GRAPHEME-BASED AUTOMATIC SPEECH RECOGNITION USING KL-HMM

Mathew Magimai.-Doss[1], *Ramya Rasipuram*[1,2], *Guillermo Aradilla*[1], *Hervé Bourlard*[1,2]

[1]Idiap Research Institute, Martigny, Switzerland
[2]Ecole Polytechnique Fédérale, Lausanne (EPFL), Switzerland
{mathew@idiap.ch, ramya.rasipuram@idiap.ch, guillermo.aradilla@gmail.com, bourlard@idiap.ch}

ABSTRACT

The state-of-the-art automatic speech recognition (ASR) systems typically use phonemes as subword units. In this work, we present a novel grapheme-based ASR system that jointly models phoneme and grapheme information using Kullback-Leibler divergence-based HMM system (KL-HMM). More specifically, the underlying subword unit models are grapheme units and the phonetic information is captured through phoneme posterior probabilities (referred as *posterior features*) estimated using a multilayer perceptron (MLP). We investigate the proposed approach for ASR on English language, where the correspondence between phoneme and grapheme is weak. In particular, we investigate the effect of contextual modeling on grapheme-based KL-HMM system and the use of MLP trained on auxiliary data. Experiments on DARPA Resource Management corpus have shown that the grapheme-based ASR system modeling longer subword unit context can achieve same performance as phoneme-based ASR system, irrespective of the data on which MLP is trained.

Index Terms— Automatic speech recognition, Graphemes, Phonemes, Kullback-Leibler divergence based hidden Markov model, Posterior features, Multilayer perceptron

1. INTRODUCTION

Standard hidden Markov model (HMM) based automatic speech recognition (ASR) systems typically use cepstral coefficients as feature vectors and phonemes/phones as subword units. The emission distribution is either modeled by Gaussian mixture models (GMMs) or multilayer perceptron (MLP).

In a more recent work, Kullback-Leibler divergence based HMM (KL-HMM) system was proposed [1], where the phoneme class conditional probabilities, referred as *posterior features*, is directly used as feature observation and each emission state is modeled by a multinomial distribution. The emission score is estimated as the KL-divergence between posterior feature observation and state multinomial distribution.

This paper proposes a novel grapheme[1] based ASR system in the framework of KL-HMM. It can be seen as a system where, first a relationship between the acoustic feature (e.g., cepstral features) and phoneme is modeled through a posterior feature estimator (more precisely, MLP). Then a soft correspondence between phonemes and graphemes is modeled/learned through the state multinomial distribution of KL-HMM system. Through ASR studies on English language, where the correspondence between phoneme and grapheme is weak, using DARPA Resource Management (RM) corpus, we show that in KL-HMM framework the grapheme-based ASR system can yield same performance as phoneme-based ASR system.

Section 2 givens an overview on grapheme-based ASR. Section 3 introduces KL-HMM based acoustic modeling and discusses briefly its potential use in the context of grapheme-based ASR. Section 4 presents the experimental setup followed by Section 5 which provides insight into the effect of contextual modeling of grapheme units in the KL-HMM framework. Section 6 presents the ASR studies. Finally, in Section 7 we conclude.

2. GRAPHEME-BASED ASR

As mentioned earlier, standard HMM-based ASR systems typically use phonemes as subword units. However, there has been a constant interest in using graphemes as subword units [2, 3, 4, 5] for reasons, such as: a) Ease to create lexicon, i.e., pronunciation of words can be derived from orthographic transcription. In case, of phonemes it is usually a semi-automatic process. This advantage particularly comes handy for tasks, such as spoken term detection, where the query term (or word) can be a word that is not present in the pronunciation lexicon and letter-to-sound rules to generate pronunciation may not be the best [6], b) Grapheme subword units, such as Roman alphabets could be shared across many languages. This gives the possibility of sharing data resources from different languages when training acoustic models [3, 7] as well as to port efficiently acoustic models trained on one language to other languages [3], and c) ASR performance could be improved by using both phoneme and grapheme subword units [4, 8].

Despite the above mentioned advantages modeling grapheme subword is not a trivial task. One of the main reason for this is that standard cepstral features which capture the envelop of the magnitude spectrum of short-term signal mainly depict characteristics of phonemes, and the correspondence between phonemes and graphemes depend upon language. In languages, such as English the correspondence between phonemes and graphemes is weak. i.e. graphemes can map to different phonemes. For instance, grapheme [C] maps to phoneme /k/ in *CAT*, to phoneme /s/ in word *CITE*, and to phoneme /ch/ in word *CHURCH*. While, in languages, such as Finnish, Spanish the correspondence is strong, i.e. close to one-to-one mapping.

