INTEGRATED PRONUNCIATION LEARNING FOR AUTOMATIC SPEECH RECOGNITION USING PROBABILISTIC LEXICAL MODELING

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

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

ABSTRACT

Standard automatic speech recognition (ASR) systems use phoneme-based pronunciation lexicon prepared by linguistic experts. When the hand crafted pronunciations fail to cover the vocabulary of a new domain, a grapheme-to-phoneme (G2P) converter is used to extract pronunciations for new words and then a phoneme-based ASR system is trained. G2P converters are typically trained only on the existing lexicons. In this paper, we propose a grapheme-based ASR approach in the framework of probabilistic lexical modeling that integrates pronunciation learning as a stage in ASR system training, and exploits both acoustic and lexical resources (not necessarily from the domain or language of interest). The proposed approach is evaluated on four lexical resource constrained ASR tasks and compared with the conventional two stage approach where G2P training is followed by ASR system development.

Index Terms— Probabilistic lexical modeling, pronunciation lexicon, grapheme subwords, phoneme subwords, grapheme-to-phoneme conversion.

1. INTRODUCTION

Automatic speech recognition (ASR) systems model words as a sequence of subword units which are further modeled as a sequence of hidden Markov model (HMM) states. This is primarily done to address data sparsity issues and achieve generalization towards unseen words. The sequence of subword units for a word is given by its pronunciation model as specified in the pronunciation lexicon. All the components in an ASR system presume the availability of a subword unit set and a pronunciation lexicon. Therefore, in practice, ASR system development can be seen as a two stage process: development of pronunciation lexicon followed by ASR system training.

Typically, ASR systems use linguistically motivated phonemes as subword units. Phoneme pronunciations are typically obtained from a hand-built lexicon which the linguistic experts have prepared. The hand crafted phoneme lexicon will provide optimum performance for ASR. Most often, the existing hand labeled phoneme lexicon (or seed lexicon) may not provide complete coverage for a new domain (target domain) for which we are interested to build an ASR system. A commonly adopted way to generate or augment the phoneme pronunciation lexicon is through automatic grapheme-to-phoneme (G2P) conversion [1, 2, 3]. The two main requirements of automatic G2P conversion systems are: a seed lexicon and a method to capture the G2P relationship observed in the seed lexicon. The G2P converter is used to augment the training vocabulary. The augmented lexicon is then used to build an ASR system [4, 5, 6]. G2P converters are also used to augment the recognition vocabulary.

Other alternatives for subword units are graphemes [7, 8, 9, 10, 11] which make the pronunciation lexicon development easy. Graphemes are the units of written language, e.g., letters of English. However, modeling graphemes for speech recognition is a challenging task for two reasons. Firstly, G2P relationship can be ambiguous as languages continue to evolve after their spelling has been standardized. Secondly, typically ASR systems directly model the relationship between graphemes and acoustic features; and the acoustic features depict the envelope of speech, which is related to phonemes. The studies until now have shown that for languages such as English and French that have an irregular G2P relationship, grapheme-based ASR systems perform worse compared to phoneme-based ASR systems [8, 9, 10].

In this paper, we propose a grapheme-based ASR approach in the framework of probabilistic lexical modeling [12, 13] where first, acoustic-to-phoneme relationship is modeled with available acoustic and lexical resources (not necessarily from the domain of interest), and then a probabilistic G2P relationship is learned given the transcribed speech data of the target domain (Section 2). As the parameters of the model capture a probabilistic G2P relationship learned through acoustic data, the approach integrates pronunciation learning implicitly as a phase in ASR system training. In this paper, on four pronunciation lexicon resource constrained ASR tasks, the proposed grapheme-based system is compared with conventional approach where first, G2P conversion is performed and then a phoneme-based system is trained (Sections 3 and 4). Furthermore, the proposed grapheme-based ASR approach is evaluated against other grapheme-based ASR approaches proposed in the literature [8, 9, 10].

