



Semi-parametric Trajectory Modelling Using Temporally Varying Feature Mapping for Speech Recognition

Khe Chai SIM and Shilin LIU

School of Computing, National University of Singapore, Singapore

{simkc, shilin}@comp.nus.edu.sg

Abstract

Recently, trajectory HMM has been shown to improve the performance of both speech recognition and speech synthesis. For efficiency, state sequence is required to compute likelihood for trajectory HMM which limits its use to N -best rescoring for speech recognition. Motivated by the success of models with temporally varying parameters, this paper proposes a Temporally Varying Feature Mapping (TVFM) model to transform the feature vector sequence such that the trajectory information as modelled by trajectory HMM is suppressed. Therefore, TVFM can be perceived as an implicit trajectory modelling technique. Two approaches for estimating the TVFM parameters are presented. Experimental results for phone recognition on TIMIT and word recognition on Wall Street Journal show that promising results can be obtained using TVFM.

Index Terms: state clustering, conditional random field, complexity control

1. Introduction

Hidden Markov Models (HMMs) are commonly used in speech recognition to represent the phone units. HMMs comprise a set of states where observation probability density functions are defined for each state and transition probabilities are defined between the states. HMMs made two fundamental assumptions which are typically not valid for speech data: 1) the first-order state transition which states that the probability of making a transition to the next state dependent only on the current state; 2) the observation independence assumption which states that the probability of making an observation depends only on the current state. These assumptions of HMMs lead to poor trajectory modelling since the trajectory within each state is constant. Commonly used approaches to circumvent this issue include appending the static feature vectors with the dynamic coefficients to capture the instantaneous trajectory information and using Gaussian Mixture Models (GMMs) to yield a piece-wise constant trajectory with a better resolution.

There have also been many attempts to model trajectory explicitly. Various trajectory modelling techniques are reviewed in [1, 2]. Recently, trajectory HMM has been shown to outperform conventional HMM systems by reformulating HMM as a generative model which takes into consideration explicitly the dynamic parameters learned by the HMM models [2]. This results in a smooth trajectory which has been found to be useful for parameter generation for speech synthesis [3]. However, state sequence is required to compute the likelihood for trajectory HMM which limits its use to N -best rescoring for speech recognition. Recently, several models with temporally varying parameters have been proposed, which can be perceived as semi-parametric trajectory modelling techniques [1]. In this pa-

per, a Temporally Varying Feature Mapping (TVFM) model is proposed to transform original feature sequence to suppress low order dynamic information so that they can be better modelled by standard HMMs. Two parameter estimation approaches are described.

The remaining of this paper is organised as follows. Section 2 describes the trajectory HMM model and how smooth trajectory can be obtained. Section 3 formulates the proposed Temporally Varying Feature Mapping (TVFM) and two parameter estimation approaches for TVFM are presented in Section 4. Experimental results are presented in Section 5.

2. Trajectory HMM

In speech recognition, static feature vectors such as the Mel Frequency Cepstral Coefficients (MFCC) and Perceptual Linear Prediction (PLP) are typically appended with the dynamic parameters to incorporate the trajectory information indirectly into the estimation of the HMM parameters. Up to the second order dynamic parameters are used such that the final feature vector is given by:

$$\mathbf{o}_t = [\mathbf{c}_t^T \quad \Delta \mathbf{c}_t^T \quad \Delta^2 \mathbf{c}_t^T]^T \quad (1)$$

where

$$\begin{aligned} \Delta \mathbf{c}_t &= \frac{\sum_{i=1}^{\delta} i(\mathbf{c}_{t+i} - \mathbf{c}_{t-i})}{2 \sum_{i=1}^{\delta} i^2} \\ \Delta^2 \mathbf{c}_t &= \frac{\sum_{i=1}^{\delta} i(\Delta \mathbf{c}_{t+i} - \Delta \mathbf{c}_{t-i})}{2 \sum_{i=1}^{\delta} i^2} \end{aligned}$$

δ is the window half-length for computing the dynamic parameters. Standard HMM assumes that these feature vectors, \mathbf{o}_t , are independent given the HMM state sequence, which is clearly invalid. This results in inconsistency between the static and dynamic parameters.