Figure 1(a) illustrates a graphical model representation of grapheme-based ASR system. In literature, most of the studies on grapheme-based ASR have focussed on contextual modeling of grapheme subword units [2, 3, 4, 5], where the implicit assumption is

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 colleague John Dines for fruitful discussions.

[1]Grapheme is the fundamental unit in a written language, e.g. English alphabets.

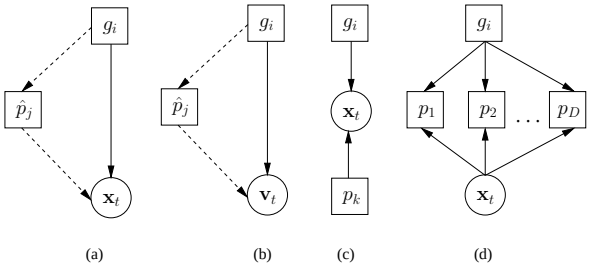

Fig. 1. Graphical model representations for different grapheme-based ASR systems. $\mathbf{x}_t$ represents acoustic feature, g_i represents grapheme subword i, p_k represents phoneme subword k, $\mathbf{v}_t$ represents Tandem feature, $p_1, p_2, \cdots, p_D$ in (d) represent phonemes. D is number of phonemes. The dotted line in (a) and (b) indicates that with contextual modeling of grapheme subword g_i may implicitly map to phoneme subword $\hat{p}_j$.

that context-dependent grapheme subword unit may map to unique phoneme, as illustrated by dotted line in Figure 1(a). These studies have shown that such systems can achieve performance comparable to phoneme-based ASR system for languages that have stronger correspondence between graphemes and phonemes (e.g., Dutch, Spanish, German), and poor performance for languages that have weaker correspondence (e.g., English). In [4], it was found that for languages like English, where the correspondence between grapheme and phoneme is weak, use of Tandem/MLP features can help in bridging the gap between the performance of phoneme-based ASR system and grapheme-based ASR system. The grapheme-based ASR system using MLP features is illustrated in 1(b). In addition to modeling only grapheme subword units, there have been studies where the ASR system uses both phone and grapheme subword units [8, 4]. Figure 1(c) illustrates a phoneme-grapheme system [8], where during training grapheme and phone subword units are jointly modeled, and during recognition decoding is performed using either one subword unit or both. It has been found that such systems could improve ASR performance.

3. KL-HMM ACOUSTIC MODELING

In a recent work, the use of posterior probabilities of phonemes directly as feature observation was proposed for HMM-based ASR [1]. As depicted in Figure 2, in this system each state i of the HMM is characterized by a reference multinomial distribution $\mathbf{y}_i = [y_i^1, \cdots, y_i^D]^T$, where D is the number of phonemes. Given an estimate of phoneme posterior feature vector $\mathbf{z}_t$ at time frame t,

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

the local score at each HMM state is estimated as Kullback-Leibler divergence between $\mathbf{y}_i$ and $\mathbf{z}_t$, i.e.,

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

where, $\mathbf{x}_t$ is the acoustic feature (such as cepstral feature) at time frame t, $\mathbf{y}_i$ is the reference distribution, $\mathbf{z}_t$ is the test distribution, and $p_1, p_2, \cdots, p_D$ are the phonemes. We denote this local score as KL.

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

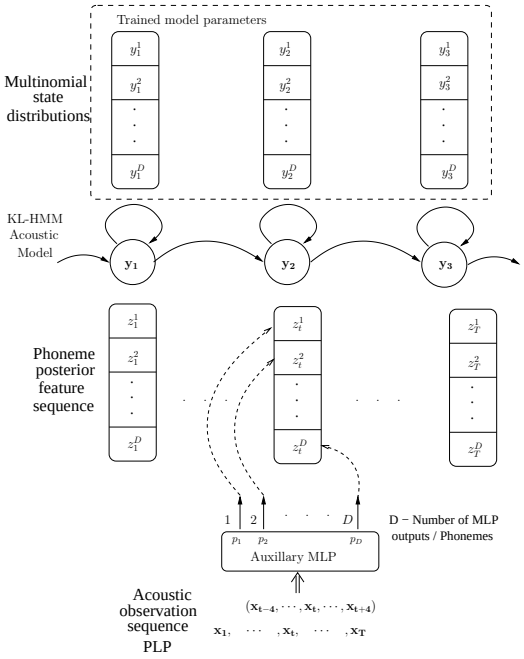

Fig. 2. Illustration of KL-HMM acoustic model

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 parameters of the HMM states (i.e., multinomial distributions) are trained using Viterbi expectation maximization algorithm with one of the local scores as the cost function. The decoding is performed using standard Viterbi decoder.

