INTERSPEECH
2009 BRIGHTON

Hidden Conditional Random Field with Distribution Constraints for Phone Classification

Dong Yu, Li Deng, Alex Acero

Speech Research Group, Microsoft Research, USA

{dongyu, deng, alexac}@microsoft.com

Abstract

We advance the recently proposed hidden conditional random field (HCRF) model by replacing the moment constraints (MCs) with the distribution constraints (DCs). We point out that the distribution constraints are the same as the traditional moment constraints for the binary features but are able to better regularize the probability distribution of the continuous-valued features than the moment constraints. We show that under the distribution constraints the HCRF model is no longer log-linear but embeds the model parameters in non-linear functions. We provide an effective solution to the resulting more difficult optimization problem by converting it to the traditional log-linear form at a higher-dimensional space of features exploiting cubic spline. We demonstrate that a 20.8% classification error rate (CER) can be achieved on the TIMIT phone classification task using the HCRF-DC model. This result is superior to any published single-system result on this heavily evaluated task including the HCRF-MC model, the discriminatively trained HMMs, and the large-margin HMMs using the same features.

Index Terms: hidden conditional random field, maximum entropy, moment constraint, distribution constraint, phone classification, TIMIT, cubic spline

1. Introduction

The recently proposed hidden conditional random field (HCRF) model [2][5] is a class of discriminative models that generalize both the hidden Markov model (HMM) and the conditional random field (CRF) model. The HCRF models are direct models that produce the most probable state sequence by estimating the conditional probability of a state sequence given the entire observed feature sequence. Different from HMMs which model the state sequence as being Markov and each observation being independent of all others given the state, HCRFs model the state sequence as being conditionally Markov given the observation sequence and is capable of representing long-range feature dependencies and incorporating highly correlated features. The HCRF model has been successfully applied to the phone classification task [2][5], the meeting segmentation task [6], and the electrocardiogram classification task [1], to name a few.

The core of the HCRF model is the maximum entropy (MaxEnt) model with moment constraints (MCs). The moment constraint requires that the expected value of each feature estimated from the model be the same as that observed in the training data. The MaxEnt principle indicates that among all the probability distributions that accord with the moment constraints, we should choose the one that maximizes the entropy. The conditional probability in the HCRF model has a nice log-linear form [2]

$$p(w|o;\boldsymbol{\lambda}) = \frac{1}{z(o;\boldsymbol{\lambda})}\sum_{s\in w} exp\left(\boldsymbol{\lambda}^T \boldsymbol{f}(w,s,\boldsymbol{o})\right), \qquad (1)$$

where $(\cdot)^T$ is the transposition of $(\cdot)$, $o = (o_1,\cdots,o_T)$ is the observation sequence, w is the state sequence without boundary information and is typically represented as a phoneme sequence or word sequence, $s = (s_1,\cdots,s_T)$ is a state sequence with boundary information, $\boldsymbol{f}(w,s,o) = [f_1(w,s,o),\cdots,f_N(w,s,o)]^T$ is the feature vector, $\boldsymbol{\lambda} = [\lambda_1,\cdots,\lambda_N]^T$ is the weight vector, and $z(o;\boldsymbol{\lambda}) = \sum_{w,s\in w} exp\left(\boldsymbol{\lambda}^T\boldsymbol{f}(w,s,o)\right)$ is the normalization factor to ensure probabilities $p(w|o;\boldsymbol{\lambda})$'s sum to 1.

Both binary and continuous features can be used in the HCRF model. For example, in the speech classification and recognition tasks [2][5] the sufficient statistics

$$f_{w'}^{(LM)}(w,s,o) = \delta(w = w') \qquad \forall w' \qquad (2)$$

$$f_{s"s'}^{(Tr)}(w,s,o) = \sum_{t=1}^{T}\delta(s_{t-1} = s")\delta(s_t = s') \ \forall s",s' \qquad (3)$$

$$f_{s'}^{(Occ)}(w,s,o) = \sum_{t=1}^{T}\delta(s_t = s') \qquad \forall s' \qquad (4)$$

$$f_{s'}^{(M1)}(w,s,o) = \sum_{t=1}^{T}\delta(s_t = s')\, o_t \qquad \forall s' \qquad (5)$$

