LANGUAGE DEPENDENT UNIVERSAL PHONEME POSTERIOR ESTIMATION FOR MIXED LANGUAGE SPEECH RECOGNITION

David Imseng[1,2]*, Hervé Bourlard*[1,2]*, Mathew Magimai.-Doss*[1]*, John Dines*[1]

[1]Idiap Research Institute, Martigny, Switzerland
[2]Ecole Polytechnique Fédérale, Lausanne (EPFL), Switzerland
{dimseng,bourlard,mathew,dines}@idiap.ch

ABSTRACT

This paper presents a new approach to estimate "universal" phoneme posterior probabilities for mixed language speech recognition. More specifically, we propose a new theoretical framework to combine phoneme class posterior probabilities in a principled way by using (statistical) evidence about the language identity. We investigate the proposed approach in a mixed language environment (SpeechDat(II)) consisting of five European languages. Our studies show that the proposed approach can yield significant improvements on a mixed language task, while maintaining the performance on monolingual tasks. Additionally, through a case study, we also demonstrate the potential benefits of the proposed approach for non-native speech recognition.

Index Terms— Speech recognition, Mixed language speech recognition, Non-native speech, Acoustic model combination, Universal phoneme set

1. INTRODUCTION

In monolingual speech recognition, the language being spoken is known during testing. On the other hand, in mixed language speech recognition, it is assumed that the language is not known a priori. One possible approach to perform mixed language speech recognition is to build multiple monolingual systems, run them, and select the word hypothesis with the maximum likelihood (among all systems). Such an approach has certain drawbacks. For instance, not all languages may have sufficient resources for building a reliable monolingual system. Also, there could be considerable degradation in performance when combining monolingual systems for the mixed language task [1]. Thus, in multilingual speech recognition literature, the use of "universal" phoneme models has been proposed [2, 3] and will be further explored here based on exploiting properties of phoneme class posterior probabilities.

The use of universal phoneme models allows the sharing of acoustic training data, can help in porting ASR systems across languages with limited resources and provides the ability to make a soft (statistical) decision at the phoneme level about the language being spoken. However, it has also been observed that with a larger quantity of monolingual training data, systems using universal phoneme models yield lower performance than systems using monolingual phoneme models. One reason for this could be that the universal phoneme set is (generally) built by merging the phonemes from multiple languages based on their phonetic symbol (e.g. IPA, SAMPA). Although different languages can have phonemes that share the same phonetic symbol, it is widely accepted that their acoustic manifestations can be significantly different, depending on the language. As a result, the same phoneme classes from different languages, as well as phoneme instances pronounced by non-native speakers, may not cluster well [4].

In this paper, we thus investigate the notion (and estimation) of universal phoneme posteriors, conditioned on the (unknown) spoken language. This is done in the context of hybrid HMM/MLP ASR systems, where multilayer perceptrons (MLP) are providing us with phoneme posterior estimates. Monolingual phoneme posterior probabilities (phoneme posteriors) for each language are thus estimated (by MLP) and then combined (using statistical evidence about language identity) to estimate *language dependent universal phoneme posteriors*. The proposed approach thus attempts to merge the benefits of the greater class discrimination provided by monolingual systems with the benefits of a soft decision offered by systems using universal phoneme models.

We formulate the notion of monolingual and universal phoneme posteriors in Section 2, before the proposed approach is investigated on SpeechDat(II) databases from five European languages in Section 3. Our findings suggest that our novel approach can yield improvement in mixed language environments. Further analysis of the results show that the proposed approach could also be potentially applied to non-native and/or accented speech recognition (Section 4). We discuss the novel approach in Section 5 and Section 6 concludes the paper.

2. PHONEME POSTERIOR ESTIMATION

In this section, we present the estimation of monolingual and universal phoneme posteriors. For the sake of clarity, we first briefly present the notation used in this paper before explaining how we estimated language-conditioned universal phoneme posteriors.

