IDIAP RESEARCH REPORT

ON JOINT MODELLING OF GRAPHEME AND PHONEME INFORMATION USING KL-HMM FOR ASR

Mathew Magimai.-Doss Guillermo Aradilla
Hervé Bourlard

Idiap-RR-24-2009

SEPTEMBER 2009

Centre du Parc, Rue Marconi 19, P.O. Box 592, CH - 1920 Martigny
T +41 27 721 77 11 F +41 27 721 77 12 info@idiap.ch www.idiap.ch

On Joint Modelling of Grapheme and Phoneme Information using KL-HMM for ASR

Mathew Magimai.-Doss [†1], Guillermo Aradilla [†2], Hervé Bourlard [†,‡3]

[†] *Idiap Research Institute, Martigny, Switzerland*
[‡] *Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland*
[1] mathew@idiap.ch, [2] guillermo.aradilla@gmail.com, [3] herve.bourlard@idiap.ch

***Abstract*—In this paper, we propose a simple approach to jointly model both grapheme and phoneme information using Kullback-Leibler divergence based HMM (KL-HMM) system. More specifically, graphemes are used as subword units and phoneme posterior probabilities estimated at output of multilayer perceptron are used as observation feature vector. Through preliminary studies on DARPA Resource Management corpus it is shown that although the proposed approach yield lower performance compared to KL-HMM system using phoneme as subword units, this gap in the performance can be bridged via temporal modelling at the observation feature vector level and contextual modelling of early tagged contextual graphemes.**

I. Introduction

State-of-the-art HMM-based automatic speech recognition (ASR) systems commonly use phoneme as subword unit. More recently, there has been growing interest in using directly the grapheme[1] (orthographic) transcription of the word (without explicit lexical phonetic level modelling). There are three main advantages in using grapheme as subword units. Firstly, dictionary generation is easy. Secondly, grapheme subword units can be shared across different languages [2] thus it may help in porting ASR system trained on languages that have large resources to languages that have resource constraints. Thirdly, unlike the phoneme-based ASR system where a word can have pronunciation variants the word representation is unique (orthographic forms of words do not vary between speakers). While the use of grapheme as subword unit limits the variability at the word representation level, the link between the acoustic waveform becomes weaker (depending on the language), as the standard acoustic features extracted from the spectrum of the speech signal characterize phonemes. For instance, in languages such as Finnish and Spanish there is good one-to-one unique correspondence between phoneme and grapheme, so when using grapheme as subword units the link between acoustic waveform can be as strong as phoneme. However, it is not the same case for languages such as English where the correspondence between phoneme and grapheme is weak. For instance, alphabet [C] can correspond to phonemes /k/ or /ch/. Thus, there has been emphasis on integrating phoneme information to yield better grapheme-based ASR system. For instance,

- In [1], the grapheme-based ASR system was trained such that both grapheme-to-phoneme mapping and HMM state tying are optimized with a single phonetically motivated decision tree. In this approach, performance comparable to phoneme-based ASR system was observed for languages such as Dutch and German but for English the performance was lower than the phoneme-based ASR system.
- Investigating different efficient state tying methods [2], [3], such as decision trees with question set that incorporate relation between phoneme and grapheme, or question set containing simple preceding and following context (singleton question set), or using bottom-up clustering for context-dependent grapheme-based ASR system was studied for different languages.
- Joint modelling of phoneme and grapheme information by training multilayer perceptron (MLP) that classifies both phoneme and grapheme [4]. During decoding, either phoneme or grapheme information is hidden by marginilizing the posterior distribution and decoding is performed with grapheme subword units or phoneme subword units, respectively or joint decoding in both spaces, i.e., phoneme subword units and grapheme subword units. The latter approach consistently yielding better performance.
- In [5], it was shown that using tandem features which can carry information more specific to the speech sound, and less susceptible to speaker and environment, the gap between phoneme-based ASR system and grapheme-based ASR system can be effectively reduced.
- Introducing grapheme context information through early tagging which can avoid capturing of gross distributions when modelling context-independent grapheme units and reducing ambiguity as well [6].

Apart from the above mentioned studies where the motivation has been to integrate phoneme information, in a more recent study it was shown that performance similar to phoneme-based ASR system can be achieved with grapheme-based ASR system "by sufficient context modelling and using enough

[1]Grapheme is a written symbol that is used to represent words.

