AN HMM-BASED FORMALISM FOR AUTOMATIC SUBWORD UNIT DERIVATION AND PRONUNCIATION GENERATION

Marzieh Razavi[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
{marzieh.razavi, mathew}@idiap.ch

ABSTRACT

We propose a novel hidden Markov model (HMM) formalism for automatic derivation of subword units and pronunciation generation using only transcribed speech data. In this approach, the subword units are derived from the clustered context-dependent units in a grapheme based system using maximum-likelihood criterion. The subword unit based pronunciations are then learned in the framework of Kullback-Leibler divergence based HMM. The automatic speech recognition (ASR) experiments on WSJ0 English corpus show that the approach leads to 12.7% relative reduction in word error rate compared to grapheme-based system. Our approach can be beneficial in reducing the need for expert knowledge in development of ASR as well as text-to-speech systems.

Index Terms— automatic subword unit derivation, pronunciation generation, hidden Markov model, Kullback-Leibler divergence based hidden Markov model

1. INTRODUCTION

In order to build an automatic speech recognition (ASR) system, typically some expert knowledge is required to define the phonological subword units and the pronunciation lexicon. However, the subword units can be derived automatically from the speech signal which can possibly help in better handling of pronunciation variations [1]. Moreover, with growing interest in development of ASR systems for under-resourced languages, attempts have been made to automatically derive subword units as well as the pronunciations based on such units.

Towards those lines, several approaches have been proposed for automatic derivation of subword units and pronunciation generation. For automatic derivation of subword units, a typical approach in the literature is through segmentation and clustering of speech signal as done in [2, 3]. In another work, a non-parametric Bayesian approach is proposed to jointly learn the segmentation, clustering and subword modeling [4]. Other existing works present a spectral clustering based approach for learning the subword units [5, 6].

For pronunciation generation based on linguistically motivated units, a typical approach is to apply grapheme-to-phoneme (G2P) conversion techniques which are either knowledge-based or data-driven. In the knowledge-based method, the pronunciation rules are derived from linguistic knowledge. In data-driven approaches, G2P relationships are learned from an initial *seed lexicon* that is typically derived from the knowledge-based approach [7, 8, 9, 10]. For generating pronunciations based on automatically derived subword units (ASWUs), however, these techniques cannot be applied as the subword units are not linguistically known and a seed lexicon based on ASWUs is not available. Therefore, several approaches have been proposed for ASWU-based pronunciation generation [3, 11, 12]. A typical approach is to infer pronunciations based on the acoustic evidences from the collected spoken samples for a given word.

There are also approaches which investigate joint determination of subword units and pronunciations. In [13, 14], approaches based on maximum likelihood criterion are proposed. In [15], the authors provide a hierarchical Bayesian model to jointly learn the subword units and pronunciations.

In this paper, we propose a novel hidden Markov model (HMM) based formalism for automatic derivation of subword units and generating pronunciations by assuming only the availability of word-level transcribed speech data (*Section* 2). In this approach, the subword units are derived from the clustered context-dependent graphemes of the HMM/Gaussian mixture model (HMM/GMM) system. To generate pronunciations based on the ASWUs, first the relationship between the graphemes and ASWUs is learned in the framework of grapheme-based Kullback-Leibler divergence based HMM (KL-HMM) [16, 17] through acoustic data. Then using the orthography of each word together with the learned grapheme-to-ASWU relationship, the pronunciation for each word is generated. This is done by employing a recently proposed acoustic G2P conversion approach in the KL-HMM framework [18] (*Section* 3). The experimental results on WSJ0 English corpus show that the proposed approach leads to about 12.7% relative reduction in word error rate (WER) compared to the grapheme-based system (*Sections* 4 and 5).

Compared to the previous approaches, the proposed method for automatic derivation of subword units is fairly simple in the sense that it fits within the HMM framework which is widely known in the community. Moreover, unlike some of the previous approaches [13], our proposed method is not limited to generate pronunciations only for the words that are seen during the training.

