IMPLICIT TRAJECTORY MODELLING USING TEMPORALLY VARYING WEIGHT REGRESSION FOR AUTOMATIC SPEECH RECOGNITION

Shilin LIU and Khe Chai SIM

School of Computing, National University of Singapore, Singapore
{shilin,simkc}@comp.nus.edu.sg

ABSTRACT

Recently, implicit trajectory modelling using temporally varying model parameters has achieved promising gains over the discriminatively trained standard HMM system. However, these works only focus on the temporally varying means or precisions explicitly. It is interesting to explore the capability of temporally varying weights, since the effect of time varying Gaussian parameters can be achieved by adjusting the weights of Gaussian Mixture Models (GMM) for different observation. This paper proposes a Temporally Varying Weight Regression (TVWR) model to learn the importance of different Gaussian components under different temporal contexts. Technically, TVWR factorizes the HMM state likelihood such that the contextual information can be modelled using time varying weights. Additionally, approximate constraints are derived to ensure a valid probabilistic model for TVWR. Experimental results for continuous speech recognition on Wall Street Journal show consistent improvements with varying system complexity and about 12% relative significant improvements in the best case.

Index Terms— trajectory modelling, complexity control, regression, nonlinear constrained optimization

1. INTRODUCTION

Hidden Markov Models (HMMs) are commonly used in speech recognition to represent a phone unit. Two fundamental assumptions are made so that efficient training and decoding algorithms can be implemented: 1) the first-order state transition probability only depends on the current state; 2) the observation output probability is independent of other states and observations given the current state. However, these assumptions are not valid for speech data which has strong temporal correlations. In order to relax these limitations, the state of the art HMM system always uses the features with dynamic coefficients and GMM to achieve a better resolution.

Several works have been attempted for trajectory modelling explicitly or implicitly [1, 2, 3, 4, 5]. fMPE [3] and RDLT [4] are successful examples implemented by temporally varying feature transformation, and these models can be explained as temporally varying means and precisions in the model space [1]. The success of these works comes from the temporally varying transformation for each frame in the feature or model space based on the rich information of temporal context. These motivate us adjusting the GMM weights to achieve the time varying emission distribution instead of modifying features or Gaussian means and/or precisions according to the different context.

The remaining of this paper is organized as follows. Section 2 gives an overview of implicit trajectory modelling. Section 3 formulates the proposed TVWR and constraints are derived for valid modelling. Section 4 presents the details of parameter estimation. Experimental results are reported in Section 5.

2. OVERVIEW OF IMPLICIT TRAJECTORY MODELLING

The independent output probability assumption can be circumvented by some trajectory modelling in the HMM framework. Instead of explicitly modelling the trajectory of the speech signal, implicit trajectory modelling is more popular for speech recognition in terms of efficiency and performance. The aim of trajectory modelling for speech recognition is to add the dependence in the model space [1, 2] or remove the dependence in the feature space [3, 4, 5] explicitly or implicitly by some kind of transformations. In this section, only model space view of implicit trajectory modelling is reviewed.

The state j output probability in many trajectory models with Gaussian mixture models can be written as follows:

$$p(\mathbf{o}_t|\tau_t, j) = \sum_{m=1}^{M} \underbrace{p(m|\tau_t, j)}_{c_{jmt}} \underbrace{p(\mathbf{o}_t|\tau_t, j, m)}_{\mathcal{N}(\mathbf{o}_t;\mu_{jmt},\mathbf{\Sigma}_{jmt})} \quad (1)$$

where τ_t is the context of observation $\mathbf{o}_t$, for simplicity, only limited surrounding observations are used as a approximation, denoted as $\tau_t = \{\mathbf{o}_{t-\delta}, \ldots, \mathbf{o}_{t-1}, \mathbf{o}_{t+1}, \ldots, \mathbf{o}_{t+\delta}\}$, where δ is the context expansion size, and the time varying parameters $\{c_{jmt}, \mu_{jmt}, \mathbf{\Sigma}_{jmt}\}$ are a function of $\{\tau_t, \mathbf{o}_t\}$.