[2]Roman alphabet is the most widely used. It covers languages from most of the European nations, all nations of the America and Oceania, most of the African nations, and a few Asian nations as well.

training data" [7].

In this paper, we propose a simple approach to jointly model both grapheme and phoneme information using the flexibility of recently proposed Kullback-Leibler divergence-based HMM system[8], [9]. More specifically, grapheme is used as subword unit (i.e., pronunciation of each word is represented in terms of its orthographic transcription) and, the phoneme information is modelled through phoneme posterior probabilities estimated at each time frame using an MLP, which are used as observation feature vectors.

We demonstrate the viability of this approach for English language through preliminary studies on DARPA Resource Management (RM) corpus. Our studies using an "off-the-shelf" MLP trained on RM corpus show that this approach can achieve performance closer to phoneme-based ASR system with still a gap of about 1% absolute word error rate. However, we also show that this gap can be reduced by improving contextual modelling either at feature level (i.e., by using a sequence of phoneme posterior probabilities as observation vector) or by improving contextual modelling by early tagging as proposed in [6]. Further more, we also show that these findings generalize to the case where the MLP is trained on an entirely different database/corpus.

The rest of the paper is organized as follows. In Section II, we briefly summarize the KL-HMM system and motivate the proposed approach in Section III. Section IV presents the experimental studies and in Section VI we conclude.

II. Kullback-Leibler Divergence Based HMM

In KL-HMM, the observation feature vector is a posterior probability vector $\mathbf{z} = [z(1), \cdots z(d), \cdots z(D)]^T$ of dimension D, and emission distribution of a state i is modelled by a multinomial distribution $\mathbf{y}^i = [y^i(1), \cdots y^i(d), \cdots y^i(D)]^T$. The posterior probability vector can be for instance the phoneme posteriors estimated by an MLP. The local score is estimated using KL-divergence. KL-divergence being an asymmetric measure the local score can be estimated in different ways, such as,

$$KL(\mathbf{y}^i, \mathbf{z}) = \sum_{d=1}^{D} y^i(d) \cdot \log(\frac{y^i(d)}{z(d)}) \quad (1)$$

$$RKL(\mathbf{z}, \mathbf{y}^i) = \sum_{d=1}^{D} z(d) \cdot \log(\frac{z(d)}{y^i(d)}) \quad (2)$$

$$SKL(\mathbf{y}^i, \mathbf{z}) = \frac{1}{2}[KL(\mathbf{y}^i, \mathbf{z}) + RKL(\mathbf{z}, \mathbf{y}^i)] \quad (3)$$

Table I compares the capabilities of KL-HMM with different types of state-of-the-art HMM-based ASR systems.

In our earlier studies [9], [10] using phoneme as subword units it was shown that (a) KL-HMM is simpler (fewer number of parameters) and flexible (it relaxes the tying between each MLP output unit and HMM-state which otherwise limits the capability of HMM/MLP) (b) KL-HMM can be trained with a fewer amount of training data using standard embedded-Viterbi-like framework, (c) KL-HMM can achieve performance comparable to state-of-the-art ASR systems, and (d) Depending upon the type of local score used $KL(\mathbf{y}^i, \mathbf{z})$ or $RKL(\mathbf{y}^i, \mathbf{z})$ if the reference distribution is reduced to delta distribution then KL-HMM is equivalent to hybrid HMM/MLP or discrete HMM, respectively.

TABLE I
Comparison of capabilities of KL-HMM with standard HMM-based systems. Notations: GMM - Gaussian mixture models, EM - Expectation maximization, EV - Embedded Viterbi, CI - Context-independent, CD - Context-dependent

System	HMM/GMM	Hybrid HMM/ MLP	Tandem	KL-HMM
Feature	Spectral-based	Spectral-based	Processed posterior Probabilities	Posterior probabilities
Emission distribution	GMM	MLP	GMM	Multinomial distribution
Local score	Likelihood	Scaled-likelihood	Likelihood	KL-divergence
Training method	EM/EV	EV/EM	EM/EV	EV / EM-like (possible)
Decoding	Viterbi	Viterbi	Viterbi	Viterbi
Subword modelling	CI and CD	CI	CI and CD	CI and CD

III. Motivation

The use of posterior probability estimates of elementary speech sound units such as phoneme as feature observation in KL-HMM provides the flexibility of choosing the subword unit level representation. In this work, we are interested in the case where the subword unit is grapheme. In doing so, dictionary generation is easy and each word has a single pronunciation and, as the feature space is phoneme posterior probabilities we can expect to keep the relation to acoustics intact via contextual modelling.