Recently, HMM has been reformulated as an explicit trajectory model such that the dynamic information is properly modelled in a generative sense. This reformulated model is referred to as trajectory HMM [2]. One of the important features of trajectory HMM is the ability to generate a maximum likelihood estimate of the trajectory of a feature vector sequence that obeys the consistency between the static and dynamic parameters given an HMM model. Let \mathbf{W} be the matrix that transforms the static features to include the dynamic parameters, such that

$$\mathbf{o} = \mathbf{W} \mathbf{c} \quad (2)$$

where $\mathbf{o} = [\mathbf{o}_1^T \mathbf{o}_2^T \dots \mathbf{o}_T^T]$ and $\mathbf{c} = [\mathbf{c}_1^T \mathbf{c}_2^T \dots \mathbf{c}_T^T]^T$. Thus, the generative model given by trajectory HMM can be expressed as [2]

$$p(\mathbf{o}|\mathbf{q}, \lambda) = \mathcal{N}(\mathbf{W} \mathbf{c}; \boldsymbol{\mu}_q^{(\mathbf{o})}, \boldsymbol{\Sigma}_q^{(\mathbf{o})}) \propto \mathcal{N}(\mathbf{c}; \tilde{\boldsymbol{\mu}}_q^{(\mathbf{c})}, \tilde{\boldsymbol{\Sigma}}_q^{(\mathbf{c})}) \quad (3)$$

where $\mathbf{q} = [q_1, q_2, \dots, q_T]$ is the state sequence of the utterance, which can be obtained using a Viterbi forced-alignment. $\boldsymbol{\mu}_q$ and $\boldsymbol{\Sigma}_q$ are the concatenated mean and covariance matrix of the entire utterance based on the state alignment, \mathbf{q} :

$$\begin{aligned}\boldsymbol{\mu}_q^{(o)} &= \left[\boldsymbol{\mu}_{q_1}^{(o)T} \quad \boldsymbol{\mu}_{q_2}^{(o)T} \quad \dots \quad \boldsymbol{\mu}_{q_T}^{(o)T} \right]^T \\ \boldsymbol{\Sigma}_q^{(o)} &= \text{diag} \left(\boldsymbol{\Sigma}_{q_1}^{(o)} \quad \boldsymbol{\Sigma}_{q_2}^{(o)} \quad \dots \quad \boldsymbol{\Sigma}_{q_T}^{(o)} \right)\end{aligned}$$

$\tilde{\boldsymbol{\mu}}_q^{(c)}$ and $\tilde{\boldsymbol{\Sigma}}_q^{(c)}$ are the corresponding mean vector and covariance matrix of trajectory HMM with respect to the static feature sequence, \mathbf{c} . They can be written in terms of the parameters of the HMM model as follows [2]:

$$\tilde{\boldsymbol{\mu}}_q^{(c)} = \tilde{\boldsymbol{\Sigma}}_q^{(c)} \mathbf{W}^T \boldsymbol{\Sigma}_q^{(o)-1} \boldsymbol{\mu}_q^{(o)} \quad (4)$$

$$\tilde{\boldsymbol{\Sigma}}_q^{(c)} = \left(\mathbf{W}^T \boldsymbol{\Sigma}_q^{(o)} \mathbf{W} \right)^{-1} \quad (5)$$

$\tilde{\boldsymbol{\mu}}_q^{(c)}$ is the maximum likelihood estimate of the trajectory of the static feature vector sequence given the trajectory HMM model, λ and $\tilde{\boldsymbol{\Sigma}}_q^{(c)}$ is the uncertainty associated with the estimate. Note that trajectory HMM requires the state sequence, \mathbf{q} , which limits its use to N -best rescoring for efficiency [2].