Many works can be explained in this framework. If only the mean is temporally varying, Eq.1 becomes fMPE [3] or

RDLT [4], which can be formulated in the model space by a semi-parametric representation [1]:

$$\mu_{jmt} = \sum_{i=1}^{N} p(i|\mathbf{o}_t, \tau_t)\Big(\mathbf{A}^{(i)}\mu_{jm} + \mathbf{b}^{(i)}\Big)$$

where $p(i|\mathbf{o}_t, \tau_t)$ is the posterior of class i given the observation and its context.

If only the covariance matrix is temporally varying, Eq.1 becomes pMPE [1], a similar formulation can be used

$$\mathbf{\Sigma}_{jmt}^{-1} = \sum_{i=1}^{N} p(i|\mathbf{o}_t, \tau_t)^2 \mathbf{C}^{(i)} \mathbf{\Sigma}_{jm}^{-1} \mathbf{C}^{(i)T}$$

If both the weight and mean are temporally varying, Eq.1 becomes Buried Markov Model (BMM) [2]:

$$p(\mathbf{o}_t|\tau_t, j) = \sum_{m,i} \underbrace{p(m|i, \tau_t, j)p(i|\tau_t, j)}_{c_{jmit}} \underbrace{p(\mathbf{o}_t|m, i, \tau_t, j)}_{\mathcal{N}(\mathbf{o}_t;\mu_{jmit},\mathbf{\Sigma}_{jmi})}$$

3. FORMULATION OF TVWR

As only the temporally varying weight is modelled and other Gaussian parameters are constant within the state, the formulation of Eq.1 is not appropriate for the discussion of TVWR. Otherwise, the component output probability $p(\mathbf{o}_t|\tau_t, j, m)$ in Eq.1 has to be assumed to be independent of the context. Instead, the joint probability of observation $\mathbf{o}_t$ and the context τ_t is used for derivation. If no temporally varying parameters exist, this approach is just a standard HMM system using a long span of observations as features:

$$p(\mathbf{o}_t, \tau_t|j) = \sum_{m=1}^{M} p(m|j)p(\mathbf{o}_t, \tau_t|j, m)$$

This system can benefit from the better temporal correlation modelling but suffer from the highly increasing system complexity. TVWR is formulated based on the factorization of this joint probability:

$$p(\mathbf{o}_t, \tau_t|j) = \sum_{m=1}^{M} \underbrace{p(m|j)p(\tau_t|\mathbf{o}_t, j, m)}_{c_{jmt}} p(\mathbf{o}_t|j, m) \quad (2)$$

where the temporally varying weight is modelled by a product of two factors: 1). the static part, $p(m|j)$; 2) the dynamic part, $p(\tau_t|\mathbf{o}_t, j, m)$. More specifically, considering the convexity of the logarithm likelihood function and the positive requirement, the conditional density function is modelled as a regression function of some posteriors with respect to exponentials:

$$p(\tau_t|\mathbf{o}_t, j, m) = \frac{1}{Z}\sum_{i=1}^{N} h(i|\mathbf{o}_t, \tau_t)\exp\{\mathbf{w}_{jm}(i)\} \quad (3)$$

where $h(i|\mathbf{o}_t, \tau_t)$ is the posterior probability of class i given $\{\mathbf{o}_t, \tau_t\}$, which can be estimated by Neural Networks if supervised by phone label or Gaussian classifier if using unsupervised clustering, $\mathbf{w}_{jm}$ is the regression parameter to learn, Z is a global constant normalizer to be discussed later with more details. The constraint for the static part of c_{jmt} in Eq.2 can be easily derived:

$$\sum_{m=1}^{M} p(m|j) = 1 \quad \forall j \quad (4)$$

In order to make Eq.3 a valid density function, the following constraint is implied:

$$\int_{\tau_t} p(\tau_t|\mathbf{o}_t, j, m)\,\mathrm{d}\tau_t = 1 \quad \forall j, m, \mathbf{o}_t$$
$$\Longrightarrow \frac{1}{Z}\sum_{i=1}^{N}\exp\{\mathbf{w}_{jm}(i)\}\int_{\tau_t} h(i|\tau_t, \mathbf{o}_t)\,\mathrm{d}\tau_t = 1 \quad (5)$$