2. INTEGRATED PRONUNCIATION LEARNING

In subword unit based ASR systems, HMM states represent subword or lexical units, i.e., $q_t \in \mathcal{L} = \{l^1, \dots l^i \dots l^I\}$, where q_t is the HMM state at time t and I is the number of lexical units. Typically, each context-independent or context-dependent subword is modeled with three HMM states. In the framework of probabilistic lexical modeling [12, 13], the relationship between the acoustic feature observation $\mathbf{x}_t$ and the lexical unit l^i at time t is factored through a *latent* variable a^d as follows:

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

$$= \sum_{d=1}^{D} \underbrace{p(\mathbf{x}_t|a^d, \theta_a)}_{\textbf{acoustic model}} \cdot \underbrace{P(a^d|q_t = l^i, \theta_l)}_{\textbf{lexical model}} \quad (2)$$

This work was supported by the Swiss NSF through the grants Flexible Grapheme-Based Automatic Speech Recognition (FlexASR) and by Hasler foundation through the grants Flexible acoustic data driven grapheme to acoustic unit conversion (AddG2SU).

We refer to the latent variable a^d as the acoustic unit and the set of acoustic units $\mathcal{A} = \{a^1, \ldots a^d, \ldots a^D\}$ where D is the total number of acoustic units. The relationship in Eqn (2) is as a result of the assumption that given a^d, $p(\mathbf{x}_t|a^d, q_t = l^i, \theta_a, \theta_l)$ is independent of l^i. In Eqn (2), $p(\mathbf{x}_t|a^d, \theta_a)$ is the acoustic unit likelihood and $P(a^d|l^i, \theta_l)$ is the probability of the latent variable given the lexical unit. We refer to $p(\mathbf{x}_t|a^d, \theta_a)$ as the acoustic model and $P(a^d|l^i, \theta_l)$ as the lexical model. The parameters of the acoustic likelihood estimator Θ_A encompass the *acoustic model* (θ_a), the *pronunciation lexicon* (θ_{pr}) and the *lexical model* (θ_l) parameters, therefore, $\Theta_A = \{\theta_a, \theta_{pr}, \theta_l\}$.

In the literature, there are two main approaches for acoustic modeling, namely, Gaussian mixture models (GMMs) and artificial neural networks (ANNs). In [13], we have shown that in standard HMM-based ASR approaches like HMM/GMM and hybrid HMM/ANN, the relationship between lexical units and acoustic units is one-to-one and the lexical model is deterministic. Furthermore, it was shown that in Kullback-Leibler divergence based hidden Markov model (KL-HMM) [14] and tied posterior [15] approaches, the relationship between lexical and acoustic units is probabilistic (probabilistic lexical modeling).

2.1. Grapheme-based ASR Approach

In the framework of probabilistic lexical modeling, the modeling of the relationship between graphemes and acoustic features can be factored into two parts through acoustic units:

1. *The acoustic model* where the relationship between acoustic units and acoustic features is modeled.
2. *The lexical model* where a probabilistic relationship between acoustic units and graphemes is modeled.

In this paper, we use the KL-HMM approach for probabilistic lexical modeling. The KL-HMM approach assumes that an acoustic unit set $\mathcal{A}$ is defined and a trained acoustic model is available. Therefore, in the first step a standard HMM-based ASR system i.e., either an HMM/GMM system or a hybrid HMM/ANN system is trained. The acoustic model (GMMs in the case of HMM/GMM or ANN in the case of hybrid HMM/ANN) is used with the pronunciation lexicon and acoustic data to train the parameters of the lexical model.

In the KL-HMM approach, the parameters of the lexical model are learned through acoustic unit posterior probability estimates. Given the acoustic model, acoustic unit probability sequences of training data are estimated. The acoustic unit probability sequences are used as feature observations to train an HMM with graphemes as lexical units using the KL-HMM approach. In the KL-HMM approach, HMM states are parameterized by categorical distributions and model a probabilistic relationship between a lexical unit and D acoustic units.

In the proposed approach with graphemes as lexical units and phonemes as acoustic units, the lexical model parameters capture a probabilistic relationship between graphemes and phonemes. Furthermore, the probabilistic G2P relationship is learned on acoustic data. Thus, the proposed grapheme-based ASR approach integrates lexicon learning as a phase in ASR system training and could potentially remove the necessity of training an explicit G2P converter. In [11][1], where the acoustic units were phonemes it was elucidated that the lexical model parameters in the proposed grapheme-based ASR approach indeed capture a probabilistic G2P relationship.