Therefore, the formulation of the standard and trajectory HMMs for the static vectors can be written as:

$$\mathbf{c}_t \sim \mathcal{N} \left(\boldsymbol{\mu}_{q_t}^{(c)}, \boldsymbol{\Sigma}_{q_t}^{(c)} \right) = \boldsymbol{\mu}_{q_t}^{(c)} + \mathbf{e}_t^{\text{hmm}} \quad (6)$$

$$\mathbf{c}_t \sim \mathcal{N} \left(\tilde{\boldsymbol{\mu}}_{q_t}^{(c)}, \tilde{\boldsymbol{\Sigma}}_{q_t}^{(c)} \right) = \tilde{\boldsymbol{\mu}}_{q_t}^{(c)} + \mathbf{e}_t^{\text{trajhmm}} \quad (7)$$

where $\boldsymbol{\mu}_{q_t}^{(c)}$ and $\tilde{\boldsymbol{\mu}}_{q_t}^{(c)}$ represent the trajectory as modelled by the standard and trajectory HMMs respectively. $\mathbf{e}_t^{\text{hmm}} = \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{q_t}^{(c)})$ and $\mathbf{e}_t^{\text{trajhmm}} = \mathcal{N}(\mathbf{0}, \tilde{\boldsymbol{\Sigma}}_{q_t}^{(c)})$ are the trajectory residuals from the standard and trajectory HMMs respectively. The distribution of these residuals are represented by the state covariance matrices of the respective models. In a standard HMM, $\boldsymbol{\mu}_{q_t}^{(c)}$ represents a piece-wise constant trajectory. Hence, the residual, $\mathbf{e}_t^{\text{hmm}}$, contains both *intra-frame* and *inter-frame* correlations. The former can be modelled using full covariance matrices or structured precision matrices [4]; while the latter is poorly modelled due to the *observation independence* assumption. Trajectory HMM, on the other hand, models a smoother trajectory using the low order dynamic information. Therefore, the slowly-evolving patterns are captured by trajectory HMM in $\tilde{\boldsymbol{\mu}}_{q_t}^{(c)}$, which suppresses *inter-frame* correlations in its residual, $\mathbf{e}_t^{\text{trajhmm}}$.

In this paper, a Temporally Varying Feature Mapping (TVFM) model is proposed to transform the static feature sequence such that low order dynamic information (such as the first two derivatives) is suppressed. The purpose of such feature transformation is to produce a static feature sequence whose trajectory is closer to a *piece-wise constant* trajectory, which can be better modelled by a standard HMM. A piece-wise constant trajectory can be obtained as follows:

$$\tilde{\mathbf{c}}_{q_t} = \mathbf{c}_t - \tilde{\boldsymbol{\mu}}_{q_t}^{(c)} + \boldsymbol{\mu}_{q_t}^{(c)} = \boldsymbol{\mu}_{q_t}^{(c)} + \mathbf{e}_t^{\text{trajhmm}} \quad (8)$$

where the trajectory HMM mean sequence is subtracted from the original static feature sequence and the standard HMM mean sequence is added to that. The resultant transformed feature sequence, $\tilde{\mathbf{c}}_{q_t}$, is simply a piece-wise constant trajectory of a standard HMM corrupted by the residual from trajectory

HMM where the inter-frame correlations have been reduced. In order to eliminate the dependence of $\tilde{\mathbf{c}}_{q_t}$ on the state sequence, q_t , TVFM will be used to approximate the transformation given in Equation 8 using only the original static feature sequence, \mathbf{c}_t . This leads to an implicit trajectory modelling technique, which will be described in the next section.