Although Z is a global constant, which can be ignored during training and decoding without affecting the performance, it has to be defined explicitly to make the constraint Eq.5 workable. Setting $\mathbf{w}_{jm} = \mathbf{0}$, which works as a reasonable starting point of TVWR, can give a valid definition:

$$Z = \sum_{i=1}^{N}\int_{\tau_t} h(i|\tau_t, \mathbf{o}_t)\,\mathrm{d}\tau_t \quad (6)$$

Note that Z is independent of $\mathbf{o}_t$ though $\mathbf{o}_t$ occurs in the above expression, given $\sum_{i=1}^{N} h(i|\tau_t, \mathbf{o}_t) = 1$. After Z is known, the Eq.5 can be rewritten as follows:

$$\sum_{i=1}^{N}\exp\{\mathbf{w}_{jm}(i)\}p(i|\mathbf{o}_t, h) = 1 \quad \forall j, m, \mathbf{o}_t \quad (7)$$

where

$$p(i|\mathbf{o}_t, h) = \frac{\int_{\tau_t} h(i|\tau_t, \mathbf{o}_t)\,\mathrm{d}\tau_t}{\sum_{i=1}^{N}\int_{\tau_t} h(i|\tau_t, \mathbf{o}_t)\,\mathrm{d}\tau_t} \quad (8)$$

and h is used to add some possible dependence on the mapping function from $\{\mathbf{o}_t, \tau_t\}$ to the class posteriors. There are two critical issues to apply constraints by Eq.7. The first one is the estimation of $p(i|\mathbf{o}_t, h)$, which can't be estimated directly from the training data.The second one is that there could be infinite number of constraints if $p(i|\mathbf{o}_t, h)$ has a strong dependence on $\mathbf{o}_t$. Therefore, an approximation of the constraints is made by dropping the dependency on $\mathbf{o}_t$, which gives $p(i|\mathbf{o}_t, h) \approx p(i|h)$ and $p(i|h)$ can be obtained by following two ways:

1). sample based approximation

$$p(i|h) \approx \frac{\sum_t h(i|\mathbf{o}_t, \tau_t)}{\sum_i\sum_t h(i|\mathbf{o}_t, \tau_t)} \quad (9)$$

2). uniform distribution assumption

$$p(i|h) = \frac{1}{N} \quad (10)$$

4. PARAMETER ESTIMATION

In this section, only the time varying weight parameter estimation is discussed, while other Gaussian parameter estimation is standard and not covered here. Instead of directly optimizing the likelihood function. its auxiliary function is to be maximized:

$$Q^{ML} = \sum_{t,j,m} \gamma_{jm}(t) \log(c_{jmt})$$

$$= \sum_{t,j,m} \gamma_{jm}(t) \Big(\log(c_{jm}) + \log \big(\sum_i h(i|\hat{\tau}_t) \exp\{\mathbf{w}_{jm}(i)\} \big) \Big)$$

subject to constraints by Eq.4 and Eq.7 for all $j = 1, 2, \ldots, S$, where $\gamma_{jm}(t)$ is the Gaussian component posterior, $c_{jm} = p(m|j)$, and $\hat{\tau}_t = \{\tau_t, \mathbf{o}_t\}$ is used to simplify the notation.. Since the static parameter c_{jm} constrained by Eq.4 can be optimized using the method by the standard system, we only focus on the optimization of the $\mathbf{w}_{jm}$,

The above objective function is hard to optimize since there is a summation inside the logarithm function. Therefore, a lower bound is derived based on Jensen's inequality so that increasing such lower bound can guarantee the increase or no change of the original function.

$$Q^{ML} \geq G = \sum_{t,j,m} \gamma_{jm}(t) \sum_i h(i|\hat{\tau}_t) \mathbf{w}_{jm}(i) + K_{const}$$

$$= \sum_{t,j,m,i} \gamma_{jm}(t) h(i|\hat{\tau}_t) \mathbf{w}_{jm}(i) + K_{const}$$

where K_{const} is the term independent of $\mathbf{w}_{jm}$.

