ACOUSTIC DATA-DRIVEN GRAPHEME-TO-PHONEME CONVERSION USING KL-HMM

Ramya Rasipuram[†,‡] *and Mathew Magimai.-Doss*[†]

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

ABSTRACT

This paper proposes a novel grapheme-to-phoneme (G2P) conversion approach where first the probabilistic relation between graphemes and phonemes is captured from acoustic data using Kullback-Leibler divergence based hidden Markov model (KL-HMM) system. Then, through a simple decoding framework the information in this probabilistic relation is integrated with the sequence information in the orthographic transcription of the word to infer the phoneme sequence. One of the main application of the proposed G2P approach is in the area of low linguistic resource based automatic speech recognition or text-to-speech systems. We demonstrate this potential through a simulation study where linguistic resources from one domain is used to create linguistic resources for a different domain.

Index Terms— Kullback-Leibler divergence based HMM, Lexicon, grapheme, phoneme, grapheme-to-phoneme converter, multilayer perceptron.

1. INTRODUCTION

Grapheme-to-phoneme (G2P) converters are used in automatic speech recognition (ASR) systems and text-to-speech synthesis (TTS) systems to generate pronunciation variants/models. In literature, different G2P approaches have been proposed, where statistical models such as, decision trees [1], joint multigram models [2], or conditional random fields [3] are used to learn pronunciation rules. All these approaches invariantly assume access to "prior" linguistic resources (e.g., phoneme set, pronunciation dictionary) or in simple terms access to parallel data consisting of sequences of graphemes and their corresponding sequences of phonemes. Such resources may not be readily available for all languages/domains.

More recently, we proposed a grapheme-based ASR system in the framework of Kullback-Leibler divergence based hidden Markov model (KL-HMM) [4, 5]. In this system, the relationship between acoustics and grapheme subword units is modeled in two steps. First, a multilayer perceptron (MLP) is trained to capture the relationship between acoustic features, such as cepstral features and phoneme classes. Then, by using the phoneme posterior probabilities (also referred to as posterior features) estimated by the MLP as feature observation, probabilistic relationship between graphemes and phonemes is captured via the state multinomial distributions of the KL-HMM system.

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). The authors would like to thank their colleagues Prof. Hervé Bourlard, David Imseng, Dr. John Dines, and Lakshmi Saheer for fruitful discussions.

This paper builds upon the above described previous work on grapheme-based ASR to propose a novel acoustic data-driven G2P conversion approach. More specifically, this approach exploits the probabilistic relationship between graphemes and phonemes captured by the KL-HMM system and the sequence information in the orthographic transcription of the word to extract pronunciation models/variants. One of the main application of this approach can be seen in the context of languages/domains that may not have prior linguistic resources. In that respect, this paper pursues a line of investigation to demonstrate the potential of the proposed approach where,

1. The MLP used to extract posterior features is trained on auxiliary/out-of-domain data. This can be likened to the scenario where the MLP is trained to classify phonemes using data from languages/domains that have prior linguistic resources.
2. A grapheme-based KL-HMM system is then trained for a task (that is assumed to not have prior linguistic resources) with posterior features extracted from in-domain acoustic data. In this case, the state multinomial distributions of KL-HMM system captures the relationship between the phonemes in the linguistic resources of auxiliary data and the graphemes.
3. Finally, a phoneme-based lexicon is built for the task from scratch automatically using the orthographic transcription of words and the KL-HMMs of grapheme subword units. The phoneme-based lexicon thus obtained is analyzed by comparing it with an existing phoneme-based lexicon at three different levels, namely, phoneme error level, word error level, and ASR system performance level.

The paper is organized as follows. Section 2 briefly introduces the KL-HMM system and summarizes our previous work on grapheme-based ASR. Section 3 presents the proposed grapheme-to-phoneme conversion approach followed by presentation of experimental studies in Section 4. Finally, we conclude in Section 5.