3. TVFM Formulation

The aim of the aforementioned implicit trajectory modelling technique is to find a mapping function that maps the original static feature sequence to a transformed feature sequence with the inter-frame correlations suppressed. Motivated by the semi-parametric trajectory modelling using temporally varying mean and precision matrix parameters [1], a Temporally Varying Feature Mapping (TVFM) model is formulated to achieve the mapping as follows:

$$\tilde{\mathbf{c}}_{q_t} \approx \mathbf{y}_t = f(\mathbf{c}) = \mathbf{A}(t)\mathbf{o}_t + \mathbf{b}(t) = \mathbf{M}(t)\boldsymbol{\xi}_t \quad (9)$$

where \mathbf{y}_t denotes the prediction of $\tilde{\mathbf{c}}_{q_t}$ using $\mathbf{A}(t)$ and $\mathbf{b}(t)$, which are the Temporally Varying Feature Mapping matrix and bias vector respectively. The transformation matrix and bias vector can be grouped into a single matrix, $\mathbf{M}(t)$, which corresponds to the linear transformation of the augmented feature vectors, $\boldsymbol{\xi}_t$ such that:

$$\mathbf{M}(t) = \left[\begin{array}{cc} \mathbf{A}(t) & \mathbf{b}(t) \end{array} \right] \quad \text{and} \quad \boldsymbol{\xi}_t = \left[\begin{array}{c} \mathbf{o}_t \\ 1 \end{array} \right] \quad (10)$$

The TVFM is modelled as a weighted linear superposition of S basis matrices as follows:

$$\mathbf{M}(t) = \sum_{s=1}^S P(s|t) \mathbf{M}_s \quad (11)$$

where \mathbf{M}_s denotes the s th basis matrix and $P(s|t)$ is the temporally varying superposition weights. The TVFM model in Equation 11 is equivalent in formulation to several existing models such as the Region-dependent Linear Transform (RDLT) [5], fMPE [6], pMPE [1] and SPLICE [7]. Typically, all these related models use a large number of basis, S , which may correspond to the number of Gaussian components in a large vocabulary continuous speech recognition system and the temporally varying weights are given by the posterior probabilities of these Gaussian components given the feature vector:

$$P(s|t) = \frac{p(\mathbf{o}_t|g_s)P(g_s)}{\sum_{r=1}^S p(\mathbf{o}_t|g_r)P(g_r)} \quad (12)$$

The basis matrices, \mathbf{M}_s , are typically estimated using discriminative criteria, such as Minimum Phone Error (MPE) [8].

Although TVFM shares the same formulation as several other related work, the motivation behind TVFM is primarily based on implicit trajectory modelling using trajectory information obtained using the trajectory HMM formulation, as described in Section 2. In particular, TVFM aims at removing the low order derivatives from the *trajectory* of the feature vector sequence to obtain a *pseudo* piece-wise constant trajectory, which can be better modelled by the standard HMM. Furthermore, the basis matrices are associated with the context-independent HMM states of the acoustic model. This leads to a much smaller number of basis, S . These basis weights are predicted from a long span of feature vectors around t using a feedforward neural network. In the next section, the parameter estimation of the TVFM will be described.

4. Parameter Estimation

The purpose of TVFM is to predict the trajectory-suppressed feature sequence from the original feature sequence, where the feature vector of each frame is transformed using Equation 11: Two parameter estimation approaches based on linear regression will be presented in this paper, namely the frame-based and model-based linear regression. These approaches will be described next.

4.1. Frame-based Linear Regression (FLR)

In a frame-based linear regression approach, linear regression is performed based on the pairs of feature sequences, $(\mathbf{c}_t, \boldsymbol{\xi}_t)$. A least squares estimate of the TVFM model can be obtained by minimising the following objective function:

$$\mathcal{Q}^{\text{LSE}} = \sum_{t=1}^T \sum_{s=1}^S P(s|t) (\mathbf{y}_t - \mathbf{M}_s \boldsymbol{\xi}_t)^T (\mathbf{y}_t - \mathbf{M}_s \boldsymbol{\xi}_t) \quad (13)$$