$$f_{s'}^{(M2)}(w,s,o) = \sum_{t=1}^{T}\delta(s_t = s')\, o_t^2 \qquad \forall s' \qquad (6)$$

were used as features, where $\delta(x) = 1$ if x is true, and $\delta(x) = 0$ otherwise. $f_{w'}^{(LM)}(w,s,o)$ are language model (LM) features, $f_{s"s'}^{(Tr)}(w,s,o)$ are state transition features, $f_{s'}^{(Occ)}$ are state occupation features, and $f_{s'}^{(M1)}(w,s,o)$ and $f_{s'}^{(M2)}(w,s,o)$ are the first and second-order statistics generated from the observations. Among these features, $f_{w'}^{(LM)}(w,s,o)$ are binary features, $f_{s"s'}^{(Tr)}(w,s,o)$ and $f_{s'}^{(Occ)}(w,s,o)$ are multi-valued discrete features, and $f_{s'}^{(M1)}(w,s,o)$ and $f_{s'}^{(M2)}(w,s,o)$ are continuous features.

As we have shown in our recent work [9], the moment constraint is a strong one for binary features since knowing the mean of the binary feature is equivalent to knowing its probability distribution in full. However, the moment constraint is very weak for continuous features and as a result we learn less than should from the training data if only moment constraints are used for the continuous features. In this paper, we advance the HCRF model by replacing the moment constraints with the distribution constraints (DCs). We show that under distribution constraints the HCRF model is no longer log-linear but with non-linear functions which embed the model parameters to be estimated. We provide a solution to this significantly more difficult optimization problem by converting it to a log-linear problem at a higher-dimensional space using cubic spline. We demonstrate that a 20.8% classification error rate (CER) can be achieved on the TIMIT phone classification task using the HCRF-DC model, which is 0.5% better than the best result obtained using the HCRF-MC model [5] and 0.3% better than the best single-system result reported on the same task which was achieved using the large-margin technique [7].

The rest of the paper is organized as follows. In Section 2, we derive the improved HCRF model with distribution constraints and show how we can solve the resulting optimization problem that contains non-linear functions as parameters by converting it to a log-linear problem at a higher-dimensional space. In Section 3, we describe the experimental results on the TIMIT phone classification task. We conclude the paper in Section 4.

2. HCRF with Distribution Constraints

In this section, we derive the HCRF model with distribution constraints and provide a solution to the more complicated model estimation problem. For the sake of clarity, we simplify $f_i(w, s, o)$ as f_i in the rest of the paper.

2.1. Distribution constraints

The distribution constraint of a continuous feature can be approximated with bucketing approaches, with which each continuous feature f_i in the range of $[l_i, h_i]$ is converted into K binary features in the form of

$$f_{ik} = \begin{cases} \frac{h_{ik} + l_{ik}}{2} & if\ f_i \in [l_{ik}, h_{ik}] \\ 0 & otherwise \end{cases} \tag{7}$$

where $l_{ik} = h_{i(k-1)} = (k-1)(h_i - l_i)/K + l_i$. The HCRF model with moment constraints on the quantized features has the form of

$$p(w|o;\boldsymbol{\lambda}) = \frac{1}{z(o;\boldsymbol{\lambda})} \sum_{s \in w} exp\left(\sum_{i=1}^{C} \sum_{k=1}^{K} \lambda_{ik} f_{ik} + \sum_{j=1}^{B} \lambda_j f_j \right) \tag{8}$$

where B is the number of binary features and C is the number of continuous features. This approximation is very rough. Now, let us assume we have infinite number of training samples so that we can increase the number of buckets K to infinity. By noting that only one f_{ik} is non-zero and the non-zero f_{ik} takes the value of f_i, we get

$$\lim_{K \to \infty} \sum_k \lambda_{ik} f_{ik} = \lambda_i(f_i) f_i. \tag{9}$$

Eq. (8) thus becomes

$$p(w|o;\boldsymbol{\lambda}) = \frac{1}{z(o;\boldsymbol{\lambda})} \sum_{s \in w} exp\left(\sum_{i=1}^{C} \lambda_i(f_i) f_i + \sum_{j=1}^{B} \lambda_j f_j \right). \tag{10}$$

Note that the weight $\lambda_i(f_i)$ for the continuous feature f_i is no longer a single value but a (generally nonlinear) function of the continuous feature f_i.

