speech recognition. The basic principle of PMC uses a log normal approximation to transform statistics of the data distribution between the cepstral domain and the linear spectral domain. Typically, further approximations are needed to compensate the dynamic parameters separately. In this paper, Trajectory PMC (TPMC) is proposed to compensate both the static and dynamic parameters. TPMC uses the explicit relationships between the static and dynamic features to transform the static and dynamic parameters into a sequence (trajectory) of static parameters, so that the log normal approximation can be applied. Experimental results on WSJCAM0 database corrupted with additive babble noise reveals that the proposed TPMC method gives promising improvements over PMC and VTS.**

I. Introduction

Hidden Markov Model (HMM) [1] is widely used as a statistical model of the acoustic patterns for speech recognition. Typically, Gaussian Mixture Models (GMMs) are used to represent the distribution of the acoustic features for each HMM state. Mel Frequency Cepstral Coefficient (MFCC) [2] is commonly used as acoustic features, with which acoustic models are trained. Therefore, the mean vector and the covariance matrix of each Gaussian component correspond to the statistics of the distribution of the acoustic features in the *cepstral* domain.

In order to achieve good performance with statistical models, such as the HMMs, it is important to have a matched condition between the acoustic data used to estimate the model parameters and the acoustic data to be observed during recognition. Any mismatch in acoustic conditions will lead to performance degradation, the severity of which depends on the degree of mismatch. One of the most common sources of acoustic mismatch is the presence of environmental noise. In practice, the type of noise present during recognition is not known *a priori*. Therefore, it is not practical to train an acoustic model for each noise condition. Existing methods for improving the performance of noisy speech recognition include Parallel Model Combination (PMC) [3] and Vector Taylor Series (VTS) [4].

Standard PMC technique uses *log normal approximation* to transform the statistics between the cepstral and linear spectral domains, such that the statistics of the clean speech model and the noise model can be combined easily in the linear spectral domain. However, log normal approximation cannot be applied directly to dynamic parameters without further approximations (*e.g.* continuous-time approximation [3] and data-driven approximation [5]). This paper proposes an extension to standard PMC which offers a unified compensation scheme for both the static and dynamic parameters. The proposed method is referred to as Trajectory PMC (TPMC). TPMC uses the explicit relationships between the static and dynamic features to transform the distribution in the observed (static and dynamic) space into an equivalent distribution in the cepstral trajectory domain. Since the statistics in the cepstral trajectory domain involves only the static parameters (including the temporal correlations), log normal approximation can be applied directly to compensate both the static and dynamic parameters in a unified manner.

The remaining of this paper is organised as follows. Section II introduces the related work, including PMC and VTS. Section III describes the trajectory HMM formulation. Section IV presents the formulation for the proposed TPMC method. Section V discusses the properties of TPMC. Finally, experimental results are given in Section VI.

II. Related Work

Model compensation techniques are widely used to adapt acoustic models trained on clean data to a new acoustic environment. Two state-of-the-art model-based noise compensation methods will be described in the following, using MFCC as the acoustic features.

A. Parallel Model Combination (PMC)

Parallel Model Combination (PMC) [3] uses the *log normal approximation* to transform the statistics of the speech data between the *cepstral* and *linear spectral* domains such that the statistics of the clean model and the noise model can be easily combined to yield the noisy speech model. Let $\boldsymbol{C}$ and $\boldsymbol{C}^{-1}$ be the DCT and inverse DCT matrices respectively. The

conversion formula from *cepstral* to *linear spectral* domains for the mean and covariance statistics are given by [6]:

$$\boldsymbol{\mu} = \exp\left(\boldsymbol{C}^{-1}\boldsymbol{\mu}^{(c)} + \frac{1}{2}\,\mathrm{diag}^{-1}\left(\boldsymbol{C}^{-1}\boldsymbol{\Sigma}^{(c)}\boldsymbol{C}^{-\top}\right)\right) \quad (1)$$

$$\boldsymbol{\Sigma} = \boldsymbol{M}\left(\exp\left(\boldsymbol{C}^{-1}\boldsymbol{\Sigma}^{(c)}\boldsymbol{C}^{-\top}\right) - 1\right)\boldsymbol{M} \quad (2)$$

where $\mathrm{diag}^{-1}(\cdot)$ denotes the operation of extracting the diagonal elements of a matrix as a column vector and $\boldsymbol{M}$ is a diagonal matrix such that $\boldsymbol{\mu} = \mathrm{diag}^{-1}(\boldsymbol{M})$. The corresponding conversion formula from *linear spectral* to *cepstral* domains are given by:

$$\boldsymbol{\mu}^{(c)} = \boldsymbol{C}\left(\log(\boldsymbol{\mu}) - \frac{1}{2}\log\left(\mathrm{diag}^{-1}(\boldsymbol{V}) + 1\right)\right) \quad (3)$$

$$\boldsymbol{\Sigma}^{(c)} = \boldsymbol{C}\left(\log\left(\boldsymbol{V} + 1\right)\right)\boldsymbol{C}^{\top} \quad (4)$$

where $\boldsymbol{V} = \boldsymbol{M}^{-1}\boldsymbol{\Sigma}\boldsymbol{M}^{-1}$. Based on the assumption that speech and noise are independent and addictive in the linear spectral domain, the corrupted-speech parameters in the same domain are:

$$\hat{\boldsymbol{\mu}} = g \cdot \boldsymbol{\mu} + \tilde{\boldsymbol{\mu}} \qquad \text{and} \qquad \hat{\boldsymbol{\Sigma}} = g^2 \cdot \boldsymbol{\Sigma} + \tilde{\boldsymbol{\Sigma}}$$

where g is a gain matching term introduced to account for level differences between the clean and the noisy speeches[1]. Unfortunately, the above log normal approximation cannot be directly applied to dynamic parameters without further approximations such as continuous-time approximation [3] and Data-driven PMC (DPMC) [5].