Differentiating the above equation and equating to zero leads to the following least squares estimation:

$$\mathbf{M}_s = \mathbf{K}_s \mathbf{G}_s^{-1} \quad (14)$$

where

$$\begin{aligned} \mathbf{G}_s &= \sum_{t=1}^T P(s|t) \boldsymbol{\xi}_t \boldsymbol{\xi}_t^T = \boldsymbol{\Sigma}_s^{(xx)} + \boldsymbol{\mu}_s^{(x)} \boldsymbol{\mu}_s^{(x)T} \\ \mathbf{K}_s &= \sum_{t=1}^T P(s|t) \mathbf{y}_t \boldsymbol{\xi}_t^T = \boldsymbol{\Sigma}_s^{(yx)} + \boldsymbol{\mu}_s^{(y)} \boldsymbol{\mu}_s^{(x)T} \end{aligned}$$

where $\boldsymbol{\mu}_s^{(x)}$ and $\boldsymbol{\mu}_s^{(y)}$ are the mean vectors of $\boldsymbol{\xi}_t$ and \mathbf{y}_t . Similarly, $\boldsymbol{\Sigma}_s^{(xx)}$ and $\boldsymbol{\Sigma}_s^{(yx)}$ denote the self covariance matrix of $\boldsymbol{\xi}_t$ and cross-covariance matrix of \mathbf{y}_t and $\boldsymbol{\xi}_t$ respectively. Hence, we have

$$\begin{aligned} \boldsymbol{\mu}_s^{(x)} &= \begin{bmatrix} \boldsymbol{\mu}_s^{(o)} \\ 1 \end{bmatrix} \\ \boldsymbol{\Sigma}_s^{(xx)} &= \begin{bmatrix} \boldsymbol{\Sigma}_s^{(oo)} & \boldsymbol{\mu}_s^{(o)} \\ \boldsymbol{\mu}_s^{(o)T} & 1 \end{bmatrix} \\ \boldsymbol{\Sigma}_s^{(yx)} &= \begin{bmatrix} \boldsymbol{\Sigma}_s^{(yo)} & \boldsymbol{\mu}_s^{(y)} \\ \boldsymbol{\mu}_s^{(o)T} & 1 \end{bmatrix} \end{aligned}$$

Manipulation of Equation 14 and decomposing \mathbf{M}_s as $\begin{bmatrix} \mathbf{A}_s & \mathbf{b}_s \end{bmatrix}$ yields

$$\mathbf{A}_s = \boldsymbol{\Sigma}_s^{(yo)} \boldsymbol{\Sigma}_s^{(oo)^{-1}} \quad (15)$$

$$\mathbf{b}_s = \boldsymbol{\mu}_s^{(y)} - \mathbf{A}_s \boldsymbol{\mu}_s^{(o)} \quad (16)$$

Incidentally, the transformation leads to the conditional expectation of \mathbf{y}_t given \mathbf{c}_t . This can be shown by substituting Equations 15 and 16 into Equation 11 to yield:

$$\begin{aligned} \mathbf{y}_t &= \sum_{s=1}^S P(s|t) \boldsymbol{\mu}_s^{(y|o)} \\ &= \sum_{s=1}^S P(s|t) \left\{ \boldsymbol{\mu}_s^{(y)} + \boldsymbol{\Sigma}_s^{(yo)} \boldsymbol{\Sigma}_s^{(oo)^{-1}} \left(\mathbf{o}_t - \boldsymbol{\mu}_s^{(o)} \right) \right\} \end{aligned}$$

Therefore, this approach is similar to the stereo-based stochastic mapping scheme presented in [9].