2.1. Notation

Each language $l \in \{1, \ldots, N\}$ (N is the number of different languages) has its own phoneme set consisting of monolingual phonemes, m_l^k, where $k \in \{1, \ldots, K_l\}$ (K_l being the number of phonemes in language l). Each phoneme, m_l^k, corresponds to a SAMPA[1] symbol. By merging all of the phoneme sets and assuming

This research was supported by the Swiss NSF through the project MultiModal Interaction and Multimedia Data Mining under contract number MULTI-200020-122062, through the National Center of Competence in Research on "Interactive Multimodal Information Management" (www.im2.ch) and by the European Community's Seventh Framework Programme (FP7/2007-2013) grant agreement 213845 (the EMIME project: www.emime.org).

[1]http://www.phon.ucl.ac.uk/home/sampa/

that the same SAMPA symbol represents the same phoneme class across languages, we can build one universal phoneme set consisting of universal phonemes u^i, where $i \in \{1, \ldots, I\}$ (I being the number of universal phonemes).

We distinguish between monolingual phoneme posteriors, $P(m^k_{l,t}|x_t)$, (for ease of notation we omit the conditional variable l, because the language is implicit to m^k_l) and universal phoneme posteriors, $P(u^i_t|x_t)$. The term x_t represents the acoustic vector observation at time t. The expressions $m^k_{l,t}$ and u^i_t stand for the statistical events $m_{l,t} = k$ (the system of language l is in class k at time t) and $u_t = i$ (the system is in class i at time t), respectively.

2.2. Monolingual phoneme posteriors

Conventional approaches to hybrid HMM-based speech recognition use monolingual phoneme posteriors. In a multilingual environment and in the case of HMM/MLP ASR systems, this requires training one MLP with K_ℓ outputs per language, ℓ, to estimate monolingual phoneme posteriors, $P(m^k_{l,t}|x_t)$. As we know from many previous experiments, MLPs yield very good estimates of these phoneme posteriors [5].

2.3. Universal phoneme posteriors

We compare two different approaches to estimate universal phoneme posteriors: (1) a conventional language independent approach, based on the training of one single multilingual MLP, and (2) a new language dependent approach, where universal language-conditioned phoneme posteriors are first obtained and then combined.

2.3.1. Language independent approach

In this approach, an MLP with I outputs is trained with the data from all available languages to estimate universal phoneme posteriors, $P(u^i_t|x_t)$.

2.3.2. Language dependent approach

In the language dependent approach, each language contributes to the estimation of the universal phoneme posteriors through proper combination of monolingual phoneme posteriors:

$$P(u^i_t|x_t) = \sum_{l=1}^{N} P(u^i_t, l|x_t) \quad (1)$$

$$= \sum_{l=1}^{N} P(u^i_t|x_t, l)P(l|x_t) \quad (2)$$

which suggests to weight the language-conditioned universal phoneme posteriors, $P(u^i_t|x_t, l)$, by the corresponding frame-based language posterior probability, $P(l|x_t)$.

One way to estimate the language-conditioned universal phoneme posterior, $P(u^i_t|x_t, l)$, is by a linear combination of monolingual phoneme posteriors:

$$P(u^i_t|x_t, l) = \sum_{k=1}^{K_l} P(m^k_{l,t}, u^i_t|x_t) \quad (3)$$

$$= \sum_{k=1}^{K_l} P(u^i_t|x_t, m^k_{l,t})P(m^k_{l,t}|x_t) \quad (4)$$

As previously noted from past literature [4], language independent acoustic modeling techniques tend to degrade the performance of ASR systems due to the greater within phoneme class variability across languages. Thus, we may hypothesize that a linear combination of language dependent posteriors may provide a better estimate of universal phoneme posterior probabilities, where Equation (4) shows that we should weight the monolingual phoneme posteriors, $P(m^k_{l,t}|x_t)$, by the conditional probabilities, $P(u^i_t|x_t, m^k_{l,t})$.