To be more specific, unlike earlier studies in grapheme-based ASR where there has been more emphasis to model the relation between grapheme subword unit representation and acoustic feature directly[2]. This can be difficult depending upon the language and also for the reason that acoustic features can be susceptible to undesirable variabilities (e.g., speaker variation, channel variation). Here in the proposed approach, this task is split into two independent steps, a phoneme posterior estimator that can learn the relation to acoustics in a better way, and KL-HMM that learns the relation between observed phoneme evidence and grapheme subword units. Furthermore, the use of discriminative classifiers such as MLP to estimate posterior features helps in reducing the effect of undesirable variabilities.

There is also flexibility in the way one would like to model phoneme information. For instance, the observation feature space of KL-HMM can be sub-phonemic posterior feature estimated using MLP [11], articulatory feature posterior prob-

abilities estimated using MLP [12], using universal phoneme set (e.g. WorldBet), or a combination of them.

It is known through letter-to-sound rule studies that to map a character in a particular word to a phoneme variable number of neighboring grapheme context information may be required. Given this, it is fairly easy to see that the conventional approach of modelling a fixed context in grapheme based ASR system such as modelling one preceding and one following context similar to triphone modelling can bring the subword representation closer to phoneme by reducing ambiguity but this model may not be as powerful as conventional triphone model in phoneme-based ASR system. Modelling fixed larger grapheme context may not necessarily solve this problem. In KL-HMM, the contextual modelling in grapheme-based ASR can be improved by for instance (a) using a time sequence of phoneme posterior probability estimates as feature observation, (b) modelling larger contexts at subword unit level using early tagging method [6], or (c) both. The methods (a)-(c) can help in striking a good balance between number of models and size of feature space. To this end, the ability to train KL-HMM with fewer amount of training data for reasons such as posterior features being linearly separable [11], having lesser variability can also be put to better use. We demonstrate this by investigating the first two methods i.e. (a) and (b) in this paper.

Though the above discussion has mainly focussed on the use of both grapheme and phoneme information, it is important to note that some of the above discussed methods/approaches are applicable to phoneme-based ASR system system as well.

IV. Experiments

We study the proposed approach to model jointly the phoneme and grapheme information using KL-HMM for English language, which has a weaker correspondence between phoneme and grapheme. We perform ASR studies on DARPA RM corpus. The RM corpus consists of read queries on the status of Naval resources [13]. The task is artificial in many aspects such as speech type, range of vocabulary and grammatical constraint. The training set consists of 3,990 utterances spoken by 109 speakers corresponding to approximately 3.8 hours of speech. Of this, we use 2,880 utterances for training and 1,100 for cross validation and development. 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 which is used for recognition.

We use an "off-the-shelf" MLP originally trained on RM corpus and used for an earlier study. For more details about the setup and state-of-the-art results reported using this MLP the reader may refer to [5].

It can be observed that for KL-HMM we need an already trained phoneme posterior feature estimator. This can be trained on the intended task. In such a case, a question can arise i.e., if we need to train a phoneme posterior estimator than why not simply use phoneme based ASR system. Moreover, there may not be always sufficient resources to train a reliable posterior feature estimator on the intended task. So, it may be better to have a posterior feature estimator trained on a large auxiliary database/corpus. In such a case, this may possibly lead to the problem of mismatched conditions. So we can possibly ask, what is the effect of using a posterior feature estimator trained on a different database on the performance of the proposed system. We address this issue by using another "off-the-shelf" MLP which was originally trained on Wall Street Journal (WSJ) corpus for phoneme-based KL-HMM studies reported in [9] (with roughly 80 hours of speech).

Both the MLPs i.e. the one trained on RM corpus (RM-MLP) and the one trained on WSJ corpus (WSJ-MLP) output phoneme posterior probabilities (also referred to as posterior features) of dimension 45.

In the KL-HMM, each subword unit is modelled by a 3 state left-to-right HMM.

In the reminder of this section we present our studies using the proposed system and compare it with phoneme-based KL-HMM system. Our main interest here is not to out-perform the phoneme-based system but to understand better how the gaps between the two systems can be reduced. We note that when ever the terms grapheme-based and phoneme-based are used we refer to the type of subword unit being used in the lexicon/dictionary.