KL-HMM establishes a framework that unifies different types of acoustic models, such as discrete HMM and HMM/MLP through the use of different local scores. For instance, the system using local score KL can be linked to hybrid HMM/MLP system, and the system using local RKL can be linked to discrete HMM system. For more, details and additional interpretations the reader is referred to [9, Chapter 6].

In the context of grapheme-based ASR, KL-HMM provides certain advantages which can be potentially exploited along with the benefits of using grapheme subword units described earlier in Section 2, such as

1. Fewer parameters: In KL-HMM, fewer parameters, i.e. a D dimensional multinomial distribution per state needs to be trained. This can be effectively exploited to model longer grapheme subword contexts.
2. Choice of posterior feature space: The posterior feature can be monolingual phoneme class conditional probabilities, multilingual/universal phoneme class conditional probabilities, or articulatory features [10].
3. Choice of posterior feature estimator: In this work, we use MLP to estimate posterior features which, in addition to directly estimating a posteriori probabilities of output classes,

also provides robustness towards undesirable variation, such as speaker and environment. However, one could use other estimators, such as GMMs.

4. Transfer learning: The posterior feature estimator could be trained on an auxiliary corpus. The use of universal phoneme posterior features or articulatory features also allows the flexibility to use data from multiple languages. For transfer learning, both MLPs and GMMs [11] could be used.

Figure 1(d) illustrates the proposed grapheme-based ASR system in KL-HMM framework, where RKL is used as the local score and the posterior features consists of phoneme class conditional probabilities. It can be observed that a part of this system can be interpreted as an acoustic data-driven grapheme-to-phoneme converter.

4. EXPERIMENTAL SETUP

In this paper, to study this approach, as a first step we focus our attention towards two aspects, i.e. modeling longer grapheme subword units and transfer learning using cross domain data for English language ASR.

We use DARPA Resource Management (RM) corpus for speaker-independent speech recognition studies. The RM corpus consists of read queries on the status of Naval resources [12]. The setup is exactly same as reported in [4]. The training set and development set consists of 3'990 utterances spoken by 109 speakers corresponding to approximately 3.8 hours of speech data. The test set contains 1'200 utterances amounting to 1.1 hours in total. The test set is completely covered by a word pair grammar included in the task specification.

The lexicon consists of 991 words. The phoneme-based lexicon was obtained from UNISYN dictionary. There are 42 context-independent phonemes including silence. The grapheme-based dictionary was transcribed using 29 context-independent graphemes (including silence, symbols).

We use *off-the-shelf* MLPs trained on RM corpus [4] and Wall Street Journal (WSJ) corpus [1] to classify 45 context-independent phonemes for estimating posterior features. The WSJ MLP is used to study the transfer learning aspect. Both these MLPs were trained with 39 dimensional perceptual linear prediction cepstral coefficients. For the context-dependent studies, we only model word internal context. Each sub-word unit is modeled by a 3 state left-to-right HMM. The tuning parameters, such as insertion penalty, language scaling factor were tuned on the development data. For each of the subword units (i.e., phoneme and grapheme), we built three systems:

1. *mono*: System that only models context-independent subword units.
2. *tri*: System that models context-dependent subword units with single preceding subword and single following subword as context (2269 models for phoneme and 1912 for grapheme).
3. *quint*: System that models context dependent subword units with two preceding subwords and two following subwords as context (3982 models for phoneme and 4112 models for grapheme).

5. ANALYSIS OF KL-HMM MODELS

In this section, we provide insight into the grapheme models that are estimated for System *mono*, System *tri*, and System *quint* when trained using WSJ MLP posterior features with SKL as local score. For this purpose, we consider models of consonant grapheme [C] and vowel grapheme [A].

5.1. Context-independent subword unit modeling

Table 1 shows, the first two components of the multinomial state distributions of the grapheme models [C] and [A], arranged in descending order, along with the corresponding phoneme label and phoneme posterior probability value. It can be observed that different states of the multinomial state distributions capture different phonemes. Grapheme model [C] (representing consonant grapheme) captures three phonemes /k/, /ch/ and /s/ in three different states and the grapheme model [A] (representing vowel consonant) captures /ae/, /ey/, /ax/, /eh/ in different states. In other words, the states capture gross phoneme information.

