Comparison of Smoothing Techniques for Robust Context Dependent Acoustic Modelling in Hybrid NN/HMM Systems

Guangsen WANG, Khe Chai SIM

School of Computing, National University of Singapore
COM1, 13 Computing Drive, Singapore 117417
wangguangsen@comp.nus.edu.sg, simkc@comp.nus.edu.sg

Abstract

Hybrid Neural Network/Hidden Markov Model (NN/HMM) systems have been found to yield high quality phone recognition performance. One issue with modelling the Context Dependent (CD) NN/HMM is the robust estimation of the NN parameters to reliably predict the large number of CD state posteriors. Previously, factorization based on conditional probabilities has been commonly adopted to circumvent this problem. This paper proposes two factorization schemes based on the product-of-expert framework, depending on the choice of the *experts*. In addition, smoothing and interpolation schemes were introduced to improve robustness. Experimental results on the WSJCAM0 reveal that the proposed CD NN/HMM parameter estimation techniques achieved consistent improvement compared to CI hybrid systems. The best hybrid system achieves a 21.7% relative phone error rate reduction and a 17.6% word error reduction compared to a discriminative trained context dependent triphone GMM/HMM system.

Index Terms: context dependent acoustic modelling, smoothing, hybrid system, discriminative training

1. Introduction

Gaussian Mixture Models (GMMs) are commonly used to represent the state output distributions in GMM/HMM systems. Discriminative training methods such as Maximum Mutual Information (MMI) [1] and Minimum Phone Error (MPE) [1] have been shown to yield superior performance compared to the conventional Maximum Likelihood (ML) estimation. As a discriminative model, Neural Network (NN) was proposed as an alternative to GMMs for acoustic modelling. In a hybrid NN/HMM system [2], an NN is used to predict the HMM state posterior probabilities, the "scaled likelihood" is used to model the HMM state emission probabilities. NNs offer several advantages which make them particularly attractive for ASR: 1) They naturally accommodate discriminative training; 2) They can incorporate multiple constraints and information sources; 3) The flexible architecture of NNs allows them to easily incorporate contextual inputs.

Context dependent (CD) modelling of phonemes is widely adopted in state-of-the-art large vocabulary continuous speech recognition (LVCSR) systems. Many research works have shown that hybrid systems can also be trained to take advantage of CD modelling to further boost the system performance [2, 3, 4]. However, directly predicting all CD state posteriors leads to an NN with a huge number of outputs. Both efficient computation and robust estimation of the model parameters will become issues. Therefore, factorization to smaller networks based on conditional probabilities is usually applied to circumvent this problem [2, 3].

Although a lot of research has been devoted to CD modelling for hybrid NN/HMM systems, few have investigated on how different factorization techniques and smoothing schemes will affect the system performance. This paper presents two factorization approaches, namely Context Dependent State Discriminator (CD-SD) and State Dependent Context Discriminator (SD-CD). In both cases, a single context independent NN was used to first obtain the CI state posteriors, a set of 2-layer NNs were built to non-linearly transform the CI log posteriors to context dependent state posteriors. This leads to a product-of-expert (PoE) mapping where each CI posterior is considered as an *expert*. As such, two factorization schemes can be achieved depending on the target of the PoE mapping. For CD-SD, a set of Context Dependent (CD) NNs were built to discriminate the CI states, one for each context cluster. On the other hand, SD-CD comprises a set of State Dependent (SD) NNs to discriminate context clusters, one for each CI state. The context clusters were obtained using the conventional decision tree clustering technique [5]. To ensure robustness of the context dependent state posterior estimation, three smoothing techniques are applied: probability scaling based smoothing, interpolation with canonical state posteriors, back-off smoothing with Dirichlet prior.

The rest of the paper is organized as follows. Section 2 formulates two factorization approaches for CD NN/HMM. Section 3 describes the product-of-expert framework for transforming CI posteriors into CD posteriors using NNs. The smoothing and interpolation schemes are discussed in section 4. Experimental results are reported in section 5. Finally section 6 summarizes the findings of the work and concludes the paper.

