Improving Grapheme-based ASR by Probabilistic Lexical Modeling Approach

Ramya Rasipuram[1,2] *and Mathew Magimai.-Doss*[1]

[1] Idiap Research Institute, CH-1920 Martigny, Switzerland
[2] Ecole Polytechnique Fédérale de Lausanne (EPFL), CH-1015 Lausanne, Switzerland
{ramya.rasipuram,mathew}@idiap.ch

Abstract

There is growing interest in using graphemes as subword units, especially in the context of the rapid development of hidden Markov model (HMM) based automatic speech recognition (ASR) system, as it eliminates the need to build a phoneme pronunciation lexicon. However, directly modeling the relationship between acoustic feature observations and grapheme states may not be always trivial. It usually depends upon the grapheme-to-phoneme relationship within the language. This paper builds upon our recent interpretation of Kullback-Leibler divergence based HMM (KL-HMM) as a probabilistic lexical modeling approach to propose a novel grapheme-based ASR approach where, first a set of acoustic units are derived by modeling context-dependent graphemes in the framework of conventional HMM/Gaussian mixture model (HMM/GMM) system, and then the probabilistic relationship between the derived acoustic units and the lexical units representing graphemes is modeled in the framework of KL-HMM. Through experimental studies on English, where the grapheme-to-phoneme relationship is irregular, we show that the proposed grapheme-based ASR approach (without using any phoneme information) can achieve performance comparable to standard phoneme-based ASR approach.

Index Terms: Automatic speech recognition, hidden Markov model, Lexical modeling, Graphemes, Phonemes, Posterior features, Kullback-Leibler divergence based HMM

1. Introduction

State-of-the-art automatic speech recognition (ASR) systems use phonemes as subword units and a pronunciation lexicon is used to map orthographic transcription of words to sequence of phonemes. Thus, one of the primary resource required to build a good ASR system is a well developed phoneme pronunciation lexicon. The development of phoneme lexicon requires some minimum phonetic expertise and is usually a semi-automatic process.

An alternative to phonemes is graphemes[1], which makes lexicon development easy [1, 2, 3], [4, Chapter 4], [5, 6, 7, 8, 9]. However, modeling the relationship between graphemes and standard acoustic feature observations, such as PLP cepstral coefficients which capture phoneme related information (from envelop of short-term spectrum) is not always trivial. The reason being grapheme-to-phoneme relationship depends upon the language. For language such as Spanish the relationship is regular, while for language such as English the relationship is irregular. To overcome the problem of irregular relationship, in literature modeling of context-dependent graphemes has been explored [1, 2, 3]. The implicit assumption here being that relationship between context-independent graphemes and context-independent phonemes can be irregular, but relationship between context-dependent graphemes and context-independent phonemes could be regular[2]. However, in case of English, context-dependent grapheme based ASR systems have been still found to yield considerably lower performance when compared to context-dependent phoneme based ASR systems [1, 2, 3, 5, 6].

Kullback-Leibler divergence based hidden Markov model (KL-HMM) is a recently proposed approach, where phoneme class conditional probabilities estimated by artificial neural networks (ANN) are directly used as feature observations [12]. In a more recent work, we showed that KL-HMM system is a HMM-based ASR system, where the relationship between physical/acoustic states (modeled by ANN) and logical/lexical states (modeled by KL-HMM) is probabilistic [13]. Furthermore, we also showed that KL-HMM approach is equally applicable to both HMM/Gaussian mixture model (HMM/GMM) system and hybrid HMM/ANN system.

Building upon our recent interpretation of KL-HMM system, this paper presents a novel grapheme-based ASR approach in which the probabilistic relationship between lexical units and acoustic states is modeled rather than the deterministic relationship (as in standard HMM-based ASR system). The proposed approach is implemented in two stages. First a crossword trigraph based HMM/GMM system is trained using state tying. The tied HMM states of this system are chosen as the acoustic states and then the probabilistic relationship between context-dependent graphemes and acoustic states is learned through KL-HMM approach. ASR studies conducted on English language show that the performance of grapheme-based ASR can be significantly improved and close to state-of-the-art performance can be achieved. Most importantly, the improvement in ASR accuracy is achieved with traditional trigraph modeling and without any phonetic knowledge.