B. Vector Taylor Series (VTS)

Vector Taylor Series (VTS) [4] is another model-based noise compensation technique widely used for noisy speech recognition. This method uses Taylor series expansion to approximate the nonlinear function describing the cepstral features of the noisy data, $\hat{\boldsymbol{c}}$, given the cepstral features of the clean data, $\boldsymbol{c}$, and the noise data, $\tilde{\boldsymbol{c}}$:

$$\hat{\boldsymbol{c}} = \boldsymbol{c} + \boldsymbol{C}\log\left(g + e^{\left(\boldsymbol{C}^{-1}(\tilde{\boldsymbol{c}} - \boldsymbol{c})\right)}\right) \quad (5)$$

The VTS formulae for compensating both the static and dynamic parameters are given in [4], using the first-order approximation.

III. Trajectory HMM Formulation

Trajectory HMM reformulates the standard HMM by imposing the explicit relationships between the static and dynamic parameters [7]. Trajectory HMM has been widely used to generate speech parameters for HMM-based speech synthesis [8]. In the standard HMM formulation, the likelihood of the HMM model, with parameters Λ, observing an observation sequence, $\boldsymbol{o} = [\boldsymbol{o}_1^{\top}, \ldots, \boldsymbol{o}_T^{\top}]^{\top}$ given the state sequence, $q = \{q_1, q_2, \ldots, q_T\}$, is given by:

$$P(\boldsymbol{o}|q, \Lambda) = \prod_{t=1}^{T} \mathcal{N}(\boldsymbol{o}_t|\boldsymbol{\mu}_{q_t}, \boldsymbol{\Sigma}_{q_t}) = \mathcal{N}(\boldsymbol{o}|\boldsymbol{\mu}_q^{(o)}, \boldsymbol{\Sigma}_q^{(o)}) \quad (6)$$

[1]Parameters of the noise model and the noise-corrupted speech model are capped with $\tilde{}$ and $\hat{}$ respectively

where $\boldsymbol{\mu}_{q_t}$ and $\boldsymbol{\Sigma}_{q_t}$ are the $3M \times 1$ mean vector and the $3M \times 3M$ covariance matrix associated with state q_t and

$$\boldsymbol{\mu}_q^{(o)} = [\boldsymbol{\mu}_{q_1}^{\top}, \ldots, \boldsymbol{\mu}_{q_T}^{\top}]^{\top}, \quad \boldsymbol{\Sigma}_q^{(o)} = \mathrm{diag}[\boldsymbol{\Sigma}_{q_1}, \ldots, \boldsymbol{\Sigma}_{q_T}]$$

$(\boldsymbol{\mu}_q^{(o)}, \boldsymbol{\Sigma}_q^{(o)})$ represents the statistics in the observation trajectory domain[2]. However, the observation parameters are related to the static cepstral coefficients as $\boldsymbol{o}_t = [\boldsymbol{c}_t^{\top}, \Delta\boldsymbol{c}_t^{\top}, \Delta^2\boldsymbol{c}_t^{\top}]^{\top}$. The explicit relationships between the static and dynamic parameters can be conveniently expressed in the following matrix notation:

$$\boldsymbol{o} = \boldsymbol{W}\boldsymbol{c} \quad (7)$$

where $\boldsymbol{c} = [\boldsymbol{c}_1^{\top}, \ldots, \boldsymbol{c}_N^{\top}]^{\top}$ is the sequence of static coefficients and $\boldsymbol{W}$ is a window matrix of size $3MT \times MN$. Here, T is the number of vectors in the *observation trajectory* space; while N is the number of vectors in the *cepstral trajectory* space[3]. We should note that in [7], N is constrained to be equal to T; whereas, in our approach, they could be flexibly chosen to satisfy certain properties, which we will detail in Section V.

By imposing the constraint in Eq. (7), Eq. (6) could be rewritten as a function of c:

$$P(\boldsymbol{W}c|q, \Lambda) = \mathcal{N}(\boldsymbol{o}|\boldsymbol{\mu}_q^{(o)}, \boldsymbol{\Sigma}_q^{(o)}) = \mathcal{K}_q \cdot \mathcal{N}(\boldsymbol{c}|\boldsymbol{\mu}_q^{(c)}, \boldsymbol{\Sigma}_q^{(c)})$$

$\mathcal{K}_q$ is a normalisation constant independent of c; whereas $\boldsymbol{\mu}_q^{(c)}$ and $\boldsymbol{\Sigma}_q^{(c)}$ are a mean vector, and a covariance matrix in the cepstral trajectory domain:

$$\boldsymbol{\mu}_q^{(c)} = \boldsymbol{\Sigma}_q^{(c)}\boldsymbol{W}^{\top}{\boldsymbol{\Sigma}_q^{(o)}}^{-1}\boldsymbol{\mu}_q^{(o)} \quad (8)$$

$$\boldsymbol{\Sigma}_q^{(c)} = \left(\boldsymbol{W}^{\top}{\boldsymbol{\Sigma}_q^{(o)}}^{-1}\boldsymbol{W}\right)^{-1} \quad (9)$$

Eq. (8) and Eq. (9) form the basis for the conversion of the statistics from the *observation trajectory* domain to the *cepstral trajectory* domain. This transform plays a crucial part in the formulation of the proposed Trajectory PMC method, which will be described in detail in the following section.