2. Factorization for CD NN/HMM systems

Suppose an n-state left-to-right topology HMM is used to model the phonemes. In a CD HMM, every state is associated with a specific phone, state and context cluster. During recognition, the likelihood of a feature vector o_t given the state s_j of the monophone m_i in the context cluster c_k, $p(o_t|m_i, s_j, c_k)$, is required for each HMM state. It has been shown that the outputs of an NN are the estimates of *a posteriori* probabilities [2]. By applying the Bayes' rule and factoring the conditional probabilities, the required likelihood can be computed in terms of the *posterior* probabilities estimated by NNs in a discriminative fashion. The schematic diagram for the factorization is shown in Fig. 1. There are two stages in this factorization framework. The first stage is the extraction of the CI state log posteriors by a single CI NN, the log posteriors are used as the features to train the NN set in the following stage, hence this CI NN is referred to as FTR-NN. The second stage uses a set of NNs to

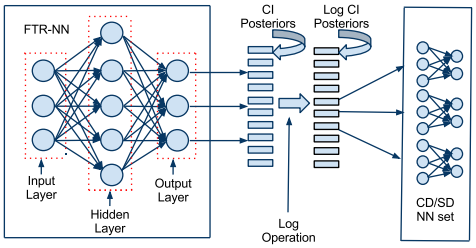

Figure 1: Schematic diagram of factorization for context dependent hybrid NN/HMM system.

non-linearly transform the log posterior probabilities of the CI states to CD/SD state posteriors. Such transform can be viewed as a PoE model [6]. By applying the Bayes' rule, we have the scaled likelihood:

$$\frac{p(o_t|m_i,s_j,c_k)}{p(o_t)} = \frac{P(m_i,s_j,c_k|o_t)}{P(m_i,s_j,c_k)} \quad (1)$$

Note that the denominator term corresponds to the prior probabilities which can be estimated by relative frequencies. The frame probability $p(o_t)$ can be dropped, since it is independent of model parameters. The posterior term in the numerator can be estimated by NNs using various factorization schemes. In [3], factorization is implemented as a hierarchical mixtures of experts (HME) [7] model. The CD posterior is factorized as a 3-level hierarchical experts by $P(c_k|m_i,s_j,o_t)\times P(m_i|s_j,o_t)\times P(s_j|o_t)$ and all experts are trained using the same acoustic features. In this paper, two factorization schemes based on the 2-level hierarchy (see Fig. 1) are proposed. In both schemes, the second stage comprises a set of CD/SD NNs trained on the CI *log* posterior features:

$$\mathbf{f}_t = \{f_{ij}(o_t) = \log P(m_i,s_j|o_t) : \forall i,j\} \quad (2)$$

2.1. Context dependent state discriminator (CD-SD)

A *global* phonetic decision tree is built to obtain the triphone context clusters for CD-SD. Given every triphone context cluster, a CD NN is trained to discriminate all CI states. Therefore, the numerator posterior in Eqn. 1 is factorized as:

$$P(m_i,s_j,c_k|o_t) = P(c_k|o_t)P(m_i,s_j|c_k,\mathbf{f}_t) \quad (3)$$

$P(c_k|o_t)$ is the output of the CI context discriminator (CI NN). $P(m_i,s_j|c_k,\mathbf{f}_t)$ is generated by a set of CD NNs, conditioned on c_k to discriminate the CI states. The inputs to this CD-SD are the CI log posteriors generated by the FTR-NN given in Eqn. 2. The product of these two posteriors is then divided by the denominator prior term in Eqn. 1, to form an estimation of the scaled observation likelihood for recognition.

2.2. State dependent context discriminator (SD-CD)

Instead of using a global decision tree, one decision tree is built for each monophone state in SD-CD. The SD-CD factorization can be expressed as:

$$P(m_i,s_j,c_k|o_t) = P(m_i,s_j|o_t)P(c_k|m_i,s_j,\mathbf{f}_t) \quad (4)$$

$P(m_i,s_j|o_t)$ is estimated by a single CI NN to discriminate the CI states. $P(c_k|m_i,s_j,\mathbf{f}_t)$ is the posterior of the context cluster, c_k, given the CI state, $\{m_i,s_j\}$ [8]. This posterior can also be estimated by a set of CI state dependent NNs (SD NNs) to discriminate the context clusters, c_k. Note that the inputs to these SD NNs are also the *log* CI state posteriors given in Eqn. 2.