The rest of the paper is organized as follows. In Section 2, we present the KL-HMM approach. The interpretation of KL-HMM as probabilistic lexical modeling approach originally proposed in [13] is relatively new and is therefore presented again for the sake of clarity in Section 3. Section 4 presents the experimental studies. Finally, in Section 5 we conclude.

2. KL-HMM

In KL-HMM approach [12], first the mapping between acoustic feature observations ($\mathbf{x}_t$) and phonemes ($p_1, \cdots, p_d, \cdots, p_D$, D - number of phoneme classes) is learned through a posterior

[1] Graphemes are alphabets of a language.

[2] Indeed, grapheme-to-phoneme conversion systems exploit this notion by building a decision tree [10] or n-gram model [11].

probability estimator which estimates class conditional probabilities of phonemes $\mathbf{z}_t$, given by,

$$\begin{aligned}\mathbf{z}_t &= [z_t^1, \cdots, z_t^d, \cdots, z_t^D]^T \\ &= [P(p_1|\mathbf{x}_t), \cdots, P(p_d|\mathbf{x}_t), \cdots, P(p_D|\mathbf{x}_t)]^T \end{aligned} \quad (1)$$

We refer to $\mathbf{z}_t$ as *posterior feature*. Generally an MLP is used to estimate posterior features. Later, the soft correspondence between HMM states and phonemes (p_d) is modeled by using posterior features $\mathbf{z}_t$ as feature observations in HMM system. Each HMM state $i \in \{1, \cdots I\}$ in KL-HMM is parameterized by a categorical distribution $\mathbf{y}_i = [y_i^1, \cdots y_i^d, \cdots, y_i^D]^{\mathrm{T}}$. The local score $S(\mathbf{y}_i, \mathbf{z}_t)$ at each HMM state i in case of KL-HMM system is the Kullback Leibler (KL) divergence between $\mathbf{y}_i$ and $\mathbf{z}_t$, i.e.,

$$S(\mathbf{y}_i, \mathbf{z}_t) = \sum_{d=1}^{D} z_t^d \log(\frac{z_t^d}{y_i^d}) \quad (2)$$

The above equation represents the case where $\mathbf{z}_t$ is the reference distribution and the local score is denoted as RKL. However, given that KL-divergence is an asymmetric measure there are other possible ways to estimate the local score [12].

2.1. Training

The KL-HMM acoustic model is fully parameterized by $\Theta = \{\{\mathbf{y}_i\}_{i=1}^I, \{a_{ij}\}_{i,j=1}^I\}$ where each state i is represented by categorical distribution $\mathbf{y}_i$, a_{ij} is the transition probability from state i to state j. Given a training set of N utterances, where each training utterance n is a sequence of posterior features $Z(n) = \{\mathbf{z}_1(n), \cdots, \mathbf{z}_t(n), \cdots, \mathbf{z}_{T(n)}(n)\}$ of length $T(n)$, the parameters Θ are estimated by Viterbi expectation maximization algorithm which minimizes the cost function,

$$\min_{Q \in \mathcal{Q}} \sum_{n=1}^{N} \sum_{t=1}^{T(n)} [S(\mathbf{y}_{q_t}, \mathbf{z}_t(n)) - \log a_{q_{t-1}q_t}] \quad (3)$$

where $\mathbf{z}_t(n) = [z_t^1(n), \cdots, z_t^d(n), \cdots, z_t^D(n)]^T$, $\mathcal{Q}$ denotes the set of possible HMM state sequences. More precisely, the training process involves iteration over the segmentation and the optimization steps until convergence.

2.2. Decoding

The decoding is performed using standard Viterbi decoder. Given a sequence of posterior features $Z = \{\mathbf{z}_1, \cdots, \mathbf{z}_t, \cdots, \mathbf{z}_T\}$ and the trained parameters $\Theta = \{\{\mathbf{y}_i\}_{i=1}^I, \{a_{ij}\}_{i,j=1}^I\}$, decoding involves recognition of the underlying hypothesis $\hat{m}$:

$$\hat{m} = \arg\min_{Q \in \mathcal{Q}} \sum_{t=1}^{T} [S(\mathbf{y}_{q_t}, \mathbf{z}_t) - \log a_{q_{t-1}q_t}] \quad (4)$$

where $\mathcal{Q}$ denotes the set of possible state sequences allowed by the hypothesis m.