2. HMM-BASED FORMALISM

In a standard HMM-based ASR framework, the goal is to find the most likely sequence of words given the acoustic observation sequence $X = [\mathbf{x_1}, \ldots, \mathbf{x_t}, \ldots, \mathbf{x_T}]$ with T denoting the total number of frames. This is achieved by finding the most probable sequence of states Q by assuming an i.i.d distribution and first order Markov model:

$$\arg\max_{Q \in \mathbf{Q}} P(Q, X|\Theta) =$$

$$\arg\max_{Q \in \mathbf{Q}} \prod_{t=1}^{T} p(\mathbf{x_t}|q_t = l^i, \Theta_A) \cdot P(q_t = l^i|q_{t-1} = l^j, \Theta) \quad (1)$$

This work was supported by Hasler foundation through the grant Flexible acoustic data driven grapheme to acoustic unit conversion. The authors would like to thank Ramya Rasipuram for her valuable comments.

where $\Theta = \{\Theta_A, \Theta_L\}$ denotes the set of parameters with Θ_L corresponding to the language model parameter set and Θ_A corresponding to the set of parameters required for modeling the relation between the acoustic observations and the lexical entities. $\mathbf{Q}$ denotes the set of all possible sequences of HMM states and $Q = [q_1, \ldots, q_t, \ldots, q_T]$ indicates the sequence of lexical HMM states corresponding to a word sequence hypothesis. Each state q_t belongs to a set of possible *lexical units* $L = \{l^1, \ldots, l^i, \ldots, l^I\}$ with cardinality of I. The lexical units are based on context-independent (CI) or context-dependent (CD) subword units (e.g. phones/polyphones or graphemes/polygraphmes), or subword unit states.

In order to estimate $p(\mathbf{x_t}|q_t = l^i, \Theta_A)$ which is of our interest, standard HMM-based ASR systems implicitly model the relation between the acoustic feature $\mathbf{x}_t$ and lexical unit l^i as two components through a *latent* variable a^d [19], that is:

$$p(\mathbf{x_t}|q_t = l^i, \Theta_A) = \sum_{d=1}^{D} p(\mathbf{x_t}, a^d|q_t = l^i, \Theta_A) \quad (2)$$

$$= \sum_{d=1}^{D} p(\mathbf{x_t}|a^d, q_t = l^i, \theta_a, \theta_l) \cdot P(a^d|q_t = l^i, \theta_l) \quad (3)$$

$$= \sum_{d=1}^{D} p(\mathbf{x_t}|a^d, \theta_a) \cdot P(a^d|q_t = l^i, \theta_l) \quad (4)$$

assuming the acoustic feature $\mathbf{x}_t$ is independent of lexical unit l^i given a^d (Equation 4).

The relationship between the acoustic features and the latent variables $p(\mathbf{x_t}|a^d, \theta_a)$ is modeled through an *acoustic model* (e.g. Gaussian mixture model) where θ_a denotes acoustic model parameters. We refer to the latent variable a^d as the *acoustic unit* belonging to a set $A = \{a^1, ..., a^d, ..., a^D\}$ with cardinality of D. The relationship between the acoustic and lexical units $P(a^d|q_t = l^i, \theta_l)$ is given by a *lexical model* where θ_l denotes lexical model parameters.

In standard HMM-based ASR systems, the relation between the acoustic and lexical units $P(a^d|q_t = l^i, \theta_l)$ is a one-to-one deterministic map, that is :

$$p(\mathbf{x}_t|q_t = l^i, \Theta_A) = p(\mathbf{x}_t|a^j, \theta_a), \text{ given } l^i \mapsto a^j,\ j \in \{1, \ldots, D\}. \quad (5)$$

2.1. Automatic Subword Unit Derivation

In the CI subword unit based ASR systems, the acoustic units are defined directly from the pronunciation lexicon (i.e. knowledge driven). The relation between the acoustic and lexical units $P(a^d|q_t = l^i, \theta_l)$ is a knowledge-based one-to-one deterministic map. As standard cepstral features tend to model the spectral envelope of the short-term spectrum, which depict characteristics of phones, the CI grapheme-based ASR system performance largely relies on the grapheme-to-phoneme relationship of the language.