A. Context-independent subword unit studies

Table II presents the ASR results when context-independent grapheme and phoneme subword units are modelled. As expected the system modelling both grapheme and phoneme information (grapheme-based) performs significantly worse than the standard phoneme-based ASR system. This is mainly due to the fact that the multinomial distribution trained for each state captures gross phoneme posterior information. For instance, the multinomial distribution for grapheme model [C] can capture information about both phoneme /ch/ and /k/. It can be observed that among the different local measures namely, KL, RKL, and SKL, RKL yields the best system when using grapheme as subword units. This can be attributed to the fact that state distributions of grapheme subword units are capturing gross phoneme posterior distribution. However, the posterior estimates from the output of the MLP typically have peaky distribution (i.e., most of the probability mass falls in one particular dimension). Given this it can be seen from Eqn. (2) that in RKL the MLP posterior estimate is the reference distribution which brings the ability to select a particular dimension (phoneme information) from the gross distribution.

In the case of phoneme as subword units the performance using different local measures is consistent with previous observation [9] i.e., use of SKL as local measure yields better system than both KL and RKL.

Furthermore, it can be seen that use of an MLP trained on larger amount of out-of-domain data i.e., WSJ-MLP leads to improvement in the performance for all the systems except for grapheme-based ASR system using RKL as local measure.

TABLE II
WORD ERROR RATE (WER) EXPRESSED IN % FOR CONTEXT-INDEPENDENT GRAPHEME AND CONTEXT-INDEPENDENT PHONEME SUBWORD UNITS BASED KL-HMM SYSTEMS. RM-MLP REFERS TO THE USE OF MLP TRAINED ON RM CORPUS FOR POSTERIOR FEATURE ESTIMATION. WSJ-MLP REFERS TO THE USE OF MLP TRAINED ON WSJ CORPUS FOR POSTERIOR FEATURE ESTIMATION.

Subword unit	Local Measure	# of Models	RM-MLP	WSJ-MLP
Grapheme	KL	29	46.3	38.7
	RKL	29	25.2	25.4
	SKL	29	32.8	32.0
Phoneme	KL	42	7.6	7.4
	RKL	42	8.0	7.3
	SKL	42	7.0	6.9

B. Context-dependent subword units studies

Table III presents the performance of KL-HMM system using context-dependent graphemes as subword units and KL-HMM system using context-dependent phonemes as subword units. In this study, similar to[5] we have only modelled word internal single preceding and following context (similar to conventional triphone modelling). Modelling context-dependent unit increases the number of models/parameters and this can possibly lead to poor model estimation due to lack of sufficient data. In such a case, it may be better to do parameter sharing. However, in this study for both grapheme-based and phoneme-based system we found on the development data that it is better to use the model estimated during context-dependent model training even if the occupancy count is as low as 15 frames compared to state/model tying. Furthermore, in the case of grapheme-based KL-HMM system we also found that using context-independent models as initial model for context-dependent models training yields similar performance compared to the use of a flat model (i.e., all dimensions having equal probability) as initial model. In other words, unlike phoneme-based KL-HMM system the training of context-independent models is not always necessary for grapheme-based KL-HMM system. Thus, in this paper we report the performance of context-dependent grapheme-based KL-HMM system where a flat model is used as an initial model.

TABLE III
WER EXPRESSED IN % FOR CONTEXT-DEPENDENT GRAPHEME AND CONTEXT-GRAPHEME PHONEME SUBWORD UNITS BASED KL-HMM SYSTEM. RM-MLP REFERS TO THE USE OF MLP TRAINED ON RM CORPUS FOR POSTERIOR FEATURE ESTIMATION. WSJ-MLP REFERS TO THE USE OF MLP TRAINED ON WSJ CORPUS FOR POSTERIOR FEATURE ESTIMATION.

Subword unit	Local Measure	# of Models	RM-MLP	WSJ-MLP
Grapheme	KL	1911	7.7	7.9
	RKL	1911	6.6	6.2
	SKL	1911	6.3	6.1
Phoneme	KL	2305	5.6	5.1
	RKL	2305	5.8	5.1
	SKL	2305	5.5	5.1

The results[3] show that the modelling of grapheme context significantly reduces the performance gap between grapheme-based KL-HMM system and phoneme-based KL-HMM system. However, from the above results it can be also seen that there still exists a gap of about 1% absolute WER (comparing the best systems) between grapheme-based and phoneme-based systems with the latter being better.