3. Probabilistic Lexical Modeling

KL-HMM until now has been investigated as an approach where a posterior probabilities of phonemes can be directly used as feature observations in HMM system [12, 7, 8, 14, 9]. Recently, we showed that the KL-HMM can be seen as probabilistic lexical modeling approach applicable to both HMM/GMM and hybrid HMM/MLP based ASR systems [13]. The interpretation being relatively recent we present here again for the sake of completeness.

3.1. Standard HMM-based ASR

In HMM-based ASR, given the acoustic model, lexicon and language model, finding the most likely word sequence is achieved by finding the most likely state sequence Q^*

$$Q^* = \arg\max_{Q \in \mathcal{Q}} P(Q, X|\Theta) \quad (5)$$

$$\approx \arg\max_{Q \in \mathcal{Q}} \prod_{t=1}^{T} p(\mathbf{x}_t|q_t, \Theta_A) \cdot P(q_t|q_{t-1}, \Theta) \quad (6)$$

$$\approx \arg\max_{Q \in \mathcal{Q}} \sum_{t=1}^{T} \log p(\mathbf{x}_t|q_t, \Theta_A) + \log P(q_t|q_{t-1}, \Theta) \quad (7)$$

where $\mathcal{Q}$ denotes set of all possible HMM state sequences, $Q = \{q_1, \cdots q_t, \cdots q_T\}$ denotes a sequence of HMM states, T denotes number of frames, and $\Theta = \{\Theta_A, \Theta_L\}$ denotes the set of parameters, more specifically acoustic model and lexical model parameters set Θ_A and language model parameters Θ_L. Eqn. (6) results after *i.i.d* and first order Markov assumptions. Usually, $\log p(\mathbf{x}_t|q_t, \Theta_A)$ is referred to as *local emission score* and $\log P(q_t|q_{t-1}, \Theta)$ is referred to as *transition score*.

In HMM/GMM system, the emission likelihood $p(\mathbf{x}_t|q_t, \Theta_A)$ is estimated using GMMs where as in hybrid HMM/ANN system, the emission likelihood is estimated using ANN [15]. Though the literature is dominated by the approach of using likelihood as local emission score, in theory, HMMs can be also trained and decoded using a posteriori probability estimate $P(q_t|x_t, \Theta_A)$ as local emission score [15]. We refer to the approach of using likelihood as local emission score as *likelihood-based* ASR approach and the approach of using a posteriori probability as local emission score as *posterior-based* ASR approach.

In practice, in HMM-based ASR system there are two kinds of states, namely *acoustic states* denoted as q_t^{aco} corresponding to acoustic model and *lexical states* denoted as q_t^{lex} corresponding to lexical model. For example, in a tied-state context-dependent subword based ASR system the clustered or physical states are the acoustic states and the states of the context-dependent subword model (e.g. /k/-/ae/+/t/) also commonly referred to as logical states are the lexical states. Let $\Theta_A = \{\theta_a, \theta_l\}$, where θ_a denotes the parameters of acoustic model and θ_l denotes the parameters of lexical model. The acoustic model parameters in the case of GMMs are the Gaussian means, variance and weights of each acoustic state. In standard HMM-based ASR systems, the relationship between lexical states and acoustic states is one-to-one, i.e. *deterministic*. Thus, θ_l consists of the set of subword units, pronunciation models of words and a table that maps lexical states (corresponding to the subword units) onto acoustic states.

During both training phase and decoding phase, the emission likelihood is estimated by matching the acoustic state evidence with the lexical model. This is trivial as the relationship between the acoustic states and the lexical states is one-to-one. More precisely, given the one-to-one relationship, $p(\mathbf{x}_t|q_t^{lex} = i, \Theta_A) = p(\mathbf{x}_t|q_t^{aco} = d, \theta_a)$, where $i \in \{1, \cdots I\}$ here denotes a lexical state, $d \in \{1, \cdots D\}$ here denotes an acoustic state, I here denotes the number of lexical states and D here denotes the number of acoustic states. Here after, for simplicity we will drop the notations for parameters.

3.2. KL-HMM as Probabilistic Lexical Model

The one-to-one relation between acoustic and lexical states makes the ASR system overly rely on lexical resources, i.e., subword units and pronunciation models. One way to overcome this is to model the probabilistic relationship between lexical and acoustic states.