IV. Trajectory PMC

Trajectory PMC (TPMC) is proposed as an extension to PMC so that both the static and dynamic parameters can be compensated using the log normal approximation in a unified manner. TPMC eliminates the need to deal with dynamic parameters explicitly by transforming the observation statistics into the cepstral trajectory statistics. As such, the dynamic feature information is implicitly encoded within the cepstral trajectory space. The overall TPMC compensation algorithm is depicted in Fig. 1. There are three major steps involved: 1) transformation of statistics from the observation space to the cepstral trajectory space (*forward trajectory*); 2) combination of clean and noise statistics in the cepstral trajectory domain; and 3) transformation of statistics from the cepstral trajectory space to the observation space (*backward trajectory*). These steps will be described in the following sections.

[2]Subscript q is used to indicate a *trajectory* domain, which represents a concatenation of a sequence of vectors.

[3]We will use the superscript $^{(o)}$ to represent features in the *observation* space as opposed to those in the *trajectory* space indicated by $^{(c)}$.

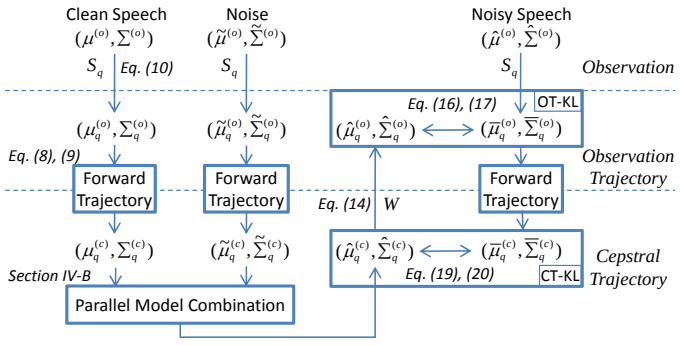

Fig. 1. A schematic diagram illustrating the Trajectory PMC process.

A. Forward Trajectory

The forward trajectory step transforms the observation statistics, $(\boldsymbol{\mu}^{(o)}, \boldsymbol{\Sigma}^{(o)})$, into the cepstral trajectory statistics, $(\boldsymbol{\mu}_q^{(c)}, \boldsymbol{\Sigma}_q^{(c)})$. First, the statistics in the observation trajectory space, $(\boldsymbol{\mu}_q^{(o)}, \boldsymbol{\Sigma}_q^{(o)})$, are needed. Since TPMC is applied per Gaussian components, these statistics are obtained by assuming a *constant* statistics within each component, leading to duplicating the observation statistics T times:

$$\boldsymbol{\mu}_q^{(o)} = \boldsymbol{S}_q \boldsymbol{\mu}^{(o)}, \quad \boldsymbol{\Sigma}_q^{(o)} = \left[\text{diag}\left(\boldsymbol{S}_q \boldsymbol{\phi}^{(o)}\right)\right]^{-1} \tag{10}$$

where $\boldsymbol{\phi}^{(o)}$ is the diagonal vector of the matrix ${\boldsymbol{\Sigma}^{(o)}}^{-1}$ and

$$\boldsymbol{S}_q = [\underbrace{1 \ldots 1}_{T \text{ times}}]^\top \otimes \boldsymbol{I}_{3M} \tag{11}$$

where $\otimes$ denotes the *Kronecker* product operator.

Next, Eq. (8) and (9) can be applied to obtain the required statistics in the cepstral trajectory domain, $(\boldsymbol{\mu}_q^{(c)}, \boldsymbol{\Sigma}_q^{(c)})$:

$$\boldsymbol{\mu}_q^{(c)} = \boldsymbol{\Sigma}_q^{(c)} \boldsymbol{W}^\top \text{diag}\left(\boldsymbol{S}_q \boldsymbol{\phi}^{(o)}\right) \boldsymbol{S}_q \boldsymbol{\mu}^{(o)} \tag{12}$$

$$\boldsymbol{\Sigma}_q^{(c)} = \left(\boldsymbol{W}^\top \text{diag}\left(\boldsymbol{S}_q \boldsymbol{\phi}^{(o)}\right) \boldsymbol{W}\right)^{-1} \tag{13}$$

Note that the process of generating the cepstral trajectory mean, $\boldsymbol{\mu}_q^{(c)}$, from the observation parameters is the same as synthesising a sequence of cepstral parameters for speech synthesis [8]. Also, synthesising data using trajectory HMM formulation for noise speech recognition has been applied to Support Vector Machines for noisy robust speech recognition [9]. However, TPMC uses the trajectory HMM formulation for statistic transformation in a rather different way.

B. Parallel Model Combination in Trajectory Domain

Applying the forward trajectory step over the two modalities – speech and noise – yields two sets of statistics in the cepstral trajectory domain: $(\boldsymbol{\mu}_q^{(c)}, \boldsymbol{\Sigma}_q^{(c)})$ and $(\tilde{\boldsymbol{\mu}}_q^{(c)}, \tilde{\boldsymbol{\Sigma}}_q^{(c)})$. Since these statistics correspond to only the static cepstral features, the *log normal* approximation approach employed in the standard PMC method, as described in Section II-A, can be applied to combine these cepstral trajectory statistics. Eq. (2) through (4) still hold for trajectory-based PMC, except that all the statistics in those equations correspond to the trajectory space (*i.e.* with subscript q) and the transformation matrices between the cepstral and log-spectral domains, $\boldsymbol{C}$ and $\boldsymbol{C}^{-1}$, are replaced by $\boldsymbol{Q}$ and $\boldsymbol{Q}^{-1}$, respectively to account for the trajectory expansion. The trajectory version of the transformation matrices are given by:

$$\boldsymbol{Q} = \boldsymbol{I}_N \otimes \boldsymbol{C} \quad \text{and} \quad \boldsymbol{Q}^{-1} = \boldsymbol{I}_N \otimes \boldsymbol{C}^{-1}$$