[1]It is important to note that in [11] the notion of probabilistic lexical modeling was not introduced.

2.2. Contributions of the Present Paper

Motivated by the analysis of the lexical model parameters and the probabilistic G2P relationship captured by them [11], in this paper we hypothesize that with probabilistic lexical modeling it is possible to build an ASR system that uses a grapheme lexicon and achieves performance as good as ASR systems, where first G2P conversion is performed to build a lexicon and then a phoneme-based ASR system is trained. To validate our hypothesis, we investigate the grapheme-based ASR approach in the following two lexical resource constrained scenarios that are commonly encountered.

In the first case, the target domain for which we are interested to build an ASR system has only word level transcribed speech (acoustic) data. Cross-domain acoustic and lexical resources are available, but they do not provide complete coverage on the target domain data. In the proposed grapheme-based ASR approach, first an acoustic model with phonemes as acoustic units and then a lexical model with graphemes as lexical units are trained. In the framework of probabilistic lexical modeling, the parameters of the acoustic model θ_a and the parameters of the lexical model θ_l can be trained on an independent set of resources [13]. Therefore, it is possible to build an ASR system where the acoustic model is trained on cross-domain acoustic and lexical resources; and the lexical model is learned on target domain data.

In the second case, there is a need to expand the recognition vocabulary. In this case, a grapheme-based ASR system can be trained where the acoustic unit set is based on phonemes and the acoustic model is trained on acoustic and lexical resources of target domain data. The lexical unit set is based on graphemes and the lexical model is also learned on target domain data. Since the lexical units are graphemes, the recognition vocabulary can be augmented easily.

Figure 1 illustrates block diagram of the conventional phoneme-based ASR system using phoneme lexicon from a G2P converter and Figure 2 illustrates the proposed grapheme-based ASR system.

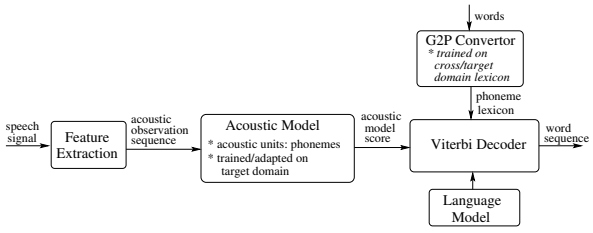

Fig. 1. The phoneme-based ASR system using a G2P converter

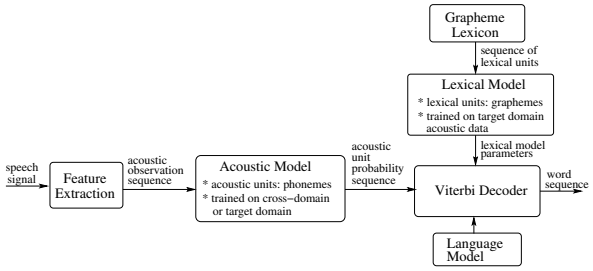

Fig. 2. The proposed grapheme-based ASR system

3. EXPERIMENTAL SETUP

The present paper investigates the hypothesis that the proposed grapheme-based ASR approach achieves performance as good as phoneme-based ASR systems using a phoneme lexicon from a G2P converter. Furthermore, the proposed grapheme-based ASR

approach incorporating a probabilistic lexical model is compared with grapheme-based HMM/GMM [8, 9] and grapheme-based Tandem [10] approaches that use a deterministic lexical model. Towards these goals, we present four different ASR studies using the following three lexica:

1. *GRAPH* - grapheme lexicon transcribed using the orthography of words.
2. *G2P* - phoneme lexicon obtained using a joint n-gram based G2P converter [2]. We used the sequitur G2P toolkit.
3. *PHONE* - well developed phoneme lexicon designed for each ASR task. This lexicon serves as an optimistic case as it is manually built and verified.

Following the previous work [12, 13], we use the KL-HMM approach for probabilistic lexical modeling. In this paper, we compare the KL-HMM, HMM/GMM and Tandem systems. All the systems model crossword context-dependent subword units (lexical units). In the KL-HMM systems acoustic units are context-independent subword units, and lexical and acoustic units are probabilistically related. In the HMM/GMM and Tandem systems, acoustic units are clustered context-dependent subword units, and lexical and acoustic units are deterministically related. The capabilities of various systems are summarized in Table 1.