In order to optimize the regression parameters for a particular state j, the problem becomes maximizing the following linear function:

$$G^{(j)} = \sum_m^M \mathbf{w}_{jm}^T \mathbf{h}_{jm}$$

subject to constraints by Eq.7, where

$$\mathbf{h}_{jm} = \sum_t^T \gamma_{jm}(t) \mathbf{h}_t$$

$$\mathbf{h}_t = [h(1|\hat{\tau}_t), h(2|\hat{\tau}_t), \ldots, h(N|\hat{\tau}_t)]^T$$

Based on the spirit of EM algorithm, the statistics $\mathbf{h}_{jm}$ are assumed to be fixed and known when $\mathbf{w}_{jm}$ is changing. This is a good news for constrained optimization problem, which requires iterative evaluation of the objective function $G^{(j)}$. Once $\mathbf{h}_{jm}$ are accumulated by one pass of training data, Lagrangian function can be applied for this maximum problem:

$$\mathcal{L}(\mathbf{w}_{j.}, \lambda_.) = \sum_m^M \mathbf{w}_{jm}^T \mathbf{h}_{jm}$$

$$+ \sum_m^M \lambda_m \Big(\sum_i^N \exp\{\mathbf{w}_{jm}(i)\} p(i|h) - 1 \Big)$$

and then solve the below equation system:

$$\nabla_{\mathbf{w}_{jm}} \mathcal{L}(\mathbf{w}_{j.}, \lambda_.) = 0 \quad \forall m$$

$$\nabla_{\lambda_m} \mathcal{L}(\mathbf{w}_{j.}, \lambda_.) = 0 \quad \forall m$$

However, the above equation system is nonlinear, whose solution is hard to be obtained. Thus, a nonlinear constrained optimization tool implemented by interior point method [6] is used.

5. EXPERIMENTAL RESULTS

In this section, experimental results are reported for word recognition on the Wall Street Journal (WSJCAM0) 5k task. This database consists of 18.30 hours of training data and 1.40 hours of testing data. The baseline system is a decision tree state-clustered triphone system with approximately 3400 distinct states. Each triphone unit is modelled by a 3-state left-to-right HMM. These models are trained on 39 dimensional MFCC features, including 12 static coefficients, `C0` energy and the first two differentials. The temporally varying class posteriors are the 40-monophone posterior probabilities obtained by using a feedforward neural network [1] where the input is $\hat{\tau}_t$ with $\delta = 4$, the hidden layer has 1000 sigmoid activated neurons, and the output layer uses softmax activation function to generate the posteriors.

For each iteration of TVWR training, only one pass over the data is enough to accumulate the statistics $\mathbf{h}_{jm}$ for the regression parameter learning, and other standard statistics. The maximum iteration of the constraint nonlinear optimizer is set to be 100, which is shown large enough for the convergence with 1e-6 tolerance. In order to make an efficient calculation of temporally varying weights, the minimum class posterior is set to be 0.01.Setting c_{jm} to be the standard weight and $\mathbf{w}_{jm}$ to be zero can give a reasonable starting point for TVWR training, i.e. the original standard HMM system. Since both approximations by Eq.9 and Eq.10 show very close performance, only the TVWR system approximated by Eq.10 are reported here. Note in the following discussion, the number in #-HMM or #-TVWR below specifies the number of Gaussian components per state within the respective system.

Firstly, the likelihood of TVWR and HMM are plotted in Figure 1. The starting point of TVWR is the standard system with 8 components per state by 4 iterations of ML training. TVWR training includes the estimation of parameters for time varying weights and all the other standard parameters. The standard system 12-HMM with the same system complexity as 8-TVWR, i.e. 4.8m parameters, shows consistent higher likelihood than 8-HMM as expected. In the case of TVWR training, the biggest likelihood improvements comes from the first iteration training, and the likelihood is always

[1] Using ICSI quicknet software package, `http://www.icsi.berkeley.edu/speech/qn.htm`

increasing with fast convergent speed. This shows us the estimation algorithm is well defined. It is not interesting to compare the likelihood between TVWR and standard HMM since the global normalizer Z by Eq.6 is ignored during training.