Interestingly for grapheme-based system when compared to context-independent system SKL yields better system compared to RKL and KL. In the case of phoneme-based system all the local measures yield similar performances.

Similar to the context-independent subword units study reported in the previous section, the WSJ-MLP trained on large amount of training data yields the best system for both grapheme-based and phoneme-based systems.

The improvement in the performance using context-dependent graphemes and the difference in the performances across different local measures for grapheme-based system suggests that though modelling single preceding and following grapheme context brings the models closer to phonemes, there still exists certain amount "ambiguous"/gross information if not in all but in few models. For instance, the average of the entropy of all multinomial distributions corresponding to the context-dependent grapheme models trained using KL local measure is 1.325 bits compared to 0.607 bits for context-dependent phoneme models trained using KL local measure. This also suggests that the context-dependent grapheme models can be limited in terms of their ability to capture phoneme level contextual information compared to context-dependent phoneme models.

In the following two subsections we present two different approaches for context-dependent grapheme models to better capture/model the contextual information. In the first approach the number of context-dependent models is preserved and observation feature dimension is increased by using a sequence of posterior feature vectors as observation feature vector. In the second approach, the observation feature dimension is preserved and the number of models are increased using early tagging of contextual graphemes [6].

C. Contextual modelling using posterior feature sequence

One way to improve the ability of context-dependent grapheme models to capture or better model the contextual information is to use a temporal sequence of posterior feature vectors as feature observation as opposed to just a single frame. In doing so, we can expect the KL-HMM system to better model the relation between the subword unit (context-dependent grapheme) and the evolution of posterior features (phoneme posterior probabilities) in a temporal neighborhood or simply said temporal contextual information. We performed preliminary ASR studies using a posterior feature sequence of length 3 (1 frame of each preceding and following context)

[3]The best performance of 5.5% WER and 6.3% WER for phoneme-based and grapheme-based system, respectively when using RM-MLP compares favorably to the best phoneme-based and grapheme-based results reported in [5].

and 5 (2 frames of each preceding and following context). In this approach, each state is modelled by a stack of multinomial distributions which are trained jointly. The number of stacks in is either 3 or 5 depending upon the length (number of frames) of the posterior feature sequence that is used as observation feature vector. Table IV presents the results of this study.

TABLE IV
WER EXPRESSED IN % FOR CONTEXT-DEPENDENT GRAPHEME AND CONTEXT-GRAPHEME PHONEME SUBWORD UNITS BASED KL-HMM SYSTEM WHERE OBSERVATION FEATURE VECTOR IS A SEQUENCE OF POSTERIOR FEATURE VECTOR. CONTEXT OF 1 REFERS ONE FRAME OF BOTH PRECEDING AND FOLLOWING TEMPORAL CONTEXT WHERE AS A CONTEXT OF 2 REFERS TO TWO FRAMES OF BOTH PRECEDING AND FOLLOWING TEMPORAL CONTEXT. RM-MLP REFERS TO THE USE OF MLP TRAINED ON RM CORPUS FOR POSTERIOR FEATURE ESTIMATION. WSJ-MLP REFERS TO THE USE OF MLP TRAINED ON WSJ CORPUS FOR POSTERIOR FEATURE ESTIMATION.

Subword unit	Local Measure	# of Models	Context	RM-MLP	WSJ-MLP
Grapheme	KL	1911	1	7.3	7.5
	RKL	1911	1	6.4	5.8
	SKL	1911	1	6.4	5.8
Phoneme	KL	2305	1	5.8	5.3
	RKL	2305	1	5.9	5.3
	SKL	2305	1	5.8	5.1
Grapheme	KL	1911	2	7.2	7.0
	RKL	1911	2	6.0	5.7
	SKL	1911	2	6.0	5.7
Phoneme	KL	2305	2	5.9	5.1
	RKL	2305	2	5.8	5.4
	SKL	2305	2	5.7	5.1

The results show that context-dependent grapheme-based KL-HMM can benefit from the information present in the temporal sequence of posterior features especially when 2 frame preceding and following posterior feature sequence is modelled. It can also be noticed that the performance of grapheme-based KL-HMM improves by increasing the temporal context. Finding the appropriate temporal context is open for future research.

However, in the case of context-dependent phoneme-based system no real improvements are seen by modelling the temporal feature sequence. On the contrary, it is hurting the system at times. It is possible that there is no useful or only redundant information is present for phoneme-based system, but it may be possible that the phoneme-based system can benefit from much longer temporal sequence (about 130-150 ms) as seen in some of the MLP-based modelling studies [11].