Table 1. Overview of different systems. CI denotes context-independent subword units, CD denotes context-dependent subword units and cCD denotes clustered context-dependent subword units. P and G denote the phoneme and grapheme lexicon, respectively. *Det* denotes the lexical model is deterministic and *Prob* denotes the lexical model is probabilistic.

System	Acoustic units $\mathcal{A}$	Lexicon	Lexical units $\mathcal{L}$	Approach
KL-HMM	CI	P or G	CD	*Prob*
Tandem	cCD	P or G	CD	*Det*
HMM/GMM	cCD	P or G	CD	*Det*

We use multilayer perceptrons (MLPs) trained to classify context-independent phonemes as the acoustic models for the KL-HMM systems. Input to all the MLPs is the 39-dimensional PLP cepstral coefficient vector with a nine-frame context. The KL-HMM systems used the symmetric KL-divergence as the local score [11]. The Tandem features were extracted by transforming the output of the MLPs (same MLPs used in the KL-HMM systems) with log transformation followed by Karhunen-Loeve transform (KLT). The 39-dimensional PLP feature vector used to train the MLP are also used to train the HMM/GMM systems.

The KL-HMM and Tandem systems are capable of exploiting cross-domain acoustic and lexical resources (if available) by using an MLP trained on cross-domain resources. The lexical model of KL-HMM systems is always trained on target-domain resources whereas both acoustic and lexical models of Tandem systems are trained on target-domain resources. The HMM/GMM system is trained on target-domain data alone. In this paper, we do not perform acoustic model adaptation of HMM/GMM systems on cross-domain resources, as it is assumed that the tasks lack only lexical resources.

In the following subsections we will describe the four investigated ASR studies that reflect practical scenarios. The first three studies are on English whereas the fourth study is on French. The G2P relationship is irregular in both English and French.

3.1. Cross-Domain ASR Study

The goal is to build an ASR system for a new domain without a phoneme pronunciation lexicon. Cross domain acoustic and lexical resources are available, however, the cross domain lexicon has a high out-of-vocabulary rate on the new domain. In that regard, we present an experimental study where the RM corpus [16] is considered as the target domain and the WSJ1 [17] corpus as the cross-domain. The standard RM setup with 3990 train utterances and 1200 test utterances is used in this study[2]. Though RM and WSJ are similar domains, among the 1000 words present in the RM task, the WSJ task includes only 568 words. That is, the RM task has 432 words that are not seen in the WSJ pronunciation lexicon. We use an *off-the-shelf* three-layer MLP [14] trained on the WSJ1 (to classify 45 context-independent phonemes) for the KL-HMM and Tandem systems.

The G2P converter trained on WSJ1 lexicon is used to estimate pronunciations for the RM words. The optimal n-gram context size was 5. The performance of the *G2P* lexicon compared to the *PHONE* lexicon given in the RM task was 92.2% phoneme accuracy.

3.2. Multi Accent Non-Native ASR Study

The goal is to build an ASR system for non-native speech including multiple accents without a phoneme pronunciation lexicon. In this study, cross-domain acoustic and lexical resources are from native language speakers. The spoken words in non-native speech are pronounced differently from native pronunciations. Capturing these variations through multiple pronunciations is not a trivial task [19]. Therefore, the approaches should implicitly handle lexical resource constraints and model the pronunciation variability.

We study multi-accent non-native speech recognition using the HIWIRE corpus [20][2]. As cross-domain resources we use the SpeechDat(II) British English corpus that includes acoustic and lexical resources from native language speakers.

A three-layer MLP [18] trained on the SpeechDat(II) British English corpus to classify 45 context-independent phonemes was used for the KL-HMM and Tandem systems. SpeechDat(II) is a telephone speech corpus, hence, the HIWIRE speech was down sampled to 8kHZ before extracting PLP cepstral features and then forward passed through the SpeechDat(II) English MLP.