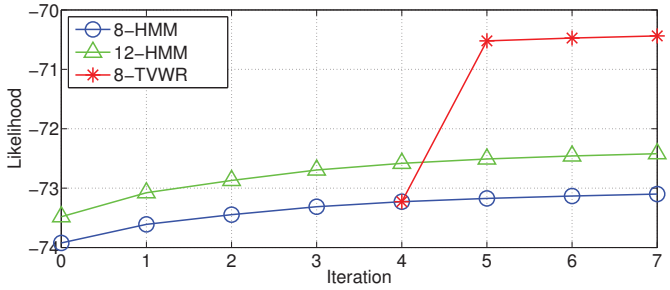

Fig. 1. *Likelihood for TVWR and standard HMM system with the same number of parameters or components.*

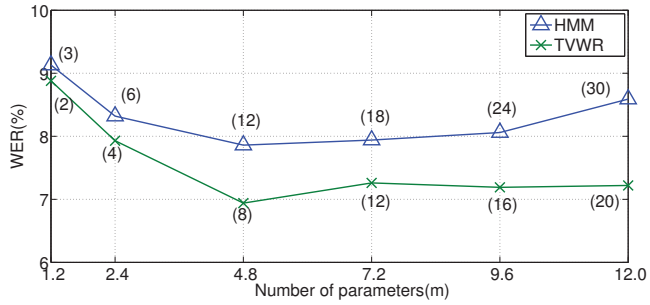

Fig. 2. *Comparison of Word Error Rate (WER%) for TVWR and standard HMM system with varying system complexity; the number of components per state for perspective system is specified in parentheses.*

Secondly, the overall performances of TVWR and standard HMM system are compared with different system complexity. Each standard HMM system is obtained by 8 iterations of Maximum Likelihood training. At the beginning, the performance of standard HMM system improves by increasing the system complexity, as shown in Figure 2. However, the performance decreases when the number of parameters is larger than 4.8m. This is expectable since the ML trained standard system is not robust for over training. Compared to this standard system, TVWR shows consistent improvements for different system complexity. This shows us the adjustment of Gaussian weights is learned properly according to the temporal context information. In addition to that, there are two more interesting findings in this figure: 1). when fewer number of components are used, the improvements of TVWR is quite small; 2). when the number of components are large enough, the performance seems converged. The first finding shows us the performance of TVWR is sensitive to the number of adjustable components weights but a optimal number exists, e.g. 8 in our set up. The second finding shows us TVWR seems to be robust for over training considering the convergent performance by increasing the system complexity. It is probably because the conditional context probability in Eq.3 is now modelled by a discriminative model, i.e. neural network, rather than a conventional generative model.

Lastly, the best results of TVWR and standard HMM system in Figure 2 are compared. 8-TVWR with 4.8m parameters shows the best performance by 6.94% WER among the TVWR systems, while 12-HMM with the same parameters also shows the best performance by 7.86% WER among the HMM systems. 8-TVWR obtained 0.92% absolute improvements or 12% relative improvements over 12-HMM, which are statistically significant at the level of p-values <0.001.

6. CONCLUSIONS

This paper has proposed an implicit trajectory model using Temporally Varying Weight Regression (TVWR) to learn the importance of Gaussian components under different temporal contexts so that the dynamic acoustic patterns can be better recognized. TVWR factorizes the HMM state likelihood such that the contextual information can be modelled using time varying weights. Parameter estimation is discussed as a constrained nonlinear optimization problem. Experimental results for continuous speech recognition on Wall Street Journal database showed that consistent improvements can be obtained using the proposed TVWR.

7. REFERENCES

[1] K.C. Sim and M.J.F. Gales, "Discriminative semi-parametric trajectory model for speech recognition," *Computer Speech & Language*, vol. 21, no. 4, pp. 669–687, 2007.

[2] J.A. Bilmes, "Buried markov models for speech recognition," in *Proc. ICASSP*. IEEE, 1999, vol. 2, pp. 713–716.

[3] D. Povey, B. Kingsbury, L. Mangu, G. Saon, H. Soltau, and G. Zweig, "fmpe: Discriminatively trained features for speech recognition," in *Proc. ICASSP*. IEEE, 2005, vol. 1, pp. 961–964.

[4] B. Zhang, S. Matsoukas, and R. Schwartz, "Discriminatively trained region dependent feature transforms for speech recognition," in *Proc. ICASSP*. IEEE, 2006, vol. 1, pp. I–I.

[5] K.C. Sim and S. Liu, "Semi-parametric trajectory modelling using temporally varying feature mapping for speech recognition," in *Proc. INTERSPEECH*. ISCA, 2010, pp. 2982–2985.

[6] A. Wächter and L.T. Biegler, "On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming," *Mathematical Programming*, vol. 106, no. 1, pp. 25–57, 2006.