4.2. Model-based Linear Regression (MLR)

Alternatively, the TVFM parameters can also be estimated using a model-based linear regression approach. According to this approach, the trajectory-suppressed features are modelled using a regular HMM:

$$\mathbf{y}_t \sim \text{HMM} \left(\boldsymbol{\mu}_{q_t}^{(y)}, \boldsymbol{\Sigma}_{q_t}^{(yy)} \right) \quad (17)$$

and linear regression is used to transform \mathbf{c}_t such that the likelihood of the HMM model generating the transformed features is maximised. This can be achieved using the Constrained Maximum Likelihood Linear Regression (CMLLR) [10], where one set of CMLLR transform is estimated for each HMM state. CMLLR can be viewed as a linear transformation of the features, which is equivalent to constraining the mean vectors and covariance matrices within the same regression class to share the same regression transform. Therefore, the resulting CMLLR transformed system is given by:

$$\mathbf{o}_t \sim \text{HMM} \left(\mathbf{A}_{q_t}^{-1} (\boldsymbol{\mu}_{q_t}^{(y)} - \mathbf{b}_{q_t}), \mathbf{A}_{q_t}^{-1} \boldsymbol{\Sigma}_{q_t}^{(yy)} \mathbf{A}_{q_t}^{-T} \right) \quad (18)$$

In other words, the parameters of TVFM is obtained by finding the state-dependent CMLLR transforms to adapt the HMMs trained on \mathbf{y}_t to \mathbf{o}_t . Note that MLR-estimated TVFM uses *speaker-independent* CMLLR transforms.

5. Experimental Results

In this section, experimental results are reported for phone recognition on TIMIT database. The training and testing data sets have 3.13 and 1.14 hours of speech data respectively. Each monophone in the acoustic models is represented by a 3-state left-to-right HMM. These models are trained on 39 dimensional MFCC feature vectors which consist of 12 static coefficients, C0 energy and the first two derivatives. The temporally varying basis weights are the state posterior probabilities obtained using a feedforward neural network¹.

Table 1: Description of systems used in this paper.

Systems	Transform Type
A	None
B	CMLLR
C	TVFM (FLR)
D	TVFM (MLR)

Firstly, the various systems used in this experiment are summarised in Table 1. System A is a standard HMM system using GMM with diagonal covariance matrix to represent the state pdfs. System B uses CMLLR transforms with HMM state-dependent regression classes. This system is an example of structured precision matrix model [4] which aims at modelling intra-frame correlations. Systems C and D are TVFM systems whose model parameters are estimated using FLR and MLR respectively. These methods are described in Sections 4.1 and 4.2 respectively. Note that system D uses a different set of CMLLR transforms from those used in system B. In system D, the HMM models are trained on the trajectory-suppressed features.

¹Trained on TIMIT using ICSI quicknet software package, <http://www.icsi.berkeley.edu/speech/qn.htm>

Table 2: Comparison of PER (%) performance for FLR and MLR estimated TVFM models with varying number of Gaussian components per state.

Number of Components	PER (%)			
	A	B	C	D
1	54.39	45.37	49.90	38.52
2	50.12	43.28	46.89	37.77
4	47.01	41.76	44.48	37.39
8	44.25	40.54	42.23	37.43
16	41.76	39.26	40.13	36.48
32	39.43	38.01	38.42	36.40

The comparison of Phone Error Rate (PER) performance of various TVFM models estimated using FLR and MLR with varying number of Gaussian components per state is given in Table 2. In general, system performance improves with increasing number of Gaussian components per state. Larger improvements are observed for system with smaller number of Gaussian components per state. These improvements reduce as the number of Gaussian components per state increases. The performance of the baseline HMM system (A) improved from 54.39% to 39.43% as the number of Gaussian components per state increases from 1 to 32. System B, a speaker-independent CMLLR system, shows 1.42%–9.02% improvements over the standard HMM system (A). Note that for a single component system, system B is effectively a full covariance matrix system since a CMLLR transform is estimated for each Gaussian components. These improvements can be viewed as contributed from structured modelling of the covariance matrices. System C, a FLR estimated TVFM system, shows consistent improvements over system A. For a single Gaussian component systems, system C gave 4.49% absolute improvement. As the number of Gaussian components increases to 32, absolute improvement of 1.01% was obtained over system A. However, this system is slightly inferior to system B. On the other hand, system D attempts to model both intra-frame and inter-frame correlations using Temporally Varying Feature Mappings. This system gave 1.61%–4.50% absolute improvements over system B. These are additional improvements due to implicit inter-frame correlation modelling using TVFM. Moreover, system D also outperforms system C, which shows the advantage of estimating TVFM parameters using MLR on models trained on trajectory-suppressed features. It was also found that MLR-estimated TVFM using models trained on trajectory-suppressed features consistently outperformed those trained using standard features by 0.79%–2.61%..