For the case of CD subword unit based ASR, the acoustic unit set A is typically derived by clustering the HMM states using decision tree methods in a data-driven manner, i.e., the acoustic units $\{a^d\}_{d=1}^D$ are the clustered context-dependent subword units. The deterministic mapping between the lexical units $\{l^i\}_{i=1}^I$ and acoustic units $\{a^d\}_{d=1}^D$ is learned during the state clustering and tying stage. With CD graphemes as lexical units, it can be observed from Equation (5) that, the likelihood of the training data is primarily maximized by acoustic unit likelihood estimate $p(\mathbf{x_t}|a^j, \theta_a)$. On the other hand, as mentioned earlier, the cepstral features are more related to phones. Therefore, in order to maximize the likelihood on the training data the clustered context-dependent grapheme units $\{a^d\}_{d=1}^D$ need to be more *phone-like*. Thus, in the present paper, we hypothesize that these clustered context-dependent grapheme units, which can be expected to be more phone-like (demonstrated later in Section 5.1.1), can be used as subword units to build ASR systems that are better than standard grapheme-based ASR systems.

2.2. Pronunciation Generation

As explained in Section 1, for generating pronunciations based on ASWUs, conventional G2P conversion approaches cannot be applied as there is no seed lexicon based on ASWUs available. In this paper, we take an alternate approach where first the relationship between the graphemes and the ASWUs is learned through the acoustic data, and then pronunciations for seen and unseen words are inferred. More precisely, this is done in the framework of KL-HMM through a recently proposed acoustic G2P conversion approach [18] where the phones are replaced by ASWUs. We refer to it as G2ASWU conversion approach and explain it in the next section.

3. G2ASWU CONVERSION APPROACH IN KL-HMM

The G2ASWU conversion approach contains two phases as illustrated in Figure 1. In the training phase, a grapheme-based KL-HMM is trained which learns the probabilistic relation between graphemes and ASWUs. In the decoding phase, given the learned relationship along with the orthography of the word, the pronunciation inference is done.

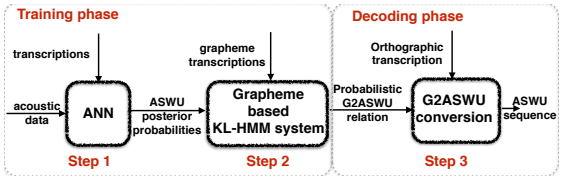

Fig. 1. Block diagram of the G2ASWU conversion approach

3.1. Training Phase

Given the ASWUs $\{a^d\}_{d=1}^D$, KL-HMM uses posterior probabilities of ASWUs $\mathbf{z}_t = [z_t^1, \ldots, z_t^d, \ldots, z_t^D]^{\mathrm{T}}$ with $z_t^d = P(a^d|\mathbf{x}_t)$ as feature observations. Thus, as a first step in the training phase, an artificial neural network (ANN) is trained to estimate the ASWU posterior features $\{\mathbf{z}_t\}_{t=1}^T$. Then as the second step, a grapheme-based KL-HMM is trained [17] in which:

1. The KL-HMM (lexical) states represent CD grapheme states. Each HMM state is parameterized by a categorical distribution $\mathbf{y}_i = [y_i^1, \ldots, y_i^d, \ldots, y_i^D]^{\mathrm{T}}$ with $y_i^d = P(a^d|l^i)$ which models the relationship between the ASWUs $\{a^d\}_{d=1}^D$ and the CD grapheme state l^i.
2. To learn the KL-HMM parameters ($\{\mathbf{y_i}\}_{i=1}^I$), a local score is defined at each state based on the KL-divergence between ASWU posterior feature $\mathbf{z}_t$ and categorical distribution $\mathbf{y}_i$:

$$S_{\mathbf{KL}}(\mathbf{z}_t, \mathbf{y}_i) = \sum_{d=1}^{D} z_t^d \log(\frac{z_t^d}{y_i^d}) \quad (6)$$

3. The parameters are estimated through Viterbi Expectation-Maximization which minimizes a cost function based on KL-divergence local score.
4. For tying KL-HMM (lexical) states, KL-divergence based decision tree state tying method proposed in [20] is applied. The decision tree state tying allows the synthesis of unseen contexts.