To summarize, the results tries to show that it can be possible to bridge the gap between grapheme-based and phoneme-based system by modelling the temporal contextual information present in the sequence of posterior features. Also, it can be observed that the use of phoneme posterior estimates from WSJ-MLP yields better system.

D. Contextual modelling using early tagging

Earlier in the Section IV-B we mentioned that the context-dependent grapheme models are initialized with a flat model and are then trained. In doing so, we in reality are performing early tagging [6]. In other words, these models can been seen as a kind of context-independent phoneme-like units that avoids capturing of gross distribution. This observation can be fairly made if we compare the performance of the context-dependent grapheme system for KL local measure in Table III with the performance of the context-independent phoneme system for different local measures in Table II. As described in [6], it is possible to better model the contextual information by building context-dependent models out of the early tagged units. We performed two studies where in the first study we modelled single preceding and following context of the early tagged units (can be seen as equivalent to modelling two preceding and two following context for each context-independent grapheme units, i.e. quintgraph), and in the second study we modelled only the following context (can be seen as equivalent to modelling one preceding context and two following context for each context-independent grapheme). During training the context-dependent early tagged units were initialized by context-independent early tagged units (i.e., the 1911 models resulting from context dependent system described in Section IV-B). Table V shows the results for these studies.

TABLE V
WER EXPRESSED IN % FOR GRAPHEME-BASED KL-HMM SYSTEM WITH CONTEXTUAL MODELLING OF EARLY TAGGED UNITS. SPF REFERS TO SINGLE PRECEDING AND FOLLOWING CONTEXT MODELLING. SF REFERS TO SINGLE FOLLOWING CONTEXT MODELLING. RM-MLP REFERS TO THE USE OF MLP TRAINED ON RM CORPUS FOR POSTERIOR FEATURE ESTIMATION. WSJ-MLP REFERS TO THE USE OF MLP TRAINED ON WSJ CORPUS FOR POSTERIOR FEATURE ESTIMATION.

Subword unit	Local Measure	# of Models	Context	RM MLP	WSJ MLP
Grapheme	KL	3208	SF	6.6	6.6
	RKL	3208	SF	6.3	5.9
	SKL	3208	SF	5.6	5.6
Grapheme	KL	4111	SPF	6.5	5.9
	RKL	4111	SPF	6.1	5.7
	SKL	4111	SPF	5.6	5.4

Contextual modelling of early tagged units in either way helps in improving the performance of the grapheme-based KL-HMM system such that the grapheme-based KL-HMM system using SKL as local measure is becoming more comparable (closer) to the phoneme-based KL-HMM system. It is also interesting to note that when single preceding and following context is modelled and WSJ-MLP is used as posterior estimator the performances of systems using different local measures are becoming comparable. This suggests that given a well trained MLP on large amount of data there is good potential in exploiting the contextual modelling of early tagged units to better capture the phoneme contextual information and improve the performance of grapheme-based KL-HMM system.

V. DISCUSSION

The studies presented in this paper shows some interesting things. For instance, simply modelling the single preceding

and single following grapheme context (like triphone systems) may not endow the grapheme-based ASR system with the same capabilities as a triphone system. However, by modelling single preceding and single following grapheme context the resulting models can be expected to have better modelling capacity than context-independent phoneme (given that we choose a right local measure). How much better this can depend upon factors such as the task at hand. In this paper, we performed studies on read speech, and we obtained better results than context-independent phoneme-based system using SKL and RKL local measure when modelling single preceding and single following grapheme context. However, on spontaneous speech it is possible to see that the performance of context-dependent (single preceding and single following grapheme) grapheme system and context-independent phoneme system are not too far apart. In such a case, we expect that the capabilities of KL-HMM system to model contextual information at the posterior feature level, and the use of early tagged units in conjunction with lesser training data requirement can have bigger roles to play in bridging the gap between context-dependent grapheme KL-HMM system and context-dependent phoneme KL-HMM system.

Furthermore, the superiority of local measures SKL and RKL over KL when modelling context-dependent grapheme systems can be more attributed to the ability of these measures to handle gross distributions (or ambiguity present in multinomial distribution) in conjunction with reliable (and "peaky") posterior estimate obtained at the output of MLP at time frame.