Table 3: Comparison of WER (%) performance for MLR estimated TVFM for triphone models on WSJ0 test sets.

System	WER (%)	
	si_dt5a	si_dt5b
A	16.44	16.42
B	14.69	15.42
D	13.45	14.76

Next preliminary experiments with continuous speech recognition were also conducted on the Wall Street Journal (WSJCAM0) 5k task. This database consists of 18.30 hours of training data. The two test sets, si_dt5a and si_dt5b have

0.73 and 0.67 hours of speech respectively. The baseline system is a decision tree state-clustered triphone system with approximately 3400 distinct states and one Gaussian component per state. The baseline system gave 16.44% and 16.42% Word Error Rate (WER) on the si_dt5a and si_dt5b respectively. System B applied speaker-independent CMLLR and obtained 1.75% and 1.00% absolute WER reduction over system A on the two test sets. Finally, system D achieved the best performance of 13.45% and 14.76%, which corresponds to 1.24% and 0.66% absolute performance improvements over system B.

6. Conclusions

This paper has proposed Temporally Varying Feature Mapping (TVFM), an implicit trajectory modelling technique which aims at transforming the feature vector sequence to suppress trajectory information so that the resulting trajectory can be better modelled by a standard HMM. The trajectory information is extracted from the feature vector sequence using trajectory HMM. A set of linear transforms associated with the HMM states are estimated to predict the trajectory-suppressed features. Two parameter estimation methods were compared, namely the frame-based and model-based approaches. Experimental results for phone recognition on TIMIT database showed improvements using TVFM for both intra-frame and inter-frame correlation modelling. Preliminary continuous speech recognition results on Wall Street Journal also showed promising improvements.

7. References

- [1] K. C. Sim and M. J. F. Gales, "Discriminative semi-parametric trajectory model for speech recognition," *Comput. Speech Lang.*, vol. 21, no. 4, pp. 669–687, 2007.
- [2] H. Zen, K. Tokuda, and T. Kitamura, "A viterbi algorithm for a trajectory model derived from HMM with explicit relationship between static and dynamic features," in *Proc. of ICASSP*, 2004, pp. 837–840.
- [3] —, "An introduction of trajectory model into HMM-based speech synthesis," in *Proc. of 5th ISCA Speech Synthesis Workshop*, 2004.
- [4] K. C. Sim and M. J. F. Gales, "Minimum phone error training of precision matrix models," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 3, pp. 882–889, 2006.
- [5] B. Zhang, S. Matsoukas, and R. Schwartz, "Discriminatively trained region dependent transforms for speech recognition," in *Proc. of ICASSP*, 2006.
- [6] D. Povey, B. Kingsbury, L. Mangu, G. Saon, H. Soltau, and G. Zweig, "fMPE: Discriminatively trained features for speech recognition," in *Proc. ICASSP*, 2005.
- [7] J. Droppo and A. Acero, "Maximum mutual information SPLICE transform for seen and unseen conditions," in *Proc of Interspeech*, 2005.
- [8] D. Povey and P. C. Woodland, "Minimum Phone Error and I-smoothing for improved discriminative training," in *Proc. ICASSP*, 2002.
- [9] M. Afify, X. Cui, and Y. Gao, "Stereo-based stochastic mapping for robust speech recognition," *Trans. Audio, Speech and Lang.*, vol. 17, no. 7, pp. 1325–1334, 2009.
- [10] M. J. F. Gales and P. C. Woodland, "Mean and variance adaptation within the MLLR framework," *Computer Speech and Languages*, vol. 10, pp. 249–264, 1996.