The G2P converter trained on SpeechDat(II) British English lexicon is used to estimate pronunciations for the HIWIRE words. The optimal width of the graphone context was found to be 6. The performance of the *G2P* lexicon compared to the *PHONE* lexicon of the HIWIRE task was 89.4% phoneme accuracy.

3.3. Lexicon Augmentation Study

In this study our goal is to augment the test vocabulary with new words that are not present in the training vocabulary. The training data includes limited word level transcribed speech data with the phoneme pronunciations of words seen in the training data. The study is performed on the PhoneBook speaker-independent task-independent 600 word isolated word recognition corpus where none of the words in the test vocabulary are present in the training vocabulary [21].

The MLP for the KL-HMM and Tandem systems was trained on limited training data of the PhoneBook corpus to classify 42 context-independent phonemes. For MLP training, we followed the same setup as in [22], where 19421 utterances are used for training and 7920 utterances for cross validation. Thus, the data used to train the MLP did not contain any of the test words.

Systems are built using both training and cross validation utterances consisting of 27341 utterances covering 2183 words. Phoneme-based KL-HMM systems used the phoneme lexicon

[2]In the previous work on RM [11] and HIWIRE [18] corpora the performance of systems using word-internal context-dependent subwords was reported.

given in the PhoneBook corpus with acoustic units as context-independent phonemes and lexical units as context-dependent phonemes. Grapheme-based KL-HMM systems are trained with acoustic units as context-independent phonemes and lexical units as context-dependent graphemes. In the case of the PhoneBook task word-internal context-dependent systems are built (as it is an isolated word recognition task).

The *G2P* lexicon for the test set was built by training a G2P converter on the training and cross-validation pronunciation lexicon (consisting of pronunciations for 2183 words). The performance of the *G2P* lexicon compared to the *PHONE* lexicon given in the PhoneBook task was 89.2% phoneme accuracy.

3.4. LVCSR Study

Finally, we study the proposed approach on a large vocabulary continuous speech recognition (LVCSR) task involving non-native speakers during recognition. Furthermore, the recognition vocabulary includes words that are not seen during training. To study such a scenario we use Mediaparl French corpus [23]. MediaParl is a bilingual corpus containing recordings of debates in Valais parliament in Switzerland in both Swiss German and Swiss French. We use only the French part of the MediaParl corpus and the experimental setup is similar to [23]. All the speakers in the training and development set are native speakers. In the test set, four speakers are German native speakers and for three speakers, French is the native language. The train vocabulary includes 12196 words and test vocabulary includes 4246 words out of which 915 words are not seen during training. We use 5-layer MLP reported in [24] trained to classify 38 context-independent phonemes for all the KL-HMM and Tandem systems. The *G2P* lexicon for the test vocabulary was built by training a G2P converter on the training and cross-validation pronunciation lexicon. The performance of the *G2P* lexicon for the test vocabulary compared to the *PHONE* lexicon given in the MediaParl French task was 97.4% phoneme accuracy.

4. RESULTS

The performance in terms of word accuracy of the various systems using three different lexica on the RM, HIWIRE, PhoneBook and MediaParl tasks is given in Tables 2, 3, 4, and 5 respectively. In the tables, boldface indicates the best system for each lexicon. The main observations from the four tasks are as follows:

- On the HIWIRE and PhoneBook tasks, the KL-HMM systems using the *GRAPH* lexicon perform better than the KL-HMM systems using the *G2P* lexicon whereas on the RM and MediaParl tasks, the KL-HMM systems using the *GRAPH* lexicon perform similar to the KL-HMM systems using the *G2P* lexicon. Furthermore, in the case of the RM and HIWIRE tasks, the KL-HMM systems using the *GRAPH* lexicon achieve performance comparable to the KL-HMM systems using the optimistic well developed *PHONE* lexicon.
- The performance of the KL-HMM systems using the *GRAPH* and *G2P* lexica is always better than that of the HMM/GMM systems. The performance of the KL-HMM system using the *PHONE* lexicon is similar to or better than that of the HMM/GMM system.
- The results show that on all the four tasks, the HMM/GMM systems using the *PHONE* lexicon perform better than the HMM/GMM systems using the *GRAPH* or *G2P* lexicon. The results also show that on the RM and HIWIRE tasks, HMM/GMM systems using the *GRAPH* lexicon and the *G2P* lexicon perform similarly. However, on the PhoneBook task where the recognition vocabulary is entirely different from the train vocabulary, the system using the *GRAPH* lexicon performs significantly better than the system using the *G2P* lexicon. On the MediaParl task with a large vocabulary, the system using the *GRAPH* lexicon performs worse than the system using the *G2P* lexicon.
- The performance of the Tandem systems is worse (for all the three lexica) than the KL-HMM systems. The performance of the Tandem systems is similar to or worse than that of the HMM/GMM systems.