2. KL-HMM SYSTEM

Figure 1 illustrates a KL-HMM system where graphemes are used as subword units and each grapheme subword unit is modeled by a single state HMM. In KL-HMM system [5, 6], posterior probabilities of acoustic classes (or simply referred to as posterior feature) is used as feature observation. For simplicity, let the acoustic classes be phonemes. Let $\mathbf{z}_t$ denote the phoneme posterior feature vector estimate at time frame t,

$$\mathbf{z}_t = [z_t^1, \cdots, z_t^D]^T = [P(p_1|\mathbf{x}_t), \cdots, P(p_D|\mathbf{x}_t)]^T$$

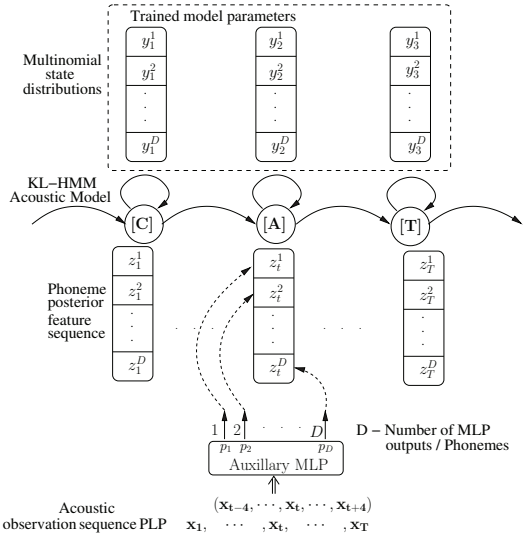

Fig. 1. Illustration of KL-HMM system using grapheme as subword units.

where $\mathbf{x}_t$ is the acoustic feature (e.g., cepstral feature) at time frame t, $\{p_1, \cdots p_d, \cdots p_D\}$ is the phoneme set, D is the number of phonemes, and $P(p_d|\mathbf{x}_t)$ denotes the a posteriori probability of phoneme p_d given $\mathbf{x}_t$. In the original work as well as in this work, $\mathbf{z}_t$ is estimated by a well trained MLP.

Each HMM state i in the KL-HMM system is parameterized by a multinomial distribution $\mathbf{y}_i = [y_i^1, \cdots, y_i^D]^T$. The local score at each HMM state is estimated as Kullback-Leibler (KL) divergence between $\mathbf{y}_i$ and $\mathbf{z}_t$, i.e.,

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

In this case, $\mathbf{y}_i$ serves as the reference distribution and $\mathbf{z}_t$ serves as the test distribution.

KL-divergence being an asymmetric measure, there are also other ways to estimate the local score,

1. Reverse KL-divergence (RKL):

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

2. Symmetric KL-divergence (SKL):

$$SKL(\mathbf{y}_i, \mathbf{z}_t) = KL(\mathbf{y}_i, \mathbf{z}_t) + RKL(\mathbf{z}_t, \mathbf{y}_i) \quad (3)$$

The HMM state parameters i.e., multinomial distributions are estimated by using Viterbi expectation maximization algorithm which minimizes a cost function based on one of the above local scores. During testing, decoding is performed using standard Viterbi decoder. For more details and interpretation of the systems resulting from different local scores the reader is referred to [5, 6].

In [4], we proposed a grapheme-based ASR system using KL-HMM. We studied this system on DARPA Resource Management (RM) corpus. More precisely, for each of the local score described earlier we trained systems that model different amount of grapheme subword unit context, namely, *context-independent*, *tri-grapheme* (single preceding and following context), *quint-grapheme* (two preceding and following contexts). We compared these systems with their respective phoneme-based systems. It was found that longer grapheme subword unit context, i.e. quint-grapheme based system yields performance comparable to phoneme-based system. Also, when modelling subword context, local score SKL was found to yield the best system for both phoneme and grapheme. Upon analysis of the trained grapheme subword unit models it was found that