An MLP classifier could be trained to estimate $P(u^i_t|x_t, m^k_{l,t})$, but this is not obvious. Alternatively, one could assume that $P(u^i_t|x_t, m^k_{l,t})$ is (1) conditionally independent of the acoustic observation, x_t (given $m^k_{l,t}$), and (2) time invariant. The resulting conditional probability, $P(u^i|m^k_l)$, can then be estimated by exploiting linguistic knowledge or by applying data-driven techniques. For this work, we study the linguistic knowledge approach where a probability of one occurs when the SAMPA symbol for monolingual and universal phoneme sets correspond:

$$P(u^i|m^k_l) = \begin{cases} 1, & \text{if } u^i = m^k_l \\ 0, & \text{if } u^i \neq m^k_l \end{cases} \quad (5)$$

Equations (2), (4) and (5) then yield the following universal phoneme posterior estimates:

$$P(u^i_t|x_t) = \sum_{l=1}^{N} P(l|x_t) \sum_{k=1}^{K_l} P(u^i|m^k_l)P(m^k_{l,t}|x_t) \quad (6)$$

3. EXPERIMENTAL SETUP

For this study, we use the application words corpus (isolated words) from the SpeechDat(II) databases of British English (EN), Spanish (ES), Italian (IT), Swiss French (SF) and Swiss German (SZ). In total there are 6.8, 0.8 and 2.5 hours of training, development and testing data, respectively. All 182 words (about 36 words per language) are transcribed in the dictionaries. For more details about the databases and the phoneme sets, the reader is referred to [1].

In this section, we first describe the different phoneme posterior estimators before we introduce the monolingual and the mixed language tasks along with the systems that are evaluated.

3.1. Phoneme posterior estimators

We investigate three different MLP-based posterior estimators (trained with Quicknet[2] software). As we usually do, all the MLPs used 39 Mel-Frequency Perceptual Linear Prediction (MF-PLP) features ($C_0 - C_{12} + \Delta + \Delta\Delta$) in a nine frame temporal context (four preceding and following frames), extracted with HTK[3], as input.

3.1.1. Monolingual estimator

To estimate monolingual phoneme posteriors, we used MLPs from previous work [1] (one MLP per language). Each MLP had one hidden layer containing 600 hidden nodes. The MLP output dimensionality corresponds to the number of phonemes in each language.

3.1.2. Universal estimators

We distinguish between the language independent and the language dependent estimation technique.

[2] http://www.icsi.berkeley.edu/Speech/qn.html

[3] http://htk.eng.cam.ac.uk/

Language independent: a universal MLP with one hidden layer containing 2620 hidden units (to keep the total number of parameters the same as in Section 3.1.1) is trained to estimate 92 universal phoneme posteriors.

Language dependent: Following Equation (6), we need three components to estimate language dependent universal phoneme posteriors. The language posteriors $P(l|x_t)$ are estimated with the recently proposed hierarchical MLP-based language identification (LID) approach [6]. The conditional phoneme posteriors, $P(u^i|m_l^k)$, are estimated according to Equation (5) and the monolingual phoneme posteriors, $P(m_{l,t}^k|x_t)$, from the monolingual classifiers discussed in Section 3.1.1 are used. The frame-based language posterior estimates tend to be noisy, therefore we smooth them by averaging over a fixed window (on which we assumed that the language stays the same) as shown below:

$$P(l|x_t) = \frac{1}{2c+1} \sum_{n=t-c}^{t+c} P(l|x_n) \tag{7}$$

The smoothing window c was tuned on the development set and fixed to $c = 21$.

3.2. Monolingual task

For the monolingual task, we assume that the language being spoken is known in advance. We compare three different systems on this task. Each system uses five isolated word recognizers (one for each language) and we select and run the recognizer associated with the known input language.

- **System mono** uses monolingual phoneme posteriors.
- **System uMLP** uses language independent phoneme posteriors estimated with the universal MLP.
- **System comb** uses language dependent phoneme posteriors obtained by combining monolingual posteriors. Since $P(l|x_t)$ is known in a monolingual environment, System comb would yield the same result as System mono. Nevertheless, also for the monolingual task, we use the estimates of $P(l|x_t)$ for System comb.

3.3. Mixed language task

For the mixed language task, we assume that the spoken language is not known a priori. On this task, we study four different systems.

- **System bbox** (blackbox) uses monolingual phoneme posteriors. It runs all five recognizers and picks the output of the recognizer with the highest likelihood[4].
- **System LID** uses monolingual phoneme posteriors. Firstly, the language is identified with our recently proposed hierarchical MLP-based LID approach [6] and secondly, the recognizer of the identified language is run.
- **System uMLP** uses language independent phoneme posteriors. One recognizer with a shared lexicon is built and run.
- **System comb** uses language dependent phoneme posteriors. One recognizer with a shared lexicon is built and run.