The results confirm our two hypotheses. Firstly, the proposed grapheme-based system can perform better than or comparably to the phoneme-based system using the phoneme lexicon from a G2P converter. Secondly, the ASR systems incorporating a probabilistic lexical model, handle the pronunciation errors inherent in the *GRAPH* and *G2P* lexica better than the standard ASR systems incorporating a deterministic lexical model.

Table 2. Word accuracies (in %) on the test set of the RM corpus.

System	*GRAPH*	*G2P*	*PHONE*
KL-HMM	**95.5**	**95.6**	**95.9**
Tandem	94.5	94.6	95.4
HMM/GMM	94.8	95.1	**95.9**

Table 3. Word accuracies (in %) on test set of the HIWIRE corpus.

System	*GRAPH*	*G2P*	*PHONE*
KL-HMM	**97.5**	**96.8**	**97.3**
Tandem	96.6	96.2	97.0
HMM/GMM	96.4	96.1	**97.2**

Table 4. Word accuracies (in %) on test set of PhoneBook corpus.

System	*GRAPH*	*G2P*	*PHONE*
KL-HMM	**93.6**	**89.1**	**97.8**
Tandem	92.7	84.9	97.4
HMM/GMM	91.0	86.7	97.0

Table 5. Word accuracies (in %) on test set of MediaParl corpus.

System	*GRAPH*	*G2P*	*PHONE*
KL-HMM	**71.7**	**71.9**	**74.1**
Tandem	65.3	64.1	66.1
HMM/GMM	68.4	70.7	73.2

5. DISCUSSION AND CONCLUSIONS

In this paper, we proposed a grapheme-based ASR approach where first acoustic-to-phoneme relationship is learned on acoustic and lexical resources and then a probabilistic grapheme-to-phoneme is learned on word level transcribed speech data from the target language or domain. The studies on four lexical resource constrained ASR tasks have shown that the proposed grapheme-based ASR approach which implicitly integrates lexicon learning performs better than or comparably to the conventional two stage approach where G2P training is followed by ASR system development. Standard G2P converters rely on language-dependent lexical resources to learn the relationship between graphemes and phonemes. The proposed grapheme-based ASR approach can exploit domain-independent (as shown in this paper) or language-independent [13] acoustic and lexical resources. The studies in this paper and the previous work [13] show that the proposed approach could potentially prevent the need for an explicit G2P convertor. In this paper, we focused mainly on the lexical model of the proposed system. In our future work we intend to bridge the gap between the proposed grapheme-based ASR system and the system using well developed phoneme lexicon by improving the acoustic model using deep ANN architectures and/or by using clustered context-dependent subword units as acoustic units.

6. REFERENCES

[1] V. Pagel, K. Lenzo, and A.W. Black, "Letter to Sound Rules for Accented Lexicon Compression," in *Proceedings of Int. Conf. Spoken Language Processing*, 1998.

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

[3] D. Wang and S. King, "Letter-to-Sound Pronunciation Prediction Using Conditional Random Fields," *Signal Processing Letters, IEEE*, vol. 18, no. 2, pp. 122–125, 2011.

[4] C. Gollan, M. Bisani, S. Kanthak, R. Schluter, and H. Ney, "Cross Domain Automatic Transcription on the TC-STAR EPPS Corpus," in *Proc. of ICASSP*, 2005, vol. 1, pp. 825–828.

[5] J. Lööf, M. Bisani, Ch, G. Heigold, Björn Hoffmeister, Ch, R. Schlüter, and H. Ney, "The 2006 RWTH parliamentary speeches transcription system," in *Proceedings of Int. Conf. Spoken Language Processing*, 2006.