C. Backward Trajectory

After going through the forward trajectory and PMC processes, we obtain the corrupted-speech model $(\hat{\boldsymbol{\mu}}_q^{(c)}, \hat{\boldsymbol{\Sigma}}_q^{(c)})$ in the cepstral trajectory space. However, statistics in the observation domain, $(\hat{\boldsymbol{\mu}}^{(o)}, \hat{\boldsymbol{\Sigma}}^{(o)})$, are needed in order to construct the noisy speech model. This imposes a *constant statistics* constraint in the observation trajectory domain, which is an intermediate space between the cepstral trajectory and observation domains (see Fig. 1). Should there be no such constraint, given the linear relationship of Eq. (7), $(\hat{\boldsymbol{\mu}}_q^{(o)}, \hat{\boldsymbol{\Sigma}}_q^{(o)})$ can be obtained easily as:

$$\hat{\boldsymbol{\mu}}_q^{(o)} = \boldsymbol{W} \hat{\boldsymbol{\mu}}_q^{(c)}, \quad \hat{\boldsymbol{\Sigma}}_q^{(o)} = \boldsymbol{W} \hat{\boldsymbol{\Sigma}}_q^{(c)} \boldsymbol{W}^\top \tag{14}$$

However, due to the constant statistics constraint in the observation space (*c.f.* Eq. (10)), it may not be possible to obtain an estimate of $(\hat{\boldsymbol{\mu}}^{(o)}, \hat{\boldsymbol{\Sigma}}^{(o)})$ that exactly represent $(\hat{\boldsymbol{\mu}}_q^{(c)}, \hat{\boldsymbol{\Sigma}}_q^{(c)})$ or $(\hat{\boldsymbol{\mu}}_q^{(o)}, \hat{\boldsymbol{\Sigma}}_q^{(o)})$. Therefore, $(\hat{\boldsymbol{\mu}}^{(o)}, \hat{\boldsymbol{\Sigma}}^{(o)})$ are estimated such that applying the forward trajectory to them yields the closest approximation to the distribution in the trajectory domains. In this work, the Kullback-Leibler (KL) divergence is employed to measure the distance between two distributions. The KL divergence between the "target" distribution $(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)$ and the "estimated" one $(\boldsymbol{\mu}_e, \boldsymbol{\Sigma}_e)$ is given by:

$$D_{\text{KL}}(t,e) = \frac{1}{2}\left[\text{Tr}\left(\boldsymbol{\Sigma}_e^{-1}\boldsymbol{\Sigma}_t\right) + \log\frac{|\boldsymbol{\Sigma}_e|}{|\boldsymbol{\Sigma}_t|} - N + \right. \\ \left. + (\boldsymbol{\mu}_e - \boldsymbol{\mu}_t)^\top \boldsymbol{\Sigma}_e^{-1}(\boldsymbol{\mu}_e - \boldsymbol{\mu}_t)\right] \tag{15}$$

Depending on the domain in which the optimisation takes place, we have the corresponding methods detailed below.

1) Observation Trajectory Space KL (OT-KL): The *target* statistics, $(\hat{\boldsymbol{\mu}}_q^{(o)}, \hat{\boldsymbol{\Sigma}}_q^{(o)})$, can be obtained by applying Eq. (14) to $(\hat{\boldsymbol{\mu}}_q^{(c)}, \hat{\boldsymbol{\Sigma}}_q^{(c)})$. The *constrained* statistics to be estimated in the observation trajectory space, $(\bar{\boldsymbol{\mu}}_q^{(o)}, \bar{\boldsymbol{\Sigma}}_q^{(o)})$, can be obtained using Eq. (10):

$$\bar{\boldsymbol{\mu}}_q^{(o)} = \boldsymbol{S}_q \hat{\boldsymbol{\mu}}^{(o)}, \quad \bar{\boldsymbol{\Sigma}}_q^{(o)} = \left(\text{diag}\left(\boldsymbol{S}_q \hat{\boldsymbol{\phi}}^{(o)}\right)\right)^{-1}$$

Hence, substituting Eq. (10), and Eq. (14) into Eq. (15) yields the KL divergence $f_{OT\text{-}KL}$ as a function of $\hat{\boldsymbol{\mu}}^{(o)}$ and $\hat{\boldsymbol{\phi}}^{(o)}$. Its partial derivatives are given by:

$$\frac{\partial f_{OT\text{-}KL}}{\partial \hat{\boldsymbol{\mu}}^{(o)}} = \boldsymbol{S}_q^\top \text{diag}\left(\boldsymbol{S}_q \hat{\boldsymbol{\phi}}^{(o)}\right)\left(\boldsymbol{S}_q \hat{\boldsymbol{\mu}}^{(o)} - \boldsymbol{W}\hat{\boldsymbol{\mu}}_q^{(c)}\right)$$

$$\frac{\partial f_{OT\text{-}KL}}{\partial \hat{\boldsymbol{\phi}}^{(o)}} = \frac{1}{2}\boldsymbol{S}_q^\top \text{diag}^{-1}\left(\boldsymbol{W}\hat{\boldsymbol{\Sigma}}_q^{(c)}\boldsymbol{W}^\top + \boldsymbol{v}\boldsymbol{v}^T\right) - \frac{T}{2}\boldsymbol{\omega}^{(o)}$$