2.2. Solution to the optimization problem

Since the model parameters to be estimated are functions instead of single values, the parameter estimation problem cannot be easily solved. Here we provide a solution by converting it to the standard log-linear form at a higher-dimensional space using the cubic-spline-based technique we recently developed [9][11][12][13]. The core idea is to approximate the continuous weight function (instead of the feature) with splines.

Given K evenly distributed knots $\{(f_{ik}, \lambda_{ik})\}\ k = 1, \cdots, K$ in the cubic spline with the natural boundary condition (i.e., the second derivatives at the boundaries are 0), and denote $h = f_{i(k+1)} - f_{ik} = f_{i(j+1)} - f_{ij} > 0, \forall j, k \in \{1, \cdots, K-1\}$, the value $\lambda_i(f_i)$ of a data point f_i can be estimated as

$$\lambda_i(f_i) = a\lambda_{ij} + b\lambda_{i(j+1)} + c\frac{\partial^2 \lambda_i}{\partial {f_i}^2}|_{f_i = f_{ij}} + d\frac{\partial^2 \lambda_i}{\partial {f_i}^2}|_{f_i = f_{i(j+1)}} \tag{11}$$

where

$$\begin{aligned} a &= \frac{f_{i(j+1)} - f_i}{f_{i(j+1)} - f_{ij}}, \\ b &= 1 - a, \\ c &= \frac{1}{6}(a^3 - a)\left(f_{i(j+1)} - f_{ij}\right)^2, and \\ d &= \frac{1}{6}(b^3 - b)\left(f_{i(j+1)} - f_{ij}\right)^2 \end{aligned} \tag{12}$$

are interpolation parameters, and $\left[f_{ij}, f_{i(j+1)}\right]$ is the section in which the point f_i falls. As we have shown in [9][11][12][13], $\lambda_i(f_i)$ can be written in the matrix form

$$\lambda_i(f_i) \cong \boldsymbol{a}^{\boldsymbol{T}}(f_i)\boldsymbol{\lambda}_i \tag{13}$$

where

$$\boldsymbol{\lambda}_i = [\lambda_{i1}, \cdots, \lambda_{iK}]^T \tag{14}$$

is the weight vector for feature f_i, and

$$\boldsymbol{a}^{\boldsymbol{T}}(f_i) = \boldsymbol{e}^T(f_i) + \boldsymbol{f}^T(f_i)\boldsymbol{C}^{-1}\boldsymbol{D} \tag{15}$$

is a vector with

$$\boldsymbol{e}^{\boldsymbol{T}}(f_i) = \begin{bmatrix} 0 & \cdots & \underbrace{a}_{j} & \underbrace{b}_{J+1} & \cdots & 0 \end{bmatrix}, \tag{16}$$

$$\boldsymbol{f}^{\boldsymbol{T}}(f_i) = \begin{bmatrix} 0 & \cdots & \underbrace{c}_{j} & \underbrace{d}_{J+1} & \cdots & 0 \end{bmatrix}, \tag{17}$$

$$\boldsymbol{C} = \begin{bmatrix} \frac{h}{6} & \mathbf{0} & \mathbf{0} & \cdots & \cdots & \cdots & \mathbf{0} \\ \frac{h}{6} & \frac{2h}{3} & \frac{h}{6} & \mathbf{0} & \cdots & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \cdots & \mathbf{0} & \frac{h}{6} & \frac{2h}{3} & \frac{h}{6} & \mathbf{0} & \vdots \\ \vdots & \vdots & \mathbf{0} & \vdots & \vdots & \vdots & \mathbf{0} \\ \mathbf{0} & \cdots & \vdots & \mathbf{0} & \frac{h}{6} & \frac{2h}{3} & \frac{h}{6} \\ \mathbf{0} & \cdots & \cdots & \cdots & \mathbf{0} & \mathbf{0} & \frac{h}{6} \end{bmatrix}, \tag{18}$$

$$\boldsymbol{D} = \begin{bmatrix} 0 & \mathbf{0} & \mathbf{0} & \cdots & \cdots & \cdots & \mathbf{0} \\ \frac{1}{h} & -\frac{2}{h} & \frac{1}{h} & \mathbf{0} & \cdots & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \cdots & \mathbf{0} & \frac{1}{h} & -\frac{2}{h} & \frac{1}{h} & \mathbf{0} & \vdots \\ \vdots & \vdots & \mathbf{0} & \vdots & \vdots & \vdots & \mathbf{0} \\ \mathbf{0} & \cdots & \vdots & \mathbf{0} & \frac{1}{h} & -\frac{2}{h} & \frac{1}{h} \\ \mathbf{0} & \cdots & \cdots & \cdots & \mathbf{0} & \mathbf{0} & 0 \end{bmatrix}. \tag{19}$$