[6] D. Jouvet, D. Fohr, and I. Illina, "Evaluating grapheme-to-phoneme converters in automatic speech recognition context," in *Proc. of ICASSP*, 2012, pp. 4821–4824.

[7] E.G. Schukat-Talamazzini, H. Niemann, W. Eckert, T. Kuhn, and S. Rieck, "Automatic Speech Recognition Without Phonemes," in *Proc. of European Conf. on Speech Communication and Technology (EUROSPEECH)*, 1993.

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

[9] M. Killer, S. Stüker, and T. Schultz, "Grapheme based Speech Recognition," in *Proc. of EUROSPEECH*, 2003.

[10] 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.

[11] M. Magimai.-Doss, R. Rasipuram, G. Aradilla, and H. Bourlard, "Grapheme-based Automatic Speech Recognition using KL-HMM," in *Proc. of Interspeech*, 2011, pp. 445–448.

[12] R. Rasipuram and M. Magimai.-Doss, "Probabilistic Lexical Modeling and Grapheme-based Automatic Speech Recognition," `http://publications.idiap.ch/downloads/reports/2013/Rasipuram_Idiap-RR-15-2013.pdf`, 2013, Idiap Research Report.

[13] R. Rasipuram and M. Magimai.-Doss, "Acoustic and Lexical Resource Constrained ASR using Language-Independent Acoustic Model and Language-Dependent Probabilistic Lexical Model," `http://publications.idiap.ch/downloads/reports/2014/Rasipuram_Idiap-RR-02-2014.pdf`, 2014, Idiap Research Report.

[14] G. Aradilla, H. Bourlard, and M. Magimai Doss, "Using KL-Based Acoustic Models in a Large Vocabulary Recognition Task ," in *Proc. of Interspeech*, 2008, pp. 928–931.

[15] J. Rottland and G. Rigoll, "Tied Posteriors: An Approach for Effective Introduction of Context Dependency in Hybrid NN/HMM LVCSR," in *Proc. of ICASSP*, 2000, pp. 1241–1244.

[16] P. J. Price, W. Fisher, and J. Bernstein, "The DARPA 1000-word resource management database for continuous speech recognition," in *Proc. of ICASSP*, 1988, pp. 651–654.

[17] P. C. Woodland, J. J. Odell, V. Valtchev, and S. J. Young, "Large Vocabulary Continuous Speech Recognition using HTK," in *Proc. of ICASSP*, 1994, vol. 2, pp. 125–128.

[18] 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 Automatic Speech Recognition and Understanding (ASRU)*, 2011, pp. 348–353.

[19] H. Strik and C. Cucchiarini, "Modeling pronunciation variation for ASR: A survey of the literature," *Speech Communication*, vol. 29, pp. 225–246, 1999.

[20] J.C. Segura, T. Ehrette, A. Potamianos, D. Fohr, I. Illina, P-A. Breton, V. Clot, R. Gemello, M. Matassoni, and P. Maragos, "The HIWIRE Database, a Noisy and Non-native English Speech Corpus for Cockpit Communication," `http://cvsp.cs.ntua.gr/projects/pub/HIWIRE/WebHome/HIWIRE_db_description_paper.pdf`, 2007.

[21] J. Pitrelli, C. Fong, S.H. Wong, J.R. Spitz, and H.C. Leung, "PhoneBook: a phonetically-rich isolated-word telephone-speech database," in *Proc. of ICASSP*, 1995, vol. 1, pp. 101–104.

[22] S. Dupont, H. Bourlard, O. Deroo, V. Fontaine, and J. M. Boite, "Hybrid HMM/ANN Systems for Training Independent Tasks: Experiments on 'Phonebook' and Related Improvements," in *Proc. of ICASSP*, 1997.

[23] D. Imseng, H. Bourlard, H. Caesar, P.N. Garner, G. Lecorvé, and A. Nanchen, "MediaParl: Bilingual mixed language accented speech database," in *Proc. of IEEE Workshop on Spoken Language Technology (SLTU)*, 2012, pp. 263–268.

[24] M. Razavi and M. Magimai-Doss, "On Recognition of Non-Native Speech Using Probabilistic Lexical Model," in *Proc. of Interspeech*, 2014.