3. Multiple PoE transforms

For both factorization schemes, a CD/SD NN set is used to predict CD/SD posteriors. These NNs are usually trained directly from acoustic features, e.g. MFCCs except that only training frames belonging to the clusters they represent are used during training. In this paper, instead of training CD/SD NNs with the MFCC feature like CI NN, CD/SD NNs are viewed as non-linear transforms of CI posteriors and these transforms are given by feeding the CI log posterior probabilities through a set of 2-layer NNs. Such transformations fit within the product-of-expert framework [6], where each CI posterior is regarded as an *expert*. Hence, the conditional posterior probability is given by the following PoE expression:

$$\begin{aligned} P(c_k|m_i,s_j,\mathbf{f}_t) &= \frac{1}{Z}\prod_{\forall i,j} P(m_i,s_j|o_t)^{-\omega_{kij}} \\ &= \frac{\exp\left\{-\sum_{\forall i,j}\omega_{kij}f_{ij}(o_t)\right\}}{\sum_{k'=1}^{K}\exp\left\{-\sum_{\forall i,j}\omega_{k'ij}f_{ij}(o_t)\right\}} \end{aligned}$$

where $Z=\sum_{k=1}^{K}\prod_{\forall i,j}P(m_i,s_j|o_t)^{-\omega_{kij}}$ is the normalizing constant. K denotes the total number of context clusters (*i.e.* the output dimension of the CD/SD NNs). ω_{kij} is the weight of posterior $P(m_i,s_j|o_t)$ for predicting the k-th context in the CD/SD NNs. The resulting form is a softmax function of a linear transformation of $\mathbf{f}_t$, which can be conveniently represented as a 2-layer NN with input $\mathbf{f}_t$ and a softmax output activation. In fact, the proposed PoE CD NN/HMM system is essentially an instance of canonical state model [9]. Different from [9] where a mixture of Constrained Maximum Likelihood Linear Regression (CMLLR) transforms are used to map the canonical states to the context dependent states, in our PoE framework, FTR-NN is the canonical state model and the CD/SD NNs are the discriminatively learned non-linear transforms for modeling context dependent states.

4. Robust estimation of context dependent state posteriors

In both factorizations, scaled likelihood can be computed by evaluating CI NN followed by the corresponding CD/SD NNs. Therefore, two sets of posteriors are produced for every observation frame. For a robust estimation of the conditional posteriors, a probability scaling based smoothing approach is introduced. Besides, an interpolation between the smoothed context dependent state posteriors and the canonical state posteriors can also be applied. To handle the possible data sparsity problem, yet another smoothing using Dirichlet prior is employed.

4.1. Probability scaling based smoothing

For CD-SD, the context dependent posterior $P(m_i,s_j|c_k,\mathbf{f}_t)$ produced by the CD NN is smoothed with the context discriminator posterior $P(c_k|o_t)$. On the other hand, the SD NN posterior $P(c_k|m_i,s_j,\mathbf{f}_t)$ under SD-CD is smoothed by the posterior of the monophone state $\{m_i,s_j\}$ which c_k is conditioned on. The smoothing is expressed as:

$$\mathcal{P}^{\text{CD}} = \begin{cases} P(m_i,s_j|c_k,\mathbf{f}_t)^{\alpha}\times P(c_k|o_t)^{1-\alpha} \\ P(c_k|m_i,s_j,\mathbf{f}_t)^{\alpha}\times P(m_i,s_j|o_t)^{1-\alpha} \end{cases} \quad (5)$$

where α ($0\le\alpha\le 1$) is the smooth factor. Note when α equals 0.5, $\mathcal{P}^{\text{CD}}$ becomes the *geometric mean* of $P(m_i,s_j|c_k,\mathbf{f}_t)$ and

$P(c_k|o_t)$ for CD-SD, $P(c_k|m_i, s_j, \mathbf{f}_t)$ and $P(m_i, s_j|o_t)$ for SD-CD. These are essentially the same as the posterior terms in Eqn. 3 and Eqn. 4. Therefore, this smoothing is referred to as Weighted Geometric Smoothing (WGS). Probability scaling based smoothing for SD-CD is widely used in the NN/HMM literature [2, 3], the main difference here is that the *conditional* posterior $P(c_k|m_i, s_j, \mathbf{f}_t)$ is estimated under the PoE framework instead of training a dedicated NN using acoustic features from scratch.