3.2. Decoding Phase

As explained in training phase, the KL-HMM parameters capture the probabilistic relationship between graphemes and ASWUs.[1] In the decoding phase, given the orthographic representation of the word together with the parameters of the KL-HMM, the most probable pronunciation is inferred. More precisely, the following steps are done to infer pronunciations as illustrated in Figure 2:

- First, a given word is tokenized into its context-dependent graphemes (Part (***A***)).
- Then, the tokenized context-dependent graphemes and the trained KL-HMM are put together to generate a sequence of ASWU posterior probability vectors (Part (***B***)). As a result of applying the state tying method in the KL-HMM framework, the approach in case of unseen words is capable of generating posterior probability vectors for the unseen grapheme contexts as well.
- Finally, the most probable sequence of ASWUs is inferred by decoding the sequence of ASWU posterior probabilities using an ergodic HMM, i.e., the ASWUs are connected in an ergodic fashion in the HMM (Part (***C***)).

More details about the original acoustic G2P conversion approach are provided in [18].

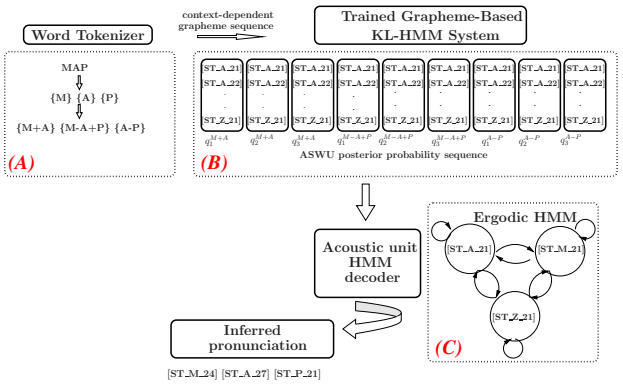

Fig. 2. Illustration of the decoding phase in G2ASWU conversion. As in this study the clustered CD graphemes are derived using the HTK toolkit [22], the ASWUs are represented in the form of HTK clustered states as $[ST\_G\_N]$, where G denotes a mono-grapheme and N denotes a natural number.

4. EXPERIMENTAL SETUP

We evaluated our approach on English using WSJ0 corpus, a 5000 word closed vocabulary task. The training set contains 7106 utterances with about 14 hours of speech and 83 speakers. The test set contains 330 utterances from 8 speakers not seen during training.

4.1. Automatic Subword Unit Derivation

In order to obtain subword units, a cross-word context-dependent grapheme-based HMM/GMM system with a minimum state duration constraint of one was trained using HTK toolkit [22]. The decision tree based clustering was done with singleton questions using maximum likelihood criterion to derive the subword units. Different number of ASWUs were obtained by adjusting the log-likelihood increase during decision-tree based state tying. In this paper we provide the results with ASWUs of size 60, 78 and 90 respectively.

[1]Instead of using the posterior based approach of KL-HMM for learning the probabilistic G2ASWU relations, it may be possible to learn this probabilistic relation in a likelihood based approach such as probabilistic classification of HMM states [21]. This is open for further research.

4.2. Pronunciation Generation

For the pronunciation generation, as the first step, a five-layer multilayer Perceptron (MLP) was trained to classify the ASWUs. We used 39-dimensional PLP cepstral features with four frames preceding context and four frames following context as MLP input. Each hidden layer had 2000 hidden units. For MLP training, about 11% of the training utterances were used for cross-validation. The MLP was trained with output non-linearity of softmax and minimum cross-entropy error criterion, using Quicknet software [23]. As the second step, in the grapheme-based KL-HMM system, context-dependent grapheme subword models were trained using posterior probabilities of ASWUs estimated through the MLP as feature observations. The parameters of the KL-HMM (categorical distributions) were estimated by minimizing a cost function based on KL-divergence local score defined in Equation (6). Each grapheme subword unit was modeled with three HMM states. In the third step, for inferring the pronunciations, each ASWU in the ergodic HMM was modeled with three left-to-right HMM states.