where $\boldsymbol{v} = \left(\boldsymbol{S}_q\hat{\boldsymbol{\mu}}^{(o)} - \boldsymbol{W}\hat{\boldsymbol{\mu}}_q^{(c)}\right)$, and $\hat{\boldsymbol{\omega}}^{(o)} = \text{diag}^{-1}\left(\hat{\boldsymbol{\Sigma}}^{(o)}\right)$. Hence, minimising $f_{OT\text{-}KL}$ with respect to $\hat{\boldsymbol{\mu}}^{(o)}$ gives:

$$\hat{\boldsymbol{\mu}}^{(o)} = \frac{1}{T}\boldsymbol{S}_q^\top\hat{\boldsymbol{\mu}}_q^{(o)} \quad (16)$$

$$\hat{\boldsymbol{\omega}}^{(o)} = \frac{1}{T}\boldsymbol{S}_q^\top\text{diag}^{-1}\left(\hat{\boldsymbol{\Sigma}}_q^{(o)} + \boldsymbol{v}\boldsymbol{v}^\top\right) \quad (17)$$

Note that the optimum estimation of the mean, $\hat{\boldsymbol{\mu}}^{(o)}$, is in fact the average of the T mean vectors in the observation trajectory space. Also, in practice, $\boldsymbol{v} \approx \boldsymbol{0}$. Hence, the optimum solution for $\hat{\boldsymbol{\omega}}^{(o)}$ is similarly the average of the variances in the observation trajectory space[4].

One main issues with the above estimation is that the resulting backward trajectory does not exactly reverse the forward trajectory step when $\boldsymbol{W}$ is not invertible (more details in Section V). This is mainly because the OT-KL ignores the temporal correlations between observation vectors in the observation trajectory space. In an attempt to suppress the *irreversibility* issue, an alternative estimation method is proposed. Suppose the covariance matrix $\hat{\boldsymbol{\Sigma}}_q^{(o)}$ is decomposed into $T\times T$ sub-blocks, $\{\hat{\boldsymbol{\Sigma}}_{ij}^{(o)}\}_{i,j=1}^{T}$, where $\hat{\boldsymbol{\Sigma}}_{ij}^{(o)}$ is a sub-block matrix of size $3M \times 3M$ at position (i, j). The optimal $\boldsymbol{\omega}^{(o)}$ given by Eq. (17) is essentially equal to $\frac{1}{T}\sum_i^T \text{diag}^{-1}(\hat{\boldsymbol{\Sigma}}_{ii}^{(o)})$. Since this computation of $\boldsymbol{\omega}^{(o)}$ only involves the diagonal sub-block matrices of $\hat{\boldsymbol{\Sigma}}_q^{(o)}$, we refer to as the *OT-KL-Diag* method.

While the above computation is theoretically justified as minimising the KL divergence, it discards the temporal information, i.e. off-diagonal sub-block matrices of the covariance matrix $\hat{\boldsymbol{\Sigma}}_q^{(o)}$. To overcome that, we suggest an alternative, *OT-KL-Full*, which considers all sub-block matrices of $\hat{\boldsymbol{\Sigma}}_q^{(o)}$, in computing $\boldsymbol{\omega}^{(o)}=\frac{1}{T}\sum_{i,j}^T \text{diag}^{-1}(\hat{\boldsymbol{\Sigma}}_{ij}^{(o)})$. We argue empirically in Section V that such approximation yields better performance.

2) Cepstral Trajectory Space KL (CT-KL): Minimising the KL divergence in the observation trajectory space ignores the explicit relationships between the static and dynamic parameters. An alternative solution obtains $(\hat{\boldsymbol{\mu}}^{(o)}, \hat{\boldsymbol{\Sigma}}^{(o)})$ by minimising $f_{CT\text{-}KL}$, the KL divergence in the cepstral trajectory space. The constrained statistics to be estimated in the cepstral trajectory space, $(\bar{\boldsymbol{\mu}}_q^{(c)}, \bar{\boldsymbol{\Sigma}}_q^{(c)})$, can be obtained using the forward trajectory transformation given in Eq. (12) and (13). The resulting KL divergence function is given as:

$$f_{CT\text{-}KL} = \frac{1}{2}\,\text{Tr}\left(\bar{\boldsymbol{\Sigma}}_q^{(c)})^{-1}\hat{\boldsymbol{\Sigma}}_q^{(c)}\right) - f\left(\hat{\boldsymbol{\mu}}^{(o)}, \hat{\boldsymbol{\phi}}^{(o)}\right) \quad (18)$$

where $f(\hat{\boldsymbol{\mu}}^{(o)}, \hat{\boldsymbol{\phi}}^{(o)})$ differs by only a constant from the log-likehood function $\log p(c|q, \Lambda)$ in [7] (Eq. 44) with $(\boldsymbol{c}, \overline{\boldsymbol{c}}_q, \boldsymbol{P}_q, \boldsymbol{\mu}_q)$ being substituted by $(\hat{\boldsymbol{\mu}}_q^{(c)}, \bar{\boldsymbol{\mu}}_q^{(c)}, \bar{\boldsymbol{\Sigma}}_q^{(c)}, \hat{\boldsymbol{\mu}}^{(o)})$. Hence, the partial derivatives of $f_{CT\text{-}KL}$ are as follows:

$$\frac{\partial f_{CT\text{-}KL}}{\partial \hat{\boldsymbol{\mu}}^{(o)}} = \boldsymbol{S}_q^\top\,\text{diag}\left(\boldsymbol{S}_q\hat{\boldsymbol{\phi}}^{(o)}\right)\boldsymbol{W}\left(\hat{\boldsymbol{\mu}}_q^{(c)} - \bar{\boldsymbol{\mu}}_q^{(c)}\right) \quad (19)$$