4.2. Interpolation with canonical state posteriors

Recall that under the PoE framework, FTR-NN is the canonical state model and the CD/SD NNs are the discriminatively learned non-linear transforms for context dependent state modeling. Therefore, it is interesting to smooth the context dependent state posterior $\mathcal{P}^{\text{CD}}$ further with the canonical state posterior $P(m_i, s_j|o_t)$ through an interpolation:

$$\mathcal{P}^{\text{INT}} = \beta \times \mathcal{P}^{\text{CD}} + (1 - \beta) \times P(m_i, s_j|o_t) \quad (6)$$

where β ($0 \leq \beta \leq 1$) is the interpolation factor which adjusts the weights of the context dependent and canonical state posteriors. The probability after interpolation $\mathcal{P}^{\text{INT}}$ is used in Eqn. 1 as the posterior term to compute the scaled likelihood.

4.3. Back-off smoothing with Dirichlet prior

Under both factorization schemes, CD/SD NNs are trained using only the frames belonging to the context clusters they represent. Therefore, data sparsity problem may exist especially when there are large number of contexts. Dirichlet prior is a Bayesian justified smoothing method for multinomial distribution widely used in language modelling [10]. Here it is used for robust context dependent acoustic modelling under the PoE framework. For Dirichlet smoothing, each CD/SD NN context is parameterized with an additional prior trained using all the training frames. The prior is expressed as:

$$p_{cs} = P(m_i, s_j|o_t) \quad \text{for CD-SD} \quad (7)$$

$$p_{sc} = P(c_k|o_t) \quad \text{for SD-CD} \quad (8)$$

Therefore, the context dependent posteriors after Dirichlet smoothing $\mathcal{P}^{\text{DIR}}$ for the two factorizations are defined as:

$$\begin{cases} [\beta p_{cs} + (1 - \beta) P(m_i, s_j|c_k, \mathbf{f}_t)]^{\alpha} P(c_k|o_t)^{1-\alpha} \\ [\beta p_{sc} + (1 - \beta) P(c_k|m_i, s_j, \mathbf{f}_t)]^{\alpha} P(m_i, s_j|o_t)^{1-\alpha} \end{cases} \quad (9)$$

Where β is the Dirichlet smoothing factor which determines how heavily the context dependent posteriors should rely on their priors, α is the smoothing factor between the CI and CD posteriors similar to Eqn. 5.

5. Experiments

5.1. Experimental setup

This section presents the experimental results for both phone and word recognition on WSJCAM0 corpus [11]. There are 18.3 hours of training data, comprising 9889 utterances. The 5k WSJ0 tasks are used for performance evaluation. The "si_dt5a" set is used as the development set to tune the smoothing and interpolation factors; "si_dt5b" is used as the test set for evaluation purpose. The phone set has 41 monophones including one silence model "sil" and one short pause model "sp". The features for CI NN and FTR-NN are the standard 39-dimensional MFCC which consists of 13 static coefficients (12 MFCC plus one C0 energy term) and its first and second derivatives. HTK[1] was adopted for decoding and NN models were trained with QuickNet[2]. Word recognition was performed using a bigram full decoding followed by a trigram rescoring.

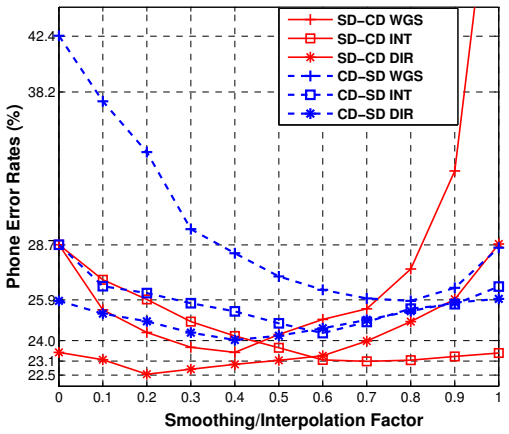

Figure 2: Effects of smoothing and interpolation schemes

The FTR-NN ($585 \times 2000 \times 120$) was trained to extract the log CI state posteriors $\mathbf{f}_t$. In order to incorporate possible acoustic contextual information between successive frames, each frame is concatenated with 7 frames on its left and right neighbour to form a 15-frame input window with 585 units for the input layer. The size of the hidden layer was chosen to optimize the frame accuracy on the development set. The output layer corresponds to the posterior probabilities of 120 physical CI monophone states (40 monophones with 3 states each; "sp" comprises one emitting state which is tied with the centre state of "sil"). This NN is the same as the CI NN of SD-CD. For CD-SD, the CI NN ($585 \times 2000 \times 51$) is trained to discriminate 51 context clusters.