From Eq. (13), we obtain

$$\lambda_i(f_i)f_i \cong \boldsymbol{a}^{\boldsymbol{T}}(f_i)\boldsymbol{\lambda}_i f_i = [\boldsymbol{a}^{\boldsymbol{T}}(f_i)f_i]\boldsymbol{\lambda}_i = \sum_k \lambda_{ik}[a_k(f_i)f_i], \tag{20}$$

where $a_k(f_i)$ is the k-th element of $\boldsymbol{a}^{\boldsymbol{T}}(f_i)$. Eq. (20) indicates that the product of a continuous feature with its continuous weight can be approximated as a sum of the products of K transformed features in the form of $a_k(f_i)f_i$ with the corresponding K single-valued weights. The conditional probability in the HCRF-DC model can thus be written as

$$p(w|o;\boldsymbol{\lambda}) = \frac{1}{z(o;\boldsymbol{\lambda})} \sum_{s\in w} exp\left(\sum_{i=1}^{C}\sum_{k=1}^{K}\lambda_{ik}f_{ik} + \sum_{j=1}^{B}\lambda_j f_j\right), \tag{21}$$

where

$$f_{ik} = a_k(f_i)f_i \tag{22}$$

only depends on the continuous-valued feature f_i and the locations of the knots, and is independent of the weights to be estimated.

Note that although Eq. (21) has the same form as Eq. (8), the definition of f_{ik} is quite different. The expanded features defined in Eq. (22) are more robust and typically performs much better than that defined in Eq. (7). In fact, spline interpolation can be considered as a filter with which the data sparseness problem can be alleviated since the difference between the model-estimated value and the observed value can be considered as the observation noise. A trade-off can thus be obtained between the accuracy of the constraint and the uncertainty of the constraint, or the discrimination ability and the generalization ability, by choosing the number of knots used: Using a small number of knots may represent the constraints with low accuracy, and as a result, reduces classification accuracy. On the other hand, increasing the number of knots forces the model obtained to follow increasingly closely to the distribution observed in the training data and may decrease the generalization ability of the model.

Eq. (21) is in the standard log-linear form and can be efficiently solved with existing algorithms for the HCRF-MC model. More specifically, we have used the RPROP [7] algorithm in our experiments since it has been shown to perform the best for the HCRF-MC model [5].

Figure 1 and Figure 2 illustrate the difference between the moment constraint and the distribution constraint. Figure 1 shows five features with the same first-order moment but different distributions. Note that although the feature distributions are very different, the model with moment constraints does not distinguish and use the distribution information. Figure 2 shows the associated weights learned using the moment constraints and the distribution constraints. With the moment constraint, the same weight (the straight line $\lambda 0$) is used for all the values of the same feature. In other words, the feature is considered having the same importance across all its value range. With the distribution constraint, the weight is a function of the feature and may have very different functional shape as the curves $\lambda 1$- $\lambda 5$ shown in Figure 2. As the result, a feature can be important within one range and less important within another range.

As observed in the HCRF-MC model [2][5], normalizing the features to the same range can typically provide better results in practical if the dynamic range of the features is vastly different, even though normalization should cause no difference in theory. This is especially true if higher-order statistics are used as features. The reason is that most optimization algorithms work best within its designed parameter range. If the optimal value is outside of the assumed range, the performance of the algorithms drastically decreases either as a speed slowdown or as an error rate increase. For example, an RPROP algorithm may have a minimum step size optimized for the designed parameter range. If a feature has a huge value, the associated optimal weight is extremely small and so the pre-set minimum step is too large to allow for a search of the optimal weight. One possible solution is to reduce the minimum step so that the optimal weight can be found. However, a smaller minimum step usually means a slower search process and may dramatically increase the training time. Two typical normalization approaches are mapping the features f in the range of $[l\ h]$ into the range of $[1\ 2]$ with $f^{'} = (f + h - 2l)/(h - l)$ and normalizing the variance of the features to unity.