The context-dependent grapheme-based KL-HMM system investigated in this paper can be contrasted with context-dependent grapheme-based HMM/GMM system using tandem features (see also Table I) such as the one studied in [5]. There are a few advantages with KL-HMM approach over tandem system. Tandem system uses transformed posterior probabilities and in the course of transformation the dimensions lose the phoneme identity information where as KL-HMM preserves it. Retaining the identity of phoneme information can be beneficial. For instance, it is possible to initialize the multinomial distribution for a context-dependent grapheme subword unit given the grapheme context information using such as letter-to-sound rules. Also, retaining the phoneme identity information may give flexibility in learning the relation between the subword units and posterior features when modelling temporal context at observation feature level, such as the way we did in this paper. Furthermore, we noted earlier that it was beneficial to retain model parameters of the KL-HMM system at the end of training even if the occupancy count for a model is as low as 15 frames. A similar flexibility may not be feasible/useful with tandem feature based system where more parameters are modelled.

VI. Conclusion

In this paper, we proposed a simple approach to jointly model both grapheme and phoneme information using KL-HMM, where, grapheme are the subword units and the phoneme posterior estimates obtained at the output of MLP are used as observation feature vector. On DARPA RM corpus the proposed approach yielded almost 1% absolute WER higher than the phoneme-based KL-HMM system. However, we showed through preliminary studies that the flexibility of the KL-HMM to model contextual information at the observation feature level and modelling of contextual information using early tagged units can help in reducing the gap between the two systems. We also showed that these findings/observations generalize to the case where the posterior feature vector is estimated using an MLP trained on an entirely different database. Furthermore, it was also found that when context-dependent grapheme units are used SKL local measure yields the best system with RKL being the next best.

Future work includes (a) finding the appropriate temporal length of posterior feature vector sequence that can be modelled effectively, (b) different ways to perform early tagging, (c) investigating the combination of both contextual modelling of early tagged units and modelling of a sequence of posterior feature vector as observation feature vector, and (d) extending the grapheme-based KL-HMM studies to more complex spontaneous conversational speech recognition task and handling of unseen contexts.

Acknowledgment

This work was supported by the Swiss National Science Foundation (SNSF) through the project MULTI and the Swiss National Center for Competence in Research (NCCR) under the project Interactive Multimodal Information Management (IM2) project.

References

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

[2] M. Killer, S. Stüker, and T. Schultz, "Grapheme based speech recognition," in *Proceedings of Eurospeech*, 2003, pp. 3141–3144.

[3] B. Mimer, S. Stüker, and T. Schultz, "Graphembasierte spracherkennung unter verwendung flexibler entscheidungbäume," in *Elektronische Sprachsignalverarbeitung ESSV*, 2004.

[4] M. Magimai.-Doss, S. Bengio, and H. Bourlard, "Joint decoding for phoneme-grapheme continuous speech recognition," in *Proceedings of ICASSP*, Montreal, Canada, 2004, pp. I–177–I–180.

[5] J. Dines and M. Magimai-Doss, "A study of phoneme and grapheme based context-dependent ASR systems," in *MLMI 2007*, ser. Lecture Notes in Computer Science No. 4892, 2008, pp. 215–226.

[6] G. Anumanchipalli, K. Prahallad, and A. Black, "Significance of early tagged contextual graphemes in grapheme based speech synthesis and recognition systems," in *Proceedings of ICASSP*, 2008.

[7] Y.-H. Sung, T. Hughes, F. Beaufays, and B. Strope, "Revisiting graphemes with increased amount of data," in *Proceedings of ICASSP*, 2009.

[8] G. Aradilla, J. Vepa, and H. Bourlard, "An acoustic model based on kullback-leibler divergence for posterior features," in *Proceedings of ICASSP*, 2007.

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

[10] G. Aradilla, "Acoustic models for posterior features in speech recognition," Ph.D. dissertation, EPFL, Lausanne , Switzerland, 2008.

[11] J. P. Pinto, H. Hermansky, B. Yegnanarayana, and M. Magimai.-Doss, "Exploiting contextual information for improved phoneme recognition," in *"Proceedings of ICASSP"*, 2008.

[12] S. King and P. Taylor, "Detection of phonological features in continuous speech using neural networks," *Computer Speech and Language*, vol. 14, no. 4, pp. 333–353, 2000.

[13] P. J. Price, W. Fisher, and J. Bernstein, "A database for continuous speech recognition in a 1000 word domain," in *Proceedings of ICASSP*, 1988.