All CD/SD NNs are trained as non-linear transforms of FTR-NN posteriors with a 2-layer topology. The log posteriors generated by FTR-NN are used as the features to train the CD/SD NNs. Therefore, CD/SD NNs have an input layer of 120 units corresponding to 120 monophone states. The CD NN set for CD-SD has 51 elements corresponding to the 51 clusters. Each CD NN is trained to discriminate all 120 monophone states given one context cluster, hence, the output layer has 120 units. On the other hand, the SD NN set for SD-CD has 117 NNs, each corresponds to one of 117 monophone states excluding "sil" states since they are modelled without contexts. Each SD NN is trained to predict the conditional posteriors for all the 50 triphone state contexts given one monophone state. Therefore, they all have 50 output units.

5.2. Investigation of smoothing and interpolation schemes

In this section, we will investigate how smoothing and interpolation would affect the system performance in terms of Phone Error Rates (PERs) on "dt5a". The PERs performance for SD-CD and CD-SD under three smoothing schemes is shown in Fig. 2: Weighted Geometric Smoothing (WGS), Interpolation (INT) and Dirichlet smoothing (DIR). Note that y-axis is in log scale, the solid lines and dotted lines correspond to the performance of SD-CD and CD-SD respectively. The FTR-NN PER (28.71%) is chosen as the baseline ($\alpha = 0$ in Fig. 2). SD-CD consistently outperforms CD-SD for all three smoothing

[1] Hidden Markov Model Toolkit, http://htk.eng.cam.ac.uk
[2] QuickNet, http://www.icsi.berkeley.edu/Speech/qn.html

schemes. For SD-CD (CD-SD), as α goes towards 1.0 (0.0) (i.e., the system behaves as a pure CD (CI) system), the PERs performance drops sharply under WGS. However, as CI posteriors are used to smooth the CD conditional posteriors, both systems can finally achieve their best WGS performance: 25.85% for CD-SD and 23.48% for SD-CD. This clearly illustrates the importance of the smoothing scheme. The best WGS systems for both SD-CD and CD-SD are chosen to produce the context dependent state posteriors $\mathcal{P}^{\mathtt{CD}}$ in Eqn. 6, these posteriors are then smoothed with the canonical state posteriors discussed in section 4.2 through an interpolation. While for the Dirichlet smoothing, the WGS smoothing factors which yield the best PERs performance for both CD-SD and SD-CD are chosen as the α values to compute $\mathcal{P}^{\mathtt{DIR}}$ in Eqn. 9, the tuning of Dirichlet smoothing factor β is shown in Fig. 2. Compared to WGS, INT leads to a further PER performance boost: 24.34% for CD-SD and 23.09% for SD-CD. The best performance is obtained with DIR smoothing: 24.03% for CD-SD and 22.54% for SD-CD.

5.3. Experimental results

Phone Error Rates (PERs) and Word Error Rates (WERs) on development set "si_dt5a" and test set "si_dt5b" of different system configurations are listed in Table 1. The parameter size

Table 1: PER (%) and WER (%) for two factorization schemes.

System		PER		WER	
		dt5a	dt5b	dt5a	dt5b
GMM-ML		30.95	32.10	11.88	13.07
GMM-MMI		29.57	30.38	10.06	10.87
CI NN		28.71	29.21	12.41	13.87
CD-SD	WGS	25.85	27.27	11.13	12.08
	INT	24.34	26.01	10.13	10.75
	DIR	24.03	25.74	9.89	10.37
SD-CD	WGS	23.48	25.06	8.22	9.74
	INT	23.09	24.03	7.83	9.24
	DIR	22.54	23.78	7.48	8.96

of the context dependent hybrid system is approximately 2.1 million. For a comparison with the standard GMM/HMM system, a triphone system with 6 components per state is build with roughly the same parameter size. The baseline GMM/HMM system is trained using both ML (GMM-ML) and MMI (GMM-MMI) criteria. The baseline GMM-ML achieved PER of 32.10% and WER of 13.07% on "dt5b". The MMI baseline GMM-MMI achieved PER of 30.38% and WER of 10.87%.