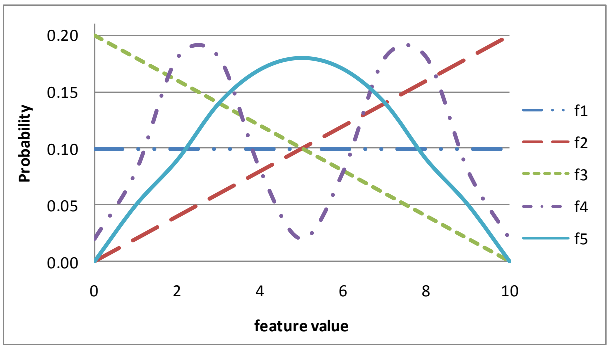

Figure 1: *Features with the same first-order moment and different distributions.*

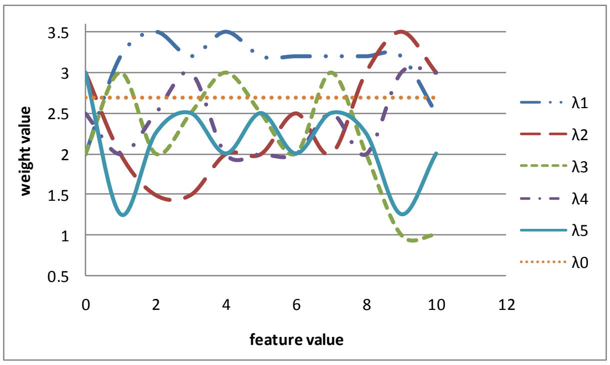

Figure 2: *Comparison between the weights obtained using the moment constraint and the distribution constraint.*

3. Experimental Results

We have evaluated the HCRF-DC model by comparing it with the HCRF-MC model on the TIMIT phone classification task. In this task, the boundaries of a segment are known and the goal is to identify the phone name of each segment. We followed the standard practice [4] of mapping the 61 TIMIT phones into 48 phones for model training, and collapsing the 48 phones to 39 phones for evaluation. All parameters used in the HCRF-MC and HCRF-DC training algorithms such as stopping point, step size, and number of knots were tuned on the MIT development set [3]. The best model parameters discovered were then used on the NIST core evaluation set.

The training, development, and evaluation sets contain 142,910, 15,334, and 7,333 phonetic segments respectively.

The mean and variance normalized 39-dimensional acoustic observations used in the experiments contain the 13-dimensional Mel-frequency cepstral coefficient (MFCC) and its first and second derivatives. To make it easier to process with HTK, the segment boundaries given in the corpus were adjusted to coincide with per-utterance segment boundaries following the practice in [2][5]. The model estimation problems of the HCRF models are non-convex and so proper initialization is crucial. In our experiments, both the HCRF-MC and HCRF-DC models were initialized from the same three-state left to right mono-phone HMM system trained with maximum likelihood (ML) criterion. The HMM system is then converted into the HCRF model based on the procedure described in [2]. Note that during the conversion, each state-Gaussian pair in the HMM system is mapped to a state in the HCRF model. With 10, 15, and 20 Gaussian components per state in the HMM system, the resulting HCRF models contain 30, 45, and 60 states per phone. The RPROP [7] algorithm was used to train the HCRF models with no additional regularizations such as those described in [9].

Table 1. *TIMIT phone classification error rate achieved with the HCRF-MC and HCRF-DC models on the MIT development set and the NIST core test set.*

# Gaussian Mixtures	HCRF-MC Dev	HCRF-MC Core	HCRF-DC Dev	HCRF-DC Core
20	19.8%	**21.3%**	19.4%	**20.8%**
15	20.3%	21.4%	19.9%	21.0%
10	20.4%	21.7%	20.3%	21.4%

Table 1 shows phone classification error rate (CER) on the MIT development set and the NIST core test set achieved with the HCRF-MC and HCRF-DC models initialized with HMM systems that contain 10, 15, and 20 Gaussian components per state. We observe from the table that the HCRF-DC model outperforms the HCRF-MC model consistently over different settings. The best result of 20.8% CER achieved with the HCRF-DC model is better than the best HCRF-MC model result of 21.3% [5], the best HMM maximum mutual information estimation (MMIE) result of 24.9%, and the best HMM ML result of 25.8 by absolute CER reduction of 0.5%, 4.1%, and 5.0% respectively. To our best knowledge, this result is also 0.3% better than the best single-system result of 21.1% reported on this task using the same features and was achieved with the large-margin technique [8]. All these improvements are statistically significant at the 5% significance level.