- Context-independent grapheme models capture gross phoneme information, i.e. the state multinomial distributions capture information about different phonemes. For instance, HMM of grapheme [C] dominantly captures the relation to phonemes /s/, /k/, /ch/. This is mainly due to the fact that in English language the correspondence between graphemes and phonemes is weak.
- Single preceding and following context models of consonant graphemes are able to dominantly capture the relation to appropriate phoneme. For instance, HMM of grapheme [C+A] [1] capture dominantly the relation to phoneme /k/, while HMM of grapheme [C+E] capture the relation to phoneme /s/. In other words, through contextual modelling the ambiguity present in context-independent grapheme models is resolved well.
- Vowel graphemes need longer context to dominantly capture the relation to appropriate phoneme. This observation is synonymous to G2P converters of English language which may need longer context to map a vowel grapheme to a unique phoneme.
- The states of the context-dependent grapheme models are also able to capture some information about preceding and following phonemes.

3. ACOUSTIC DATA-DRIVEN G2P APPROACH

One of the key issue when developing a G2P converter is how to effectively learn/capture the relation between phonemes and graphemes. As discussed briefly in the previous section, when using graphemes as subword units in KL-HMM based ASR system this relation is captured probabilistically through the state multinomial distributions. The proposed G2P approach which builds upon this observation consists of two phases:

1. Training: In this phase, a grapheme-based KL-HMM system is trained using phoneme posterior features [4]. It is to be noted that this phase also includes the training of phoneme posterior feature estimator. As mentioned earlier, in our case, it is a well trained MLP.
2. Decoding: Given the KL-HMMs of grapheme subword units and the orthographic transcription of the word, this phase involves inference of phoneme sequence.

More precisely, the decoding phase consists of the following steps:

1. The orthographic transcription of the given word is parsed to extract the (context-independent) grapheme sequence. For example, word AREA is parsed as [A] [R] [E] [A].

[1] + denotes following context and - denotes preceding context. For instance, the lexical model of word CAT with single preceding and following context is [C+A] [C-A+T] [A-T].

Word	Actual pronunciation	Extracted pronunciation
WHEN+S	/w//eh//n//z/	/w//eh//n//z/
ANCHORAGE	/ae/ /ng/ /k/ /er/ /ih/ /jh/	/ae/ /ng/ /k/ /ch/ /ao/ /r/ /ih/ /jh/
ANY	/eh/ /n/ /iy/	/ae/ /n/ /iy/
CHOPPING	/ch/ /aa/ /p/ /ih/ /ng/	/ch/ /aa/ /p/ /iy/ /ng/

Table 1. Illustration of pronunciation models extracted for a few words using tri-grapheme KL-HMMs. By Actual pronunciation, we refer to the pronunciation given in the RM dictionary.

2. The context-independent grapheme sequence is then turned into context-dependent grapheme sequence. In simple terms, context expansion. For example, the sequence [A] [R] [E] [A] is changed into a sequence [A+R] [A-R+E] [R-E+A] [E-A] (in case of tri-grapheme context).
3. A word level HMM is then created by concatenating the HMMs of the context-dependent graphemes in the sequence. A sequence of phoneme posterior probabilities is then obtained by stacking the multinomial distributions of the states in the (left-to-right) order in which the states are connected. For example, in the case of context-dependent grapheme sequence [A+R] [A-R+E] [R-E+A] [E-A] the sequence of phoneme posterior probabilities starts with multinomial distribution of the first HMM state of [A+R] followed by the multinomial distribution of the second HMM state of [A+R], and so on till the multinomial distribution of the final HMM state of [E-A]. In other words, the grapheme KL-HMM acts like a generative model where each state (in the left-to-right sequence) generates a single phoneme posterior probability vector.
4. Finally, the phoneme posterior probabilities in the sequence are used as *local scores*, exactly like in the case of hybrid HMM/MLP system [7], and decoded by a fully ergodic HMM system (that connects all the D phonemes with a uniform transition probability matrix) to infer the phoneme sequence.