KL-HMM approach can be viewed as a posterior based ASR approach which replaces the deterministic map between lexical and acoustic states in a standard HMM-based ASR system with a probabilistic map. This is achieved in two steps:

1. first, an acoustic state posterior probability estimator is trained which estimates $\mathbf{z}_t = [P(q_t^{aco} = 1|\mathbf{x}_t) \cdots P(q_t^{aco} = D|\mathbf{x}_t)]^{\mathrm{T}}$.
2. second, a KL-HMM system is trained using $\mathbf{z}_t$ as feature observations. The states of second HMM represent lexical states i.e., context-dependent subword units and are parameterized by $\{\mathbf{y}_i\}_{i=1}^{I}$ which model the probabilistic relation between lexical and acoustic states, i e., $\mathbf{y}_i = [P(q_t^{aco} = 1|q_t^{lex} = i) \cdots P(q_t^{aco} = D|q_t^{lex} = i)]^{\mathrm{T}}$

If $\mathbf{z}_t$ is estimated using ANN then KL-HMM can be seen as probabilistic lexical modeling applied to hybrid HMM/ANN system [7], where as if $\mathbf{z}_t$ is estimated using GMMs of clustered HMM states then it can be seen as probabilistic lexical modeling applied to HMM/GMM system [13]. Furthermore, compared to standard ASR system which uses deterministic lexical model, KL-HMM system does not changes the acoustic model complexity but only the lexical model complexity, where now θ_l consists of now consists of subword unit set, pronunciation model of words and $\{\mathbf{y}_i\}_{i=1}^{I}$. For relations to other probabilistic lexical modeling approaches and interpretation of previous work on KL-HMM [12, 7, 8, 14] the reader is referred to [13].

In our previous studies, we investigated probabilistic lexical modeling using KL-HMM approach for grapheme lexicon in the framework of hybrid HMM/MLP-based ASR system [7, 8, 9], where acoustic states represent context-independent phonemes and the lexical states represent context-dependent graphemes. The acoustic state probability $\mathbf{z}_t$ estimator was an MLP trained using in-domain or out-of-domain acoustic-phonetic resources. In following section, we show the potential of probabilistic lexical modeling for grapheme-based ASR, where the acoustic states are derived from the data by modeling context-dependent graphemes in the framework of conventional HMM/GMM system.

4. Experimental Setup and Results

In this section, we compare deterministic lexical modeling and probabilistic lexical modeling in the context of both grapheme-based ASR and phoneme-based ASR.

ASR studies are conducted on DARPA Resource Management task [16]. The training set consists of 3'990 utterances spoken by 109 speakers (approximately 3.8 hours speech data). The test set contains 1'200 utterances amounting to 1.1 hours of speech data (formed by combining Feb'89, Oct'89, Feb'91 and Sep'92 test sets). The test set is completely covered by a word pair grammar (perplexity 60) included in the task specification.

The phoneme lexicon was obtained from UNISYN dictionary [17]. About 35 words in phoneme lexicon have more than one pronunciation. The grapheme lexicon was transcribed using 29 context-independent graphemes (which includes symbols - and ', and silence).

4.1. Deterministic Lexical Model based ASR System

We build crossword triphone and trigraph based HMM/GMM systems with decision tree based state tying using HTK toolkit [18]. Each triphone or trigraph is modeled by 3 HMM states. The acoustic feature $\mathbf{x}_t$ is the 39 dimensional PLP cepstral feature. Phoneme-based HMM/GMM system uses phonetic question set where as grapheme-based HMM/GMM system uses singleton question set. State tying resulted in 1611 clustered/acoustic states for phoneme-based system and 1536 clustered/acoustic states for grapheme-based system.

4.2. Probabilistic Lexical Model based ASR System

Given the clustered acoustic state models of the deterministic lexical model based system, training probabilistic lexical model based system using KL-HMM approach involves,

1. estimation of acoustic state posterior feature $\mathbf{z}_t = [z_t^1 \cdots z_t^d \cdots z_t^D]^{\mathrm{T}}$ assuming equal priors for the acoustic states,

$$z_t^d = P(q_t^{aco} = d|\mathbf{x}_t) = \frac{p(\mathbf{x}_t|q_t^{aco} = d)}{\sum_{j=1}^{D} p(\mathbf{x}_t|q_t^{aco} = j)} \quad (8)$$

   where $p(\mathbf{x}_t|q_t^{aco} = d)$ is the likelihood of acoustic state d, and D is the number of acoustic states.
2. and then, estimation of $\mathbf{y}_i$ by KL-HMM approach using RKL as local score.