4. Conclusions

We have developed an HCRF model with distribution constraints. We showed that in the HCRF-DC model, the conditional probability is no longer in the simple log-linear form where each weight to be estimated is a single, constant value. Instead, the weights in the HCRF-DC model are non-linear functions of the features. We provided a solution to this optimization problem by converting it to a log-linear problem at a higher-dimensional space using cubic splines and demonstrated the effectiveness of the HCRF-DC model on the TIMIT phone classification task.

We believe the HCRF-DC model can be applied to other tasks where the HCRF-MC model has been successfully used to achieve improved performance. In addition, by incorporating additional (e.g., long-range [10][14][15]) features and introducing additional regularization terms the performance can be further improved.

5. Acknowledgements

We would like to thank Dr. Asela Gunawardana and Milind Mahajan at Microsoft research for their valuable discussions and useful helps in conducting experiments.

6. References

[1] El-Khoribi, R. A., “Hidden Conditional Random Fields for ECG Classification”, ICGST-AIML Journal, vol. 8, no. III, December 2008, pp. 25-30.

[2] Gunawardana, A., Mahajan, M., Acero, A. and Platt, J. C., “Hidden Conditional Random Fields for Phone Classification”, in Proc. of Interspeech 2005, pp. 1117—1120.

[3] Halberstadt, A. K. and Glass, J. R., “Heterogeneous acoustic measurements for phonetic classification,” in Proc. of Eurospeech 1997, pp. 401–404.

[4] Lee, K. F. and Hon, H. W., “Speaker Independent Phone Recognition Using Hidden Markov Models,” in Proc. of ICASSP 1980, pp. 1641–1648.

[5] Mahajan, M., Gunawardana, A. and Acero, A., “Training Algorithms for Hidden Conditional Random Fields’, in Proc. of ICASSP 2006, vol. I, pp. 273 – 276.

[6] Reiter, S., Schuller, B. and Rigoll, G., “Hidden Conditional Random Fields for Meeting Segmentation”, in Proc. of ICME 2007, pp. 639-642.

[7] Riedmiller, M. and Braun, H., “A direct adaptive method for faster back-propagation learning: The RPROP algorithm”, in proc. of IEEE ICNN 1993, vol. 1, pp. 586-591.

[8] Sha, F. and Saul, L.K., “Large Margin Gaussian Mixture Modeling for Phonetic Classification and Recognition,” in Proc. of ICASSP 2006, vol. I, pp. 265 – 268.

[9] Sung, Y.-H., Boulis, B., Manning, C. and Jurafsky D., “Regularization, Adaptation, and Non-independent Features Improve Hidden Conditional Random Fields for Phone Classification”, in proc. of ASRU workshop, pp. 347-352.

[10] Deng, L., Li, X., Yu, D., and Acero, A., "A Hidden Trajectory Model with Bi-directional Target-Filtering: Cascaded vs. Integrated Implementation for Phonetic Recognition", in proc. ICASSP 2005.

[11] Yu, D., Deng, L., Gong, Y. and Acero, A., “Discriminative Training of Variable-Parameter HMMs for Noise Robust Speech Recognition”, in Proc. of Interspeech 2008, vol. I, pp. 285-288.

[12] Yu, D., Deng, L., Gong, Y. and Acero, A., “A Novel Framework and Training Algorithm for Variable-Parameter Hidden Markov Models”, IEEE trans. on Audio, Speech, and Language Processing (to appear).

[13] Yu, D., Deng, L. and Acero, A., "Using Continuous Features in the Maximum Entropy Model", Pattern Recognition Letters (to appear).

[14] Yu, D., Deng, L. and Acero, A., "Structured Speech Modeling", IEEE Trans. on Audio, Speech and Language Processing. Vol. 14 No. 5, Sep 2006. pp. 1492- 1504.

[15] Yu, D., Deng, L. and Acero, A., “Evaluation of a Long-contextual-span Hidden Trajectory Model and Phonetic Recognizer Using A* Lattice Search,” in Proc. of Interspeech, 2005, pp. 553-556.