The proposed G2P approach has certain benefits some of which are inherited from KL-HMM system, such as

- The posterior feature estimator and the parameters of the KL-HMM system can be trained on independent data. As a result, the posterior feature estimator, i.e. the MLP can be trained using the data of resource rich languages/domains.
- KL-HMM system has fewer parameters, i.e. each emitting state is parameterized by a D dimensional multinomial distribution. This is particularly of interest when there is less amount of transcribed data or longer grapheme subword unit context models that may need to be trained.
- The search involved during decoding to infer the phoneme sequence is relatively simple.
- Though in this paper the focus is on phoneme, the approach could be extended to other units, such as syllable, automatically derived acoustic subword units.

4. EXPERIMENTAL STUDIES

To demonstrate the potential of the proposed approach, we consider a scenario where we have access to acoustic data from two different domains. For the first domain, we assume that we have prior linguistic resources, i.e. phoneme set and pronunciation dictionary. For the second domain, we assume that we do not have any prior linguistic resources, i.e. neither phoneme set nor pronunciation dictionary. However, we still would like to build a phoneme-based ASR system for the second domain.

We simulate this scenario by using Wall Street Journal (WSJ) corpus for the first domain, i.e. the MLP to extract posterior feature is trained on WSJ data. While, using DARPA Resource Management (RM) corpus for the second domain, i.e. the grapheme-based KL-HMM system is trained on RM and then phoneme-based pronunciation dictionary for RM task is generated using the grapheme KL-HMMs and the orthography of the words. More precisely,

- Training phase: From our previous grapheme-based ASR study [4], we selected the tri- and quint-grapheme KL-HMMs that were trained using the cost function based on local score SKL with posterior features estimated by WSJ MLP.
- Decoding phase: We generated two phoneme-based pronunciation dictionaries containing the 991 words of RM task (following the steps described earlier in Section 3). One dictionary generated using tri-grapheme models, and the other generated using quint-grapheme models. During decoding, each phoneme was modeled by a 3-state HMM. Note that multiple pronunciations could be extracted for each word using n-best decoding. However, in this study we only used 1-best decoding, i.e. single pronunciation model for each word.

For details about the setup of RM task and the WSJ MLP, the reader is referred to [4].

The main reason to design the experiment in this fashion was to have better control on the study as phoneme sets can vary from task to task. In addition, there are reference ASR systems to which the ASR studies conducted in this paper can be fairly compared. Thus, helping us in gaining better understanding. In other words, we have access to the phoneme-based lexicon obtained from UNISYN dictionary for both WSJ task and RM task. So, the extracted pronunciation models could be analyzed not only at ASR system level performance but also at phoneme error level and word error level.

Table 1 shows the extracted pronunciation models of a few words by the proposed approach along with their respective pronunciation in the RM dictionary.

4.1. Analysis at Phoneme Level and Word Level

We compared the pronunciation models extracted for each word using the tri-grapheme KL-HMMs with the respective pronunciation in the RM dictionary. We performed similar comparison for pronunciation models extracted using quint-grapheme KL-HMMs. Table 2 presents this comparison in terms of phoneme error rate (PER) and word error rate (WER).

It can be seen that the pronunciation models extracted using quint-graphemes are more closer to the actual pronunciation when compared to pronunciation models extracted using tri-graphemes.

	PER	WER
Tri-grapheme	20.1%	68.8%
Quint-grapheme	15.9%	60.4%

Table 2. Comparing the extracted pronunciation models with actual pronunciations in terms of PER and WER. Quint-grapheme denotes the case where pronunciation lexicon is extracted using quint-grapheme models. Tri-grapheme denotes the case where pronunciation lexicon is extracted using tri-grapheme models.

This is further illustrated by Table 3 which shows the distribution of words in terms of Levenshtein distance (between the extracted pronunciation and actual pronunciation). It is interesting to note that

Levenshtein distance	Quint-grapheme	Tri-grapheme
0	392	309
1	376	373
2	166	206
3	45	71
4	6	26
5	4	4
6	1	1
7	1	1

Table 3. Distribution of words in terms of Levenshtein distance between the extracted pronunciation and actual pronunciation. Quint-grapheme denotes the case where pronunciation lexicon is extracted using quint-grapheme models. Tri-grapheme denotes the case where pronunciation lexicon is extracted using tri-grapheme models.

about 77.5% of the words (in case of Quint-grapheme) and 68.8% of the words (in case of Tri-grapheme) lie with in the Levenshtein distance of one.