We train and test crossword triphone and trigraph KL-HMM systems where in both systems similar to HMM/GMM system each subword unit is modeled by 3 states. The state tying is performed using the approach proposed in [14].

4.3. Systems

We build six systems, namely,

1. BASE-PHONE: phoneme-based ASR system with deterministic lexical model, where lexical states are triphones and acoustic states are the 1611 clustered states.
2. BASE-GRAPH: grapheme-based ASR system with deterministic lexical model, where lexical states are trigraphs and acoustic states are the 1536 clustered states.
3. PROB-PHONE: phoneme-based ASR system with probabilistic lexical model, where lexical states are triphones and acoustic states are the 1611 clustered states of the system BASE-PHONE.
4. PROB-GRAPH: grapheme-based ASR system with probabilistic lexical model, where lexical states are trigraphs and acoustic states are the 1536 clustered states of the system BASE-GRAPH.
5. PROB-PHONE-CROSS: phoneme-based ASR system with probabilistic lexical model, where lexical states are triphones but the acoustic states are 1536 clustered states of the system BASE-GRAPH.
6. PROB-GRAPH-CROSS: grapheme-based ASR system with probabilistic lexical model, where lexical states are trigraphs but the acoustic states are 1611 clustered states of the system BASE-PHONE.

System PROB-GRAPH-CROSS is somewhat similar to the grapheme-based ASR system presented in [7], in the sense that both use phoneme information. More precisely, in [7] acoustic

Table 1: word error rate (WER) for different systems

	System	Lexical Model	WER
1	BASE-PHONE	deterministic	4.2
2	PROB-PHONE	probabilistic	2.9
3	BASE-GRAPH	deterministic	6.3
4	PROB-GRAPH	probabilistic	4.3
5	PROB-PHONE-CROSS	probabilistic	3.5
6	PROB-GRAPH-CROSS	probabilistic	4.1

states are context-independent phonemes learned on either in-domain data or out-of-domain data (using an MLP), where as in the system PROB-GRAPH-CROSS the acoustic states are clustered context-dependent phonemes learned on the in-domain data (using GMMs). Furthermore, system PROB-GRAPH derives the acoustic states from context-dependent graphemes, and thus does not use any phoneme information.

4.4. Results

Table 1 presents the ASR performance of different systems in terms of word error rate (WER). It can be observed show that,

- the performance of the system BASE-PHONE with deterministic lexical modeling is comparable to 4.1% WER reported in [19]. The performance of the system PROB-PHONE with KL-HMM approach is better than the system BASE-PHONE. Result shows that probabilistic lexical modeling can improve state-of-the-art ASR performance by modeling the pronunciation variation.
- the performance of the system BASE-GRAPH with deterministic lexical modeling is poor compared to the system BASE-PHONE. However, the system PROB-GRAPH improves the performance of grapheme-based ASR and achieves performance close to state-of-the-art phoneme-based ASR system BASE-PHONE.

The systems PROB-PHONE-CROSS and PROB-GRAPH-CROSS were built mainly for analysis purpose. It can be observed that the performance of the system PROB-PHONE-CROSS is worse than the system PROB-PHONE but better than the system BASE-PHONE. Furthermore, it can be seen that system PROB-GRAPH-CROSS improves over the systems PROB-GRAPH, BASE-GRAPH and BASE-PHONE. This indicates that clustered states of the system BASE-PHONE and the system BASE-GRAPH are modeling similar kind of acoustic information, and the poor performance of system BASE-GRAPH is primarily due to the use of deterministic lexical model.

5. Conclusions and Future Work

In this paper, we proposed a novel grapheme-based ASR approach in HMM/GMM framework, in which the probabilistic relationship between context-dependent graphemes (lexical states) and HMM tied states (acoustic states) is modeled. Through experimental studies we showed that by modeling the probabilistic relationship, the approach is more robust against possible pronunciation errors inherent in the grapheme lexicon. The performance of the proposed grapheme-based ASR system can be significantly improved and close to state-of-the-art phoneme-based ASR performance can be achieved without using phoneme information. There is further scope for improving the proposed grapheme-based ASR approach:

- The state clustering i.e., the acoustic states can be improved. For example, using phonetic question set [1, 2], or by enhanced tree clustering [3] which allows parameter sharing across different polygraphs, or by clustering quintgraphs.
- in our previous work in the framework of hybrid HMM/ANN system [7], we observed that the performance of grapheme-based ASR system can be improved by increasing the lexical model complexity without changing acoustic model complexity. We can do similar thing in the present approach, where the lexical states model longer grapheme contexts such as, quintgraphs and the acoustic states represent clustered trigraphs (as done in this paper).