Results from Table 1 show that the hybrid systems under both factorizations have increased the system performance compared with the CI NN system for both phone recognition and word recognition. Moreover, SD-CD outperforms CD-SD consistently, DIR smoothing yields the best results for both CD-SD and SD-CD. Compared with the GMM/HMM system, SD-CD under all three smoothing schemes outperforms both GMM-ML and GMM-MMI, CD-SD has a comparable performance with GMM-MMI under INT and DIR smoothing. The best system SD-CD with DIR smoothing achieved PER of 23.78% and WER of 8.96% on the testing set "dt5b". These translate to a 21.7% relative phone error rate reduction and a 17.6% word error rate reduction compared to GMM-MMI. Further significance test using SCTK[3] shows that the WER improvements of SD-CD DIR over the baseline GMM-MMI and SD-CD WGS are significant with a significance value of 0.003 and 0.01 respectively at the significant level of 0.05.

[3]NIST Speech Recognition Scoring Toolkit, http://www.itl.nist.gov/iad/mig//tools

6. Conclusions

In this paper, context dependent acoustic modelling of hybrid NN/HMM systems were investigated under two factorization schemes. In addition, to improve the robustness, three smoothing techniques were applied: probability scaling based smoothing, interpolation with canonical state posterior, backoff smoothing with Dirichlet prior. The factorization is implemented using a cascade of NNs. Moreover, CD/SD NNs are modelled as non-linear transforms of the FTR-NN posteriors under a product-of-expert framework. Experimental results reveal that smoothing has a significant impact on the system performance. Compared to the CI hybrid system, our best system achieves a substantial PERs and WERs improvement. With the same parameter size, the CD hybrid system outperforms significantly both the ML and MMI trained GMM/HMM systems. Also SD-CD gives a better performance than CD-SD. The best performance for both CD-SD and SD-CD is obtained by the Dirichlet prior smoothing. Future work can be expanded by training the NNs using sequence classification criteria, e.g., MMI or MPE, instead of frame-based cross-entropy criterion. Applying discriminative triphone clustering techniques other than the ML based phonetic decision tree method to CD/SD NN training is also an interesting direction.

7. Acknowledgement

This research is done for CSIDM Project No. CSIDM-200806 partially funded by a grant from the National Research Foundation (NRF) administered by the Media Development Authority (MDA) of Singapore.

8. References

[1] D. Povey, *Discriminative training for large vocabulary speech recognition*. PhD thesis, Cambridge University, 2004.

[2] N. Morgan and H. Bourlard, "Neural networks for statistical recognition of continuous speech," *Proceedings of the IEEE*, vol. 83, pp. 742–772, 1995.

[3] J. Fritsch, M. Finke, and A. Waibel, "Context-dependent hybrid HME/HMM speech recognition using polyphone clustering decision trees," in *IEEE., ICASSP*, 1997, pp. 1759–1762.

[4] A. J. Robinson, G. D. Cook, D. P. W. Ellis, E. Fosler-Lussier, S. J. Renals, and D. A. G. Williams, "Connectionist speech recognition of broadcast news," *Speech Commun.*, vol. 37, pp. 27–45, 2002.

[5] S. J. Young, J. J. Odell, and P. C. Woodland, "Tree-based state tying for high accuracy acoustic modelling," in *HLT*, 1994, pp. 307–312.

[6] K. C. Sim, "Discriminative product-of-expert acoustic mapping for cross-lingual phone recognition," in *IEEE., ASRU*, 2009, pp. 546–551.

[7] M. I. Jordan, "Hierarchical mixtures of experts and the EM algorithm," *Neural Computation*, vol. 6, pp. 181–214, 1994.

[8] H. Bourlard, N. Morgan, C. Wooters, and S. Renals, "CDNN: a context dependent neural network for continuous speech recognition," in *IEEE., ICASSP*, 1992, pp. 349–352.

[9] M. J. F. Gales and K. Yu, "Canonical state models for automatic speech recognition," in *Interspeech*, 2010, pp. 58–61.

[10] C. X. Zhai and J. Lafferty, "A study of smoothing methods for language models applied to information retrieval," *ACM Trans. Inf. Syst.*, vol. 22, pp. 179–214, 2004.

[11] T. Robinson, J. Fransen, D. Pye, J. Foote, and S. Renals, "WSJCAM0: A British English speech corpus for large vocabulary continuous speech recognition," in *IEEE., ICASSP*, 1995, pp. 81–84.