4. RESULTS

In this section, we present the experimental results of the systems presented in Section 3.

[4]The performance of System bbox has already been published in [1].

4.1. Monolingual task

Table 1 summarizes the number of errors for each system on the monolingual task (6975 test utterances in total). Although System comb is using estimates of $P(l|x_t)$, there is no significant difference in performance between the three systems (McNemar test with significance level 2.5%).

Monolingual task			
System	mono	uMLP	comb
Total errors	171	169	192
Word accuracy	97.5%	97.6%	97.2%

Table 1. Comparison of the three systems (see Section 3.2) on the monolingual task (language known during testing).

4.2. Mixed language task

Table 2 shows the number of errors for each system on the mixed language task. System bbox and System LID are both significantly outperformed by the systems that use universal phoneme posteriors. Universal phoneme posteriors should thus be preferred in mixed language environments.

Mixed language task				
System	bbox	LID	uMLP	comb
Total errors	951	430	331	308
Word accuracy	86.4%	93.8%	95.3%	95.6%

Table 2. Comparison of the four systems (see Section 3.3) on the mixed language task (language not known a priori).

System LID and System comb use the same language posterior estimator. System LID performs an utterance-based hard decision about the language, whereas System comb incorporates the smoothed language posteriors on a frame basis. System comb significantly outperforms System LID. Thus, the benefits of soft decisions taken in the universal phoneme approach are evident.

System comb performs best, but there is no significant improvement compared to System uMLP (McNemar, 2.5%). The universal MLP estimator is based on the assumption that the same SAMPA symbol represents the same phoneme across languages. Since the same assumption can be found in the binary formulation of Equation 5, it is not surprising that both estimation techniques yield similar results. However, by estimating $P(u_t^i|x_t, m_{l,t}^k)$ in a data-driven way, we could expect better results for System comb.

4.3. Non-native speech - a case study

We hypothesize that the knowledge contained in multilingual speech recognizers may even be useful in a monolingual context, especially when dealing with accented or non-native speech. In particular, the estimates of $P(l|x_t)$ in System comb enable a dynamic weighting of phoneme probabilities from different language sources that may be beneficial under such circumstances. Since the SpeechDat(II) databases contain only a few non-native speech utterances it is difficult to explore this claim in detail, however, most speakers of the Swiss French corpus volunteered to provide their mother tongue. We therefore analyzed one case (file A14417A4) and assumed that the spoken language was known a priori. The utterance consisted of the French word *stop*, read by a native Swiss German speaker. The

word *stop* was present in the dictionaries of English, Swiss German and Swiss French, but each language uses a different phonetic transcription.

In Figure 1, the smoothed language posteriors are shown. The x-axis provides the phoneme forced alignment and the y-axis shows the language posteriors. As we might expect, the posteriors of both the spoken language (Swiss French) and the mother tongue (Swiss German) dominate in the speech regions, but interestingly the Swiss German posteriors are the most important. However, on the relatively long silence at the end of the utterance, the distribution changes completely which indicates that the hierarchical MLP-based LID approach learns the language and not the channel.

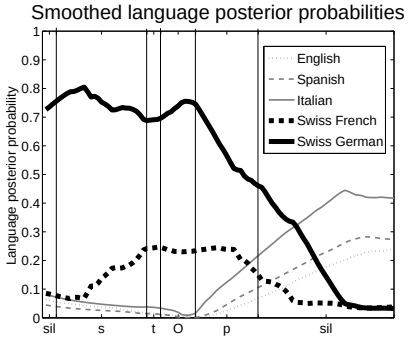

Fig. 1. Language posteriors of the considered utterance. The x-axis provides the forced aligned phoneme alignment.

Frame-based Swiss German and Swiss French phoneme posteriors (estimated by their respective monolingual MLPs) are given in Figure 2. It can be seen that the Swiss German phoneme posteriors correspond quite well with the phoneme alignments. In the case of Swiss French, *[O]* has low posterior probability for the frames with which it should normally be associated, while the phonemes *[a~]* and *[O/]* have high posterior probabilities. As a result, the Swiss French recognizer wrongly decodes the utterance as *[a~] [f] [a~]* (enfant, "child" in English), whereas System comb and System uMLP both correctly decoded the utterance. This implies that the proposed approach can be exploited to boost the performance of non-native ASR.