We will scrutinize these possible improvements in our future work in addition to applying the approach to conversational speech and accented speech.

6. Acknowledgements

This work was supported by the Swiss NSF through the grants “Flexible Grapheme-Based Automatic Speech Recognition (FlexASR)” and the National Center of Competence in Research (NCCR) on “Interactive Multimodal Information Management” (www.im2.ch).

7. References

[1] S. Kanthak and H. Ney, “Context-Dependent Acoustic Modeling using Graphemes for Large Vocabulary Speech Recognition,” in *Proc. of ICASSP*, 2002, pp. 845–848.

[2] M. Killer, S. Stüker, and T. Schultz, “Grapheme based Speech Recognition,” in *Proc. of European Conference on Speech Communication and Technology (EUROSPEECH)*, 2003.

[3] B. Mimer, S. Stüker, and T. Schultz, “Flexible Decision Trees for Grapheme based Speech Recognition,” in *Elektronische Sprachsignalverarbeitung, Cottbus, Germany*, 2004.

[4] T. Schultz and K. Kirchhoff, “Multilingual Acoustic Modeling,” in *Multilingual Speech Processing*. Academic Press, 2006.

[5] J. Dines and M. Magimai-Doss, “A Study of Phoneme and Grapheme based Context-Dependent ASR Systems,” in *Proc. of Machine Learning for Multimodal Interaction (MLMI)*, 2007, pp. 215–226.

[6] Y.-H. Sung, T. Hughes, F. Beaufays, and B. Strope, “Revisiting Graphemes with Increasing Amounts of Data,” in *Proc. of ICASSP*, 2009, pp. 4449–4452.

[7] M. Magimai.-Doss, R. Rasipuram, G. Aradilla, and H. Bourlard, “Grapheme-based Automatic Speech Recognition using KL-HMM,” in *Proc. of Interspeech*, 2011, pp. 2693–2696.

[8] D. Imseng, R. Rasipuram, and M. Magimai.-Doss, “Fast and flexible Kullback-Leibler divergence based acoustic modeling for non-native speech recognition,” in *Proc. of ASRU*, 2011, pp. 348–353.

[9] R. Rasipuram, P. Bell, and M. Magimai.-Doss, “Grapheme and Multilingual Posterior Features for Under-Resourced Speech Recognition: A Study on Scottish Gaelic,” in *Proc. of ICASSP*, 2013.

[10] V. Pagel, K. Lenzo, and A. W. Black, “Letter to sound rules for accented lexicon compression,” in *Proceedings of ICSLP*, 1998.

[11] M. Bisani and H. Ney, “Joint-Sequence Models for Grapheme-to-Phoneme Conversion,” *Speech Communication*, vol. 50, pp. 434–451, May 2008.

[12] G. Aradilla, H. Bourlard, and M. Magimai.-Doss, “Using KL-Based Acoustic Models in a Large Vocabulary Recognition Task ,” in *Proc. of Interspeech*, 2008.

[13] R. Rasipuram and M. Magimai.-Doss, “KL-HMM and Probabilistic Lexical Modeling,” http://publications.idiap.ch/downloads/reports/2013/Rasipuram_Idiap-RR-04-2013.pdf, 2013, Idiap Research Report.

[14] D. Imseng *et al.*, "Comparing different acoustic modeling techniques for multilingual boosting," in *Proc. of Interspeech*, 2012.

[15] H. Bourlard and N. Morgan, *Connectionist Speech Recognition - A Hybrid Approach*. Kluwer Academic Publishers, 1994.

[16] P. J. Price *et al.*, "A database of continuous speech recognition in a 1000 word domain," in *Proc. of ICASSP*, 1988.

[17] S. Fitt, "Documentation and User Guide to UNISYN Lexicon and Postlexical Rules," CSTR, University of Edinburgh, Tech. Rep., 2000.

[18] S. Young *et al.*, *The HTK Book (for HTK Version 3.4)*. Cambridge University Engineering Department, UK, 2006.

[19] T. Hain, "Hidden model sequence models for automatic speech recognition," PhD Dissertation, University of Cambridge, 2001.