4.3. Evaluation

We evaluated the approach by comparing HMM/GMM systems trained using the ASWUs with an HMM/GMM system trained using graphemes as subword units. In both cases, we trained cross-word context-dependent HMM/GMM systems with 39 dimensional PLP cepstral features extracted using HTK toolkit [22]. Each subword unit was modeled with three HMM states. Each HMM state was modeled by a mixture of 16 Gaussians. For tying the HMM states, only singleton questions were used.

5. EXPERIMENTAL RESULTS AND ANALYSIS

Table 1 provides the results in terms of WER. It can be observed from the results that the HMM/GMM systems using ASWUs perform better than the baseline grapheme-based system. The best performance is achieved with the ASWUs of size 78 which performs significantly better than the grapheme-based system (with 99.5% confidence). As it can be seen from the table, the number of tied states in all of the systems are roughly the same. In fact, the grapheme-based system has the largest number of tied states. So the improvement in the accuracy cannot be attributed to the increase in model complexity.

Unit type	# of units	# of tied states	WER
Grapheme	26	2072	14.2
Automatically-derived	60	2060	13.8
Automatically derived	78	2025	**12.4**
Automatically derived	90	2012	12.7

Table 1. HMM/GMM results with different subword units

The baseline phone-based system in this setup has 8% WER according to [6], so there is still room for further improvement.

5.1. Analysis

This section provides some analysis for the proposed approach.

5.1.1. Comparison to the Phonetic Subword Units

As explained in Section 2, our hypothesis in this paper was that the clustered context-dependent grapheme units are phone-like. In

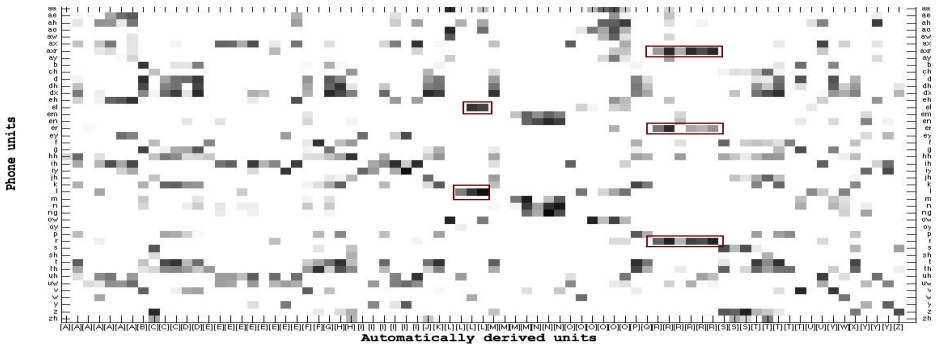

Fig. 3. Relation between the phone units and automatically derived subword units based on the KL-divergence matrix

order to analyze the validity of our hypothesis, we compute the KL-divergence between two Gaussian distributions, one modeling a mono-phone unit and the other modeling an ASWU in the HMM/GMM setup. We compute the KL-divergence between single Gaussians as this is the starting point after state clustering in the HMM/GMM setup. The KL-divergence between the Gaussian $\mathcal{N}_0(\mu_0, \Sigma_0)$ modeling a mono-phone unit as the reference distribution and the Gaussian $\mathcal{N}_1(\mu_1, \Sigma_1)$ modeling an ASWU as the measured distribution is [24]:

$$0.5\{\mathrm{Tr}(\Sigma_1^{-1}\Sigma_0) + (\mu_1 - \mu_0)^T\Sigma_1^{-1}(\mu_1 - \mu_0) - K - \ln\frac{|\Sigma_0|}{|\Sigma_1|}\}$$

Where μ , Σ and K are the mean vector, the covariance matrix and dimension of the vector space respectively.