[4]We use $\hat{\boldsymbol{\Sigma}}^{(o)}$, $\hat{\boldsymbol{\phi}}^{(o)}$, and $\hat{\boldsymbol{\omega}}^{(o)}$ interchangeably.

$$\frac{\partial f_{CT\text{-}KL}}{\hat{\boldsymbol{\phi}}^{(o)}} = \frac{1}{2}\boldsymbol{S}_q^\top\,\text{diag}^{-1}\Big[2\hat{\boldsymbol{\mu}}^{(o)}(\hat{\boldsymbol{\mu}}_q^{(c)} - \bar{\boldsymbol{\mu}}_q^{(c)})^\top\boldsymbol{W}^\top + 2\boldsymbol{W}\hat{\boldsymbol{\Sigma}}_q^{(c)}\boldsymbol{W}^\top + \boldsymbol{W}\left(\bar{\boldsymbol{\mu}}_q^{(c)}\bar{\boldsymbol{\mu}}_q^{(c)^\top} - \hat{\boldsymbol{\mu}}_q^{(c)}\hat{\boldsymbol{\mu}}_q^{(c)^\top}\right)\boldsymbol{W}^\top\Big] \quad (20)$$

Equating Eq. (19) to $\boldsymbol{0}$ results in linear equations, which yields closed-form solutions for $\hat{\boldsymbol{\mu}}^{(o)}$ when $\hat{\boldsymbol{\phi}}^{(o)}$ is known:

$$\boldsymbol{S}_q^\top\boldsymbol{W}\bar{\boldsymbol{\Sigma}}_q^{(c)}\boldsymbol{W}^\top\boldsymbol{S}_q\hat{\boldsymbol{\Sigma}}^{(o)^{-1}}\hat{\boldsymbol{\mu}}^{(o)} = \boldsymbol{S}_q^\top\boldsymbol{W}\hat{\boldsymbol{\mu}}_q^{(c)} \quad (21)$$

As $f_{CT\text{-}KL}$ is not a quadratic function of $\hat{\boldsymbol{\phi}}^{(o)}$, the optimal value is found by using a gradient method with the partial derivative from Eq. (20) given a fixed $\hat{\boldsymbol{\mu}}^{(o)}$.

V. Properties of Trajectory PMC

In this section, several properties of the TPMC method will be discussed. First, the matrix $\boldsymbol{W}$, which encodes the explicit relationships between the static and dynamic parameters, will be examined. As previously mentioned, the size of the matrix $\boldsymbol{W}$ is $3MT \times MN$, where T and N denote the *lengths* of the trajectory in the observation and cepstral domains, respectively. In the original work of trajectory HMM [7], the trajectory length in the observation and cepstral domains are chosen to be the same ($N{=}T$) where the first and last 2δ columns[5] of $\boldsymbol{W}$ are truncated. In this work, we consider $N{=}(T + 4\delta)$. This allows for the flexibility of making $\boldsymbol{W}$ a square matrix (and invertible) when $T = 2\delta$, which simplifies the forward trajectory formulae (Eq. (13) and (12)) as:

$$\boldsymbol{\mu}_q^{(c)} = \boldsymbol{W}^{-1}\boldsymbol{S}_q\boldsymbol{\mu}^{(o)} \quad (22)$$

$$\boldsymbol{\Sigma}_q^{(c)} = \left[\boldsymbol{W}^\top\,\text{diag}\left(\boldsymbol{S}_q\boldsymbol{\phi}^{(o)}\right)\boldsymbol{W}\right]^{-1} \quad (23)$$

When $\boldsymbol{W}$ is square, a reversible backward trajectory that optimises the KL divergence in both the observation and cepstral trajectory domains leads to the solution given by Eq. (16) and Eq. (17)[6]. Another interesting property of TPMC when $\boldsymbol{W}$ is square is that the compensation of the static parameters is identical to that of the standard PMC. However, when N is larger than $T + 4\delta$, OT-KL estimation does not yield a reversible statistic transformation. Fig. 2 (left) shows the average KL divergence between the original and compensated models with no noise. A zero KL divergence indicates reversibility. It was found that for OT-KL-Diag estimation, the compensated model quickly diverges from the original model as N increases, most notably for the static and delta parameters. On the other hand, OT-KL-Full yields perfect reconstruction of the static parameters for different N. The divergence is smaller for the delta parameters but much larger than OT-KL-Diag for the delta-delta parameters. The non-reversibility property of OT-KL estimation is attributed to the fact that features in the observation space are assumed to be uncorrelated. On the other hand, the CT-KL estimation uses the trajectory HMM estimation approach [7], which imposes

[5]δ is the window length on each side of the current frame when computing dynamic parameters. We use $\delta{=}1$ for both delta and delta-delta computation.

[6]In this case, $\boldsymbol{v} = \boldsymbol{0}$ in Eq. (17).