Model: [C]	State: 1	State: 2	State: 3
1st Max	/s/ (0.6)	/ch/ (0.3)	/k/ (0.9)
2nd Max	/z/ (0.1)	/t/ (0.3)	/t/ (0.02)
Model: [A]	**State: 1**	**State: 2**	**State: 3**
1st Max	/ae/ (0.64)	/ey/ (0.54)	/ax/ (0.32)
2nd Max	/eh/ (0.13)	/ax/ (0.08)	/ae/ (0.1)

Table 1. The first two components of the multinomial state distributions of the models [C] and [A], shown along with the corresponding phoneme label and phoneme posterior probability value

5.2. Context-dependent subword unit modeling

Table 2 illustrates that, by modeling single preceding and following context for the grapheme [C] ambiguity in the model is resolved for three different contexts graphemes [b-C+A], [b-C+E] and [b-C+H] (where 'b' refers to begin of the word). While, for graphemes like [A], even single preceding and following context modeling is not sufficient to capture the relevant phoneme information. The states of context-dependent grapheme subword model [b-A+R], capture more than one phoneme (/ey/, /aa/, /axr/ and /ae/).

Table 2 shows that the vowel grapheme model [b-A+R*E] representing quint-graph context model for the grapheme [A] resolves the ambiguity. We can observe that the model dominantly captures phoneme /aa/. Also, the multinomial state distribution of third state seems to model the transition information, i.e. transition to phoneme /r/.

Overall, this suggests that the modeling of subword context longer than usual single preceding and single following may yield a grapheme-based ASR system that behaves similar to a phoneme-based system. Our ASR studies presented in the following section demonstrates this.

6. RESULTS

Table 3 presents the performance, in terms of word error rate (WER), of phoneme-based ASR system and grapheme-based ASR system on the test set for three different local scores KL, RKL and SKL, when MLP trained on RM corpus is used to estimate posterior features. Similarly, Table 4 reports the performances, when MLP trained on WSJ is used to estimate posterior features.

As it can be observed, irrespective of the MLP is used, the trends are same. In the case of phonemes, System *tri* yields the best performance for all the local scores. While in the case of graphemes, System *quint* yields the best performance for all the local scores. The reason behind this is that grapheme needs more context to disambiguate between phonemes (see Section 5). Neverthe-

Model: [b-C+A]	State: 1	State: 2	State: 3
1st Max	/k/ (0.6)	/k/ (0.9)	/k/ (0.9)
2nd Max	/t/ (0.1)	/g/ (0.03)	/t/ (0.03)
Model: [b-C+E]	State: 1	State: 2	State: 3
1st Max	/s/ (0.5)	/s/ (0.8)	/s/ (0.9)
2nd Max	/z/ (0.4)	/sh/ (0.07)	/z/ (0.05)
Model: [b-C+H]	State: 1	State: 2	State: 3
1st Max	/t/ (0.5)	/ch/ (0.8)	/ch/ (0.6)
2nd Max	/ch/ (0.2)	/jh/ (0.1)	/t/ (0.3)
Model: [b-A+R]	State: 1	State: 2	State: 3
1st Max	sil (0.49)	/ey/ (0.21)	/aa/ (0.30)
2nd Max	/aa/ (0.14)	/ae/ (0.20)	/axr/ (0.23)
Model: [b-A+R*E]	State: 1	State: 2	State: 3
1st Max	/aa/ (0.24)	/aa/ (0.74)	/aa/ (0.24)
2nd Max	/t/ (0.18)	/ao/ (0.11)	/r/ (0.24)

Table 2. The first two components of the multinomial state distributions of the models [b-C+A], [b-C+E], [b-C+H], [b-A+R] and [b-A+R*E] ('b' refers to begin of the word) arranged in descending order, shown along with the corresponding phoneme label and phoneme posterior probability value

System	Phoneme			Grapheme		
	Local score			Local score		
	KL	RKL	SKL	KL	RKL	SKL
mono	7.1	8.0	7.1	42.1	25.8	32.9
tri	5.5	5.9	**5.1**	7.7	6.5	5.9
quint	5.4	5.8	5.2	6.1	5.7	**5.2**

Table 3. Word error rate expressed in % using phoneme and grapheme as subword units in KL-HMM system for three different local scores KL, RKL, and SKL. The posterior feature is estimated by MLP trained on RM corpus. Boldface indicates the best system for each of the subword units.

less, the phoneme-based and grapheme-based systems achieve the similar/same performance. Local score SKL yields the best performance across all the phoneme-based systems. However in the case of grapheme, SKL yields the best performance only when context is modeled. In case of phoneme, System *quint* yields same or poor performance compared to System *tri*. The poor performance could be due to redundant models introduced when increasing context or insufficient data available to model all the contexts. While, in the case of grapheme, the improvement could be attributed to disambiguation provided by the increased context.