Figure 3 visualizes the KL-divergence matrix using a grayscale color map. Brighter pixels represent larger values (larger KL-divergences). For simplicity, in the X axis, each ASWU is shown only by the grapheme model that it belongs to, i.e., ASWUs of the form [ST_G_N] are all shown by [G]. In this figure, the X axis presents the ASWUs of size 78. It can be observed from the figure that there exists a consistent relation between the ASWUs and phones. This relation can be easily seen in the case of consonant graphemes (such as [B], [L], [M] and [R]). For example, the ASWUs belonging to grapheme model [L] are more related to /l/ and /el/ sounds and the ASWUs belonging to grapheme model [R] are more related to /r/, /axr/, and /er/ sounds as highlighted in the figure. The analysis provided here is inline with empirical observations made in grapheme-based ASR studies reported in [25].

5.1.2. Generated Pronunciations

In order to analyze the generated pronunciations, we have provided some examples of the words together with their generated pronunciations with ASWUs of size 78 in Table 2. It can be seen from the table that the G2ASWU conversion approach has learned to distinguish different sounds of the same letter to provide a pronunciation similar to what is seen in a phone-based lexicon. For example, the letter C is mapped to [ST_C_23] when it has /k/ sound and is mapped to [ST_S_23] when it corresponds to a /s/ sound. In addition, the letter A is mapped to [ST_A_22], [ST_A_27] and [ST_A_24] depending on its corresponding sound in the context. It is also interesting to note that the letter P in the word PHONE is mapped to [ST_F_22] as it has /f/ sound while in the word UPHELD it is correctly mapped to [ST_P_21].

Word	Generated pronunciation
ACCENT	ST_A_22 ST_C_23 ST_S_23 ST_E_29 ST_N_23 ST_T_23
ACCORD	ST_A_22 ST_C_23 ST_C_22 ST_O_22 ST_R_25 ST_D_21
ALAN	ST_A_22 ST_L_24 ST_A_27 ST_N_21
ALARM	ST_A_22 ST_L_24 ST_A_24 ST_R_22 ST_M_24
PHONE	ST_F_22 ST_O_26 ST_N_21
UPHELD	ST_U_22 ST_P_21 ST_H_22 ST_L_24 ST_D_21

Table 2. Sample examples for the generated pronunciations

5.2. Comparison to Related Work

Our approach is similar to the work in [6] in the sense that they both derive subword units from context-dependent grapheme-based systems. However, in our approach, instead of spectral based clustering which requires computation of a similarity matrix between HMMs, decision-tree clustering is done which is more efficient in terms of time complexity. Moreover, the pronunciations in [6] are transformed using a statistical machine translation approach while in our approach, the pronunciations are generated using the KL-HMM framework. As the experimental setup in this paper and the work in [6] are the same, we have provided a comparison between the baseline and the best results in both works in terms of WER in table 3. It can be observed from the table that the proposed approach in this paper is leading to a better performance (1.4% reduction in WER). As the two approaches are using different clustering mechanism for deriving ASWUs and different G2ASWU conversion methods, it would be interesting to ascertain where our approach is gaining. This is part of our future work.

Approaches	Grapheme subword unit	ASWU
Approach proposed in [6]	14.5	13.8
Present work	14.2	12.4

Table 3. Comparison with the related work in [6]. The results in the first row are obtained with transformed pronunciations.

6. CONCLUSION AND FUTURE DIRECTIONS

In this paper we proposed a new HMM-based formalism for subword unit derivation and pronunciation generation. We showed that the clustering technique that is generally applied for the purpose of parameter sharing and handling unseen contexts in the HMM framework, can actually be used for automatic derivation of subword units. Our experimental results show that the proposed approach is performing significantly better than a standard grapheme based ASR approach. Our future work will focus towards a) using more contextual information (than tri-graphemes) and b) developing a criteria for determining the optimal number of subword units objectively.

7. REFERENCES

[1] K. Livescu, E. Fosler-Lussier, and F. Metze, "Subword modeling for automatic speech recognition: Past, present, and emerging approaches.," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 44–57, 2012.

[2] T. Svendsen, KK. Paliwal, E. Harborg, and P. Husoy, "An improved sub-word based speech recognizer," in *Proceedings of ICASSP*, 1989, pp. 108–111.