4.2. Analysis at ASR Performance Level

We built separate context-dependent phoneme (more precisely triphone) based ASR systems using the different pronunciation lexicons. We refer to the system using the pronunciations extracted with quint-grapheme models as *Quint-grapheme* and the system using the pronunciations extracted with tri-grapheme models as *Tri-grapheme*. We trained two types of ASR system, namely, standard HMM/Gaussian mixture model (HMM/GMM) system and KL-HMM system. In the case of KL-HMM system, we used the same WSJ MLP for posterior feature extraction and trained the system by optimizing the cost function based on local score SKL. We compare the performance of systems *Quint-grapheme* and *Tri-grapheme* to ASR system that is trained using the original RM dictionary, referred to as System *Baseline*. Table 4 presents the performance of different systems on the RM evaluation set of 1200 utterances in terms of WER.

It is interesting to note that despite only generating correct pronunciation for 39.6% words and 31.2% words systems *Quint-grapheme* and *Tri-grapheme* achieve a performance that is close to the *Baseline* system, respectively. Thus, demonstrating the potential of the proposed approach.

The ASR studies also show the superiority of quint-grapheme models over tri-grapheme models for extracting pronunciation model. It can be also observed that the ASR performance differences

System	HMM/GMM	KL-HMM
Baseline	5.7%	4.7%
Tri-grapheme	7.8%	5.9%
Quint-grapheme	7.1%	5.4%

Table 4. Performance of different ASR systems expressed in terms of WER.

with respect to system *Baseline* in the case of KL-HMM system is lower than HMM/GMM system. This may be due to matched condition effect, as the data used for pronunciation model extraction and ASR system training is same.

5. CONCLUSION

In this paper, we presented a novel acoustic data driven grapheme-to-phoneme conversion approach using KL-HMM. The main strength of the proposed G2P approach is that the relationship between phoneme and grapheme is learned through acoustics. This strength could be exploited to

- develop a lexicon from scratch given some transcribed acoustic data from the target language/domain and acoustic and linguistic resources of other languages/domains. Our experimental studies tried to demonstrate this. This approach could be further exploited for rapid development of ASR and TTS systems for languages/domains that have fewer resources, and for tasks such as proper name recognition where one usually has to take multiple languages into account when extracting pronunciation models.
- generate pronunciation variants (taking the acoustic realization aspects into account). Our experimental studies also suggests it as in the generated lexicons maximum 39.6% of the words had correct pronunciations, but the systems were still able to achieve performance that was not too far from the baseline system. Thus, the proposed G2P approach is also interesting for languages/domains that have prior linguistic resources.

Our future work includes a) improving the pronunciation variant/model extraction by using mixed context grapheme models, phoneme n-gram models, n-best list, and confidence measures, b) evaluation on unseen words, i.e. the words that are not seen during KL-HMM training and low resource ASR task.

6. REFERENCES

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

[2] M. Bisani and H. Ney, "Joint-sequence models for grapheme-to-phoneme conversion," *Speech Communication*, vol. 50, no. 5, pp. 434–451, May 2008.

[3] D. Wang and King. S, "Letter-to-sound pronunciation prediction using conditional random fields," *IEEE Signal Processing Letters*, vol. 18, no. 2, pp. 122–125, 2011.

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

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

[6] G. Aradilla, *Acoustic Models for Posterior Features in Speech Recognition*, Ph.D. thesis, Ecole Polytechnique Fédérale de Lausanne, 2008.

[7] N. Morgan and H. Bourlard, "Continuous speech recognition: an introduction to the hybrid HMM/connectionist approach," *IEEE Signal Processing Magazine*, vol. 12, pp. 24–42, May 1995.