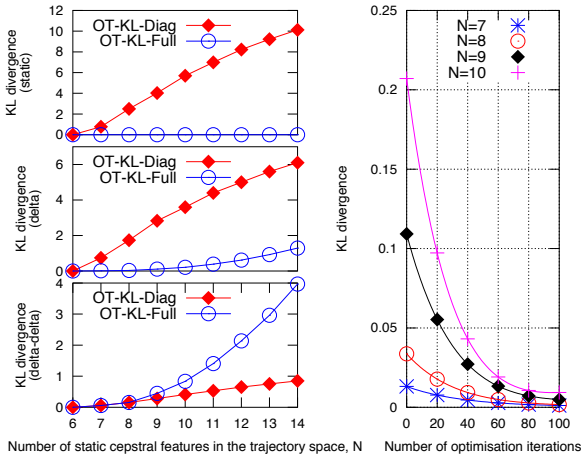

Fig. 2. *Left*: Reversibility comparison w.r.t. KL divergence for the static, delta and delta-delta parameters with increasing N. *Right*: Convergence comparison of KL divergence with increasing optimisation iterations for different N.

the explicit relationships between the static and dynamic parameters. Hence, the optimum CT-KL estimates will yield a reversible compensation when there is no noise. Fig. 2 (right) shows the convergence of the overall KL divergence between the original and compensated models for $N=7, 8, 9, 10$ with increasing optimisation iterations for the CT-KL estimation.

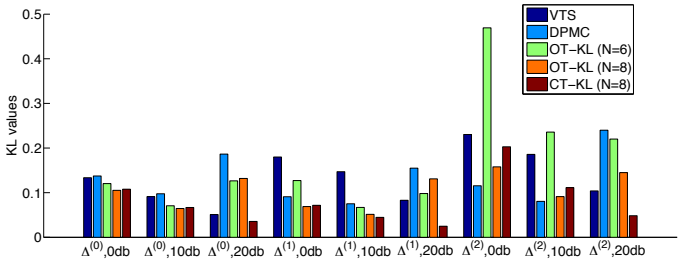

Fig. 3. KL divergences of different compensated single-component models w.r.t "reference" noisy speech methods for different SNRs and feature parts.

When noise is present, as the signal-to-noise ratios (SNRs) decrease, reversibility might not be as crucial as for high SNRs. To verify that, we compare in Fig. 3 the performance of different compensation methods (VTS and DPMC are added for completeness) across different settings. As anticipated, OT-KL with $N=8$ rivals the CT-KL when SNRs are low, e.g. 0dB and 10dB; whereas, CT-KL only yields superior performance with SNR$=$ 20dB. When $N=6$, the OT-KL method fails to compensate the delta-delta parameters, which we believe is partly due to the short trajectory length in the observation domain, i.e. $T=2$. VTS method, while demonstrates very good performance for static features, does not compensate very well for other feature parts, especially when SNRs are low. DPMC, on the other hand, yields good performance at low SNRs, but fail to perform consistently across different feature parts.

VI. Experiments

Experiments were conducted using the WSJCAM0 [10] corpus. The training data consists of 9889 utterances giving a total of 18.3 hours of data. The evaluation set is made up of the combination of the `si_dt5a` and `si_dt5b` development datasets. There are a total of 1.4 hours of test data. Both the training and test data were artificially corrupted by additive babble noise from the NOISEX database [11] to generate noisy speech data at signal-to-noise (SNR) ratios of 20dB, 10dB and 0dB. The noise data used to corrupt the training data were also used to estimate the noise model, which in this work, is a single Gaussian distribution.

All the acoustic models used in the subsequent experiments were decision-tree state clustered triphone HMM models, with approximately 4000 distinct states. These models were trained on 39-dimensional features comprising 13 static MFCC coefficients (including the `C0` term) together with the first and second order dynamic parameters. Firstly, clean speech models were trained on the original speech data provided by the WSJCAM0 corpus. In addition, noisy speech models were also trained on the artificially corrupted speech data at various SNRs to obtain a set of "reference" noisy speech models and assess the "upper bound" performance. The noisy speech models were trained by first performing Single Pass Retraining (SPR) [12] so that the initial state alignments were obtained using a clean model on speech speech data. These models were subsequently trained with three additional Baum-Welch iterations. The Word Error Rate

TABLE I
WER (%) Performance of Baseline Models

No. of Components	Model	WER (%)			
		Clean	20dB	10dB	0dB
1	Clean	17.21	27.30	57.40	95.28
	Noisy (SPR)	–	19.32	29.64	55.83
	Noisy (BW)	–	17.83	27.77	54.06
16	Clean	8.49	17.84	50.59	93.65
	Noisy (SPR)	–	9.48	17.59	44.11
	Noisy (BW)	–	9.47	16.67	44.12

(WER) performance of various models under different SNR conditions are summarised in Table I. Two sets of models with 1 and 16 Gaussian components per state were evaluated. In general, the WER performance degrades as SNR decreases. The relative WER reduction achieved by the SPR-trained noisy speech models over the clean speech models were 29.2%–48.4% for 1-component systems and 46.9%–65.2% for 16-component systems. With additional Baum-Welch retraining, further relative WER reduction of 5.2%–7.7% were observed except for the 16-component systems at SNR of 20dB and 0dB where the performance difference is very small.