[3] CH. Lee, F. K. Soong, and B. Juang, "A segment model based approach to speech recognition," in *Proceedings of ICASSP*, 1988.

[4] C. Lee and J. R. Glass, "A nonparametric Bayesian approach to acoustic model discovery," in *Proceedings of ACL*, 2012, pp. 40–49.

[5] A. Jansen and K. Church, "Towards unsupervised training of speaker independent acoustic models," in *Proceedings of Interspeech*, 2011, pp. 1693–1692.

[6] W. Hartmann, A. Roy, L. Lamel, and J. Gauvain, "Acoustic unit discovery and pronunciation generation from a grapheme-based lexicon," in *Proceedings of ASRU*, 2013, pp. 380–385.

[7] M. Bisani and H. Ney, "Joint-sequence models for grapheme-to-phoneme conversion," *Speech Communication*, vol. 50, no. 5, May 2008.

[8] V. Pagel, K. Lenzo, and A. Black, "Letter to sound rules for accented lexicon compression," in *Proceedings of ICSLP*, 1998, vol. 5, pp. 2015–2020.

[9] T. Sejnowski and C. Rosenberg, "Parallel networks that learn to pronounce English text," *Complex systems*, vol. 1, no. 1, pp. 145–168, 1987.

[10] P. Taylor, "Hidden Markov models for grapheme to phoneme conversion.," in *Proceedings of Interspeech*, 2005, pp. 1973–1976.

[11] T. Fukada, M. Bacchiani, KK Paliwal, and Sagisaka. Y., "Speech recognition based on acoustically derived segment units," in *Proceedings of ICSLP*, 1996.

[12] KK. Paliwal, "Lexicon-building methods for an acoustic sub-word based speech recognizer," in *Proceedings of ICASSP*, Apr 1990, pp. 729–732 vol.2.

[13] M. Bacchiani and M. Ostendorf, "Joint lexicon, acoustic unit inventory and model design," *Speech Communication*, vol. 29, no. 2, pp. 99–114, 1999.

[14] T. Holter and T. Svendsen, "Combined optimisation of baseforms and model parameters in speech recognition based on acoustic subword units," in *Proceedings of ASRU*, Dec 1997, pp. 199–206.

[15] C. Lee, Y. Zhang, and J. R. Glass, "Joint learning of phonetic units and word pronunciations for ASR.," in *EMNLP*. 2013, pp. 182–192, ACL.

[16] G. Aradilla, H. Bourlard, and M. Magimai-Doss, "Using KL-based acoustic models in a large vocabulary recognition task.," in *Proceedings of Interspeech*, 2008, pp. 928–931.

[17] M. Magimai-Doss, R. Rasipuram, G. Aradilla, and H. Bourlard, "Grapheme-based automatic speech recognition using KL-HMM," in *Proceedings of Interspeech*, Aug. 2011.

[18] R. Rasipuram and M. Magimai-Doss, "Acoustic data-driven grapheme-to-phoneme conversion using KL-HMM," in *Proceedings of ICASSP*, Mar. 2012.

[19] R. Rasipuram and M. Magimai.-Doss, "Acoustic and lexical resource constrained asr using language-independent acoustic model and language-dependent probabilistic lexical model," *Speech Communication*, vol. 68, pp. 23–40, Apr. 2015.

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

[21] X. Luo and F. Jelinek, "Probabilistic classification of HMM states for large vocabulary continuous speech recognition," in *Proceedings of ICASSP*, 1999, vol. 1, pp. 353–356.

[22] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book Version 3.0*, Cambridge University Press, 2000.

[23] D. Johnson et al., "ICSI Quicknet Software Package," http://www.icsi.berkeley.edu/Speech/qn.html, 2004.

[24] J. Duchi, "Derivations for linear algebra and optimization," http://www.cs.berkeley.edu/∼jduchi/projects/general_notes.pdf, 2007.

[25] R. Rasipuram and M. Magimai-Doss, "Improving grapheme-based ASR by probabilistic lexical modeling approach," in *Proceedings of Interspeech*, 2013.