As reported in [4] on exactly same setup, HMM/GMM system which is equivalent (in terms of context modeling) to System *tri* in Table 3 achieves a performance of 5.7% WER for phoneme subword unit and 7.3% WER for grapheme subword unit using PLP features, and 5.7% WER for phoneme subword unit and 6.3% for grapheme subword using Tandem features (as stand alone features). It can be seen that KL-HMM system with local score SKL is performing better than HMM/GMM system for both phoneme and grapheme subword units. Training of HMM/GMM system that is equivalent to System *quint* resulted in data sparsity issues.

Finally, the best system is obtained when MLP trained on WSJ corpus is used to estimate posterior features. This can be attributed to the fact that WSJ MLP is trained on more data ($\approx$ 80hrs). This suggests that the system could benefit from MLPs trained on large auxiliary data.

System	Phoneme			Grapheme		
	Local score			Local score		
	KL	RKL	SKL	KL	RKL	SKL
mono	7.4	8.0	6.9	39.9	25.3	32.4
tri	5.1	5.0	**4.7**	7.1	6.0	6.0
quint	5.8	5.4	**4.7**	5.8	5.7	**4.7**

Table 4. Word error rate expressed in % using phoneme and grapheme as subword units in KL-HMM system for three different local scores KL, RKL, and SKL. The posterior feature is estimated by MLP trained on WSJ corpus. Boldface indicates the best system for each of the subword units.

7. CONCLUSIONS

In the framework of KL-HMM, we proposed a novel grapheme-based ASR system in which the acoustic-phonetic information is modeled through a posterior feature estimator, such as MLP and the relationship between the grapheme and phoneme is captured via the state multinomial distributions. English language ASR studies on RM corpus showed that this system, which can model longer subword unit context efficiently by exploiting the flexibility of KL-HMM that it has fewer number of parameters to train, can achieve same performance as phoneme-based ASR system (in spite of poor correpondence between graphemes and phonemes in English and irrespective of the data on which MLP is trained). Furthermore, it is interesting to note that grapheme-based system exploits the increased capacity of the KL-HMM model (through longer context modeling) better than a phoneme-based system. In future work, we will extend the ASR studies to relatively more difficult task, more specifically to conversational speech recognition and non-native speech recognition, and in this context will also explore other posterior feature representation, such as universal phoneme posterior probabilities estimated by training a multilingual MLP.

8. REFERENCES

[1] G. Aradilla et al., "Using KL-Based Acoustic Models in a Large Vocabulary Recognition Task ," in *Proc. of Interspeech*, 2008.

[2] S. Kanthak and H. Ney, "Context-dependent acoustic modeling using graphemes for large vocabulary speech recognition," in *Proc. of ICASSP*, 2002, pp. 845–848.

[3] M. Killer et al., "Grapheme based speech recognition," in *Proc. of Eurospeech*, 2003.

[4] J. Dines and M. Magimai-Doss, "A study of phoneme and grapheme based context-dependent ASR systems," in *MLMI 2007, Lecture Notes in Computer Science No. 4892*, 2008.

[5] Y-H Sung et al., "Revisiting graphemes with increased amount of data," in *Proc. of ICASSP*, 2009.

[6] D. Wang et al., "A comparison of phone and grapheme-based spoken term detection," in *Proc. ICASSP*, 2009.

[7] S. Kanthak and H. Ney, "Multilingual acoustic modeling using graphemes," 2003, pp. 1145–1148.

[8] M. Magimai.-Doss et al., "Joint decoding for phoneme-grapheme continuous speech recognition," in *Proc. of ICASSP*, 2004.

[9] G. Aradilla, *Acoustic Models for Posterior Features in Speech Recognition*, Ph.D. thesis, EPFL, Switzerland, 2008.

[10] R. Rasipuram and M. Magimai.-Doss, "Integrating Articulatory Features using Kullback-Leibler Divergence based Acoustic Model for Phoneme Recognition," in *Proc. of ICASSP*, 2011.

[11] D. Povey et al., "The subspace Gaussian mixture model-A structured model for speech recognition," *Computer Speech and Language*, 2011.

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