First, the effect of the trajectory length of the cepstral trajectory domain, N, was investigated. Table II shows the WER performance comparison for TPMC models with N ranging from 6 to 10. As previously mentioned in Section V, the OT-KL estimation method for TPMC is *reversible* only when $\boldsymbol{W}$ is invertible (*i.e.* when N=6). Therefore, TPMC with

TABLE II
WER (%) PERFORMANCE OF 1-COMPONENT TPMC MODELS USING OT-KL-FULL ESTIMATION WITH DIFFERENT TRAJECTORY LENGTH, N

SNR (dB)	WER (%)				
	6	7	8	9	10
20	**21.61**	22.14	23.83	26.38	29.60
10	39.67	37.71	**37.51**	40.45	45.01
0	71.57	68.67	**66.70**	67.75	71.17

TABLE III
WER (%) PERFORMANCE OF 1-COMPONENT TPMC MODELS USING DIFFERENT BACKWARD TRAJECTORY ESTIMATION METHODS

N	Method	WER (%)		
		20dB	10dB	0dB
8	OT-KL-Full	23.83	37.51	66.70
	CT-KL	**22.27**	36.86	66.44
10	OT-KL-Full	29.60	45.01	71.17
	CT-KL	23.55	**36.58**	**64.93**

smaller N gave better performance in low noise conditions. With lower SNRs, $N=8$ was found to yield the best WER performance. Next, the parameter estimation methods for the TPMC reverse process are compared in Table III. In general, CT-KL estimation method outperforms OT-KL-Full since the relationships between the static and dynamic parameters are properly imposed, but at the expense of higher computational costs due to the gradient optimisation for the variance parameters. Furthermore, using the OT-KL-Full estimation with N=8 gave only marginally inferior performance compared to the best performing systems. Therefore, subsequent analyses will be based on this model.

TABLE IV
WER (%) PERFORMANCE COMPARISON OF 16-COMPONENT SYSTEMS USING VARIOUS NOISE COMPENSATION SCHEMES

Model	WER (%)		
	20dB	10dB	0dB
PMC	11.76	29.53	73.25
VTS	11.26	20.25	51.57
DPMC	17.32 (10.68)	28.42 (19.10)	58.90 (49.86)
TPMC	**10.84**	**19.68**	**50.20**

Finally, Table IV compares the WER performance of the proposed TPMC method with VTS, PMC and DPMC. Only the static parameters were compensated for PMC. DPMC models were trained by simulating an average of 500 noisy data samples per Gaussian component. In general, VTS outperforms PMC across different SNR conditions. The performance gain increases as SNR value drops because dynamic parameters were not compensated for PMC. Two sets of results were reported for DPMC. The top row refers to the results where DPMC models were estimated with fixed component alignments, *i.e.* data were simulated per Gaussian components. In this case, DPMC performed significantly worse than the other methods. However, if the data were simulated at the state level, allowing component alignments to be optimised, DPMC achieved the lowest WER (shown in parentheses). The proposed TPMC approach consistently outperformed both PMC and VTS across various SNR conditions. The performance of TPMC is also very competitive (only slightly worse) when compared to DPMC with component realignment. Nevertheless, it is worth pointing out that the proposed TPMC method is more efficient than DPMC since it does not involve synthesising noisy data and reestimation. Furthermore, TPMC is applied to each individual Gaussian components independently. Hence, there was no compensation for the component weights. Therefore, one possible extension for TPMC as a future work is to incorporate the compensation for Gaussian weight parameters.

VII. CONCLUSIONS

This paper has presented an extension to the standard Parallel Model Combination (PMC) technique that offers a solution to compensate both the static and dynamic parameters in a unified manner. The proposed method is called Trajectory PMC (TPMC) as it is motivated by the trajectory HMM formulation. The explicit relationships between the static and dynamic features are used to derive the statistics in the *cepstral trajectory* domain such that log normal approximation can be applied. The proposed TPMC method was found to yield consistently better performance compared to the standard PMC and VTS methods, both in terms of the Kullback-Leibler divergence and word error rate evaluations.

REFERENCES

[1] L. A. Rabiner, "A tutorial on hidden Markov models and selective applications in speech recognition," in *Proc. of the IEEE*, vol. 77, February 1989, pp. 257–286.

[2] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustic, Speech and Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.

[3] M. J. F. Gales, "Model-based techniques for noise robust speech recognition," Ph.D. dissertation, Gonville and Caius College, University of Cambridge, 1996.

[4] A. Acero, L. Deng, T. Kristjansson, and J. Zhang, "HMM adaptation using vector Taylor series for noisy speech recognition," in *Proc. of ICSLP*, vol. 3, 2000, pp. 869–872.

[5] M. Gales and S. Young, "A fast and flexible implementation of parallel model combination," *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, vol. 1, pp. 133–136, 1995.

[6] M. J. F. Gales and S. J. Young, "Cepstral parameter compensation for HMM recognition in noise," *Speech Commun.*, vol. 12, pp. 231–239, July 1993. [Online]. Available: http://portal.acm.org/citation.cfm?id=178489.178498

[7] H. Zen, K. Tokuda, and T. Kitamura, "Reformulating the HMM as a trajectory model by imposing explicit relationships between static and dynamic feature vector sequences," *Computer Speech & Language*, vol. 21, no. 1, pp. 153–173, 2007.

[8] ——, "An introduction of trajectory model into HMM-based speech synthesis," in *in Proc. of ISCA SSW5*, 2004, pp. 191–196.

[9] M. J. F. Gales, A. Ragni, H. AlDamarki, and C. Gautier, "Support vector machines for noise robust ASR," in *Proc. of Automatic Speech Recognition and Understanding Workshop*, 2009.

[10] T. Robinson *et al.*, *WSJCAM0 Cambridge Read News*. Linguistic Data Consortium, Philadelphia, 1995.

[11] A. Varga and H. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communication*, vol. 12, no. 3, pp. 247–251, 1993.

[12] S. J. Young *et al.*, *The HTK Book (for HTK version 3.4)*. Cambridge University, December 2006.