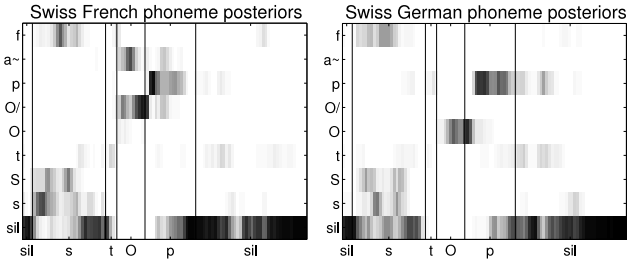

Fig. 2. Frame-based phoneme posteriors for Swiss German and Swiss French MLP classifiers. The intensity in the image is scaled between 0 and 1 (1 is black and 0 is white). For the sake of clarity, only the posteriors for a subset of all phonemes are displayed.

5. DISCUSSION

System comb aims at estimating universal phoneme posteriors while conserving language specific characteristics. The contribution is twofold: it targets to improve speech recognition in mixed language environments and it has the potential to exploit multilingual information to improve speech recognition of accented and/or non-native speech. Furthermore, it might also be useful for acoustic modeling in code-switching environments.

It is possible to draw similarities between the proposed language dependent approach and other approaches studied in a multilingual scenario. For instance, (1) the notion of estimating universal posteriors conditioned upon language identity i.e., $P(u_t^i|x_t, l)$ can be compared to the use of language questions in the decision of splitting the polyphone tree [2], and (2) the idea of weighted combination of monolingual phoneme posteriors in our case can be seen in the same light as the acoustic model interpolation approach to improve speech recognition of low-resource languages [2]. We can also find some similarity with other work which has looked at combining knowledge from various languages at the feature level [7] as opposed to the probabilistic combination presented here.

6. CONCLUSION

We have presented a novel approach to estimate universal phoneme posteriors based on a probabilistic combination of monolingual phoneme classifiers and compared this against conventional monolingual and universal phoneme posterior estimators. We observe no significant difference in using monolingual or universal phoneme posteriors in a monolingual environment. However, if the language is not known a priori, the use of universal phoneme posteriors leads to significant improvements in performance (over the use of monolingual posteriors) with the newly proposed approach yielding the best system. Furthermore, a case study on a non-native speech utterance suggests that universal phoneme posteriors could also help in improving speech recognition on non-native speech, hence pointing towards further research in that direction.

In the future, we also aim to further explore the proposed approach by using data driven methods to model the relation between monolingual phoneme units and universal phoneme units (i.e., $P(u_t^i|x_t, m_{l,t}^k)$) that relaxes the linguistic constraint that the same phonetic symbols represent the same phonemes across languages.

7. REFERENCES

[1] David Imseng, Hervé Bourlard, and Mathew Magimai.-Doss, "Towards mixed language speech recognition systems," in *Proc. of Interspeech*, 2010, pp. 278–281.

[2] Tanja Schultz and Alex Waibel, "Language independent and language adaptive acoustic modeling for speech recognition," *Speech Communication*, vol. 35, pp. 31–51, 2001.

[3] Joachim Köhler, "Multilingual phone models for vocabulary-independent speech recognition tasks," *Speech Communication*, vol. 35, pp. 21–30, 2001.

[4] Dirk Van Compernolle, "Recognizing speech of goats, wolves, sheep and...non-natives," *Speech Communication*, vol. 35, pp. 71–79, 2001.

[5] Michael D. Richard and Richard P. Lippmann, "Neural network classifiers estimate bayesian a posteriori probabilities," *Neural Computation*, vol. 3, no. 4, pp. 461–483, 1991.

[6] David Imseng, Mathew Magimai.-Doss, and Hervé Bourlard, "Hierarchical multilayer perceptron based language identification," in *Proc. of Interspeech*, 2010, pp. 2722–2725.

[7] Martin Raab, Rainer Gruhn, and Elmar Nöth, "Multilingual weighted codebooks for non-native speech recognition," in *Proc. of the 11th int. conf. on Text, Speech and Dialogue*, 2008, pp. 485–492.