Improving non-native ASR through stochastic multilingual phoneme space transformations

David Imseng[1,2], Hervé Bourlard[1,2], John Dines[1], Philip N. Garner[1] and Mathew Magimai.-Doss[1]

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

Abstract

We propose a stochastic phoneme space transformation technique that allows the conversion of conditional source phoneme posterior probabilities (conditioned on the acoustics) into target phoneme posterior probabilities. The source and target phonemes can be in any language and phoneme format such as the International Phonetic Alphabet. The novel technique makes use of a Kullback-Leibler divergence based hidden Markov model and can be applied to non-native and accented speech recognition or used to adapt systems to under-resourced languages. In this paper, and in the context of hybrid HMM/MLP recognizers, we successfully apply the proposed approach to non-native English speech recognition on the HIWIRE dataset.

Index Terms: Non-native speech recognition, universal phoneme set, multilingual acoustic modeling

1. Introduction

State-of-the-art speech recognizers typically use phonemes as sub-word units. However, training phoneme models is still a challenging task given the high pronunciation variability of words (within the same language), as well as the variability of the acoustic realization of the same phoneme (within and between languages). In this paper, we propose an approach addressing some of the issues related to acoustic modeling of phonemes and apply the proposed approach to non-native speech recognition in the framework of a hybrid HMM/MLP, using a Multilayer Perceptron (MLP) to estimate phonetic class posteriors used in Hidden Markov Models (HMM).

A phoneme set represents the sounds of spoken language and is specific to a language in the sense that two languages could share some, but usually not all, phonemes. The creation of a phoneme set and a lexicon requires linguistic expertise and resources, which include human knowledge.

To date, ASR studies have mainly focused on the recognition of speech from native speakers, while effectively recognizing speech from both native and non-native speakers is still a major challenge. Usually, pronunciation lexicons are created by only taking into account how native speakers pronounce the words. Even then, it is known that acoustic realizations of the same phoneme exhibit high variability, thus, a considerable amount of data is necessary to properly train the models. Modeling variability of the acoustic realizations becomes even more challenging if we have to deal with non-native and accented speech, the main reason being the influence of the native language on the target language sound pronunciation.

In previous work [1], we found that ASR performance on non-native speech can be improved by pooling resources from multiple languages via a universal phoneme set. In this paper, we boost non-native ASR performance by transforming multilingual class probabilities conditioned on the acoustics into monolingual class probability estimates of a target language. More specifically, we first create a universal phoneme set, and then train universal acoustic models with data from five European languages. Given an entirely new target database, along with the lexical resources, the relation between the universal phoneme set and the target phoneme set is learned on the adaptation data by using a Kullback-Leibler divergence based HMM, as presented in Section 2. The learned relation can be seen as a data-driven soft mapping between two phoneme sets that takes the acoustics into account. During recognition, the resulting stochastic mapping is then exploited to transform the conditional posterior probabilities of the universal phonemes into estimates of posterior probabilities of the phonemes belonging to the target database. With less than two minutes of non-native adaptation data, the proposed system yields significant improvement compared to a system trained on native English.

2. Stochastic phoneme space transformation

Although humans are able to produce a large variety of phones, we assume here that all those phones, across speakers and languages, share a common acoustic space $\mathcal{X}$. None or only very few languages make use of all phones. Therefore, most languages only partially cover $\mathcal{X}$.

In ASR, we usually use phonemes as sub-word units to model human speech production. A phoneme is defined as the smallest sound unit of a language that discriminates between a minimal word pair [2]. In contrast to phones, phonemes are defined in the context of a particular language. Therefore, as visualized in Figure 1, two different phoneme sets partition the same acoustic space differently. We consider:

- A source phoneme set Φ consisting of S phonemes s^k
- A target phoneme set Ψ consisting of D phonemes d^l

where $k \in \{1, \ldots, S\}$ and $l \in \{1, \ldots, D\}$

In this paper, we investigate a new approach to map conditional class probabilities of phonemes from a source phoneme set Φ to a target phoneme set Ψ, given acoustic observations. In general, we consider the source and target phoneme sets to be defined in different languages. It is evident that phoneme sets of foreign languages have a different coverage of the acoustic space $\mathcal{X}$.

More specifically, we consider the following problem: given an MLP trained to estimate source phoneme posterior

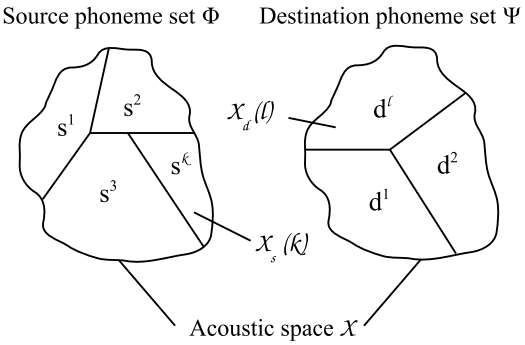

Figure 1: *Two different phoneme sets cover the same acoustic space differently. $\mathcal{X}_s(k)$ and $\mathcal{X}_d(l)$ are acoustic subspaces associated with phonemes s^k and d^l respectively.*

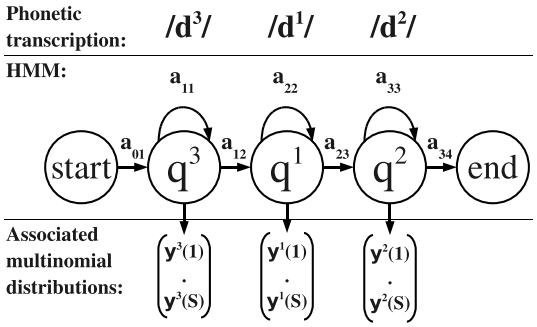

Figure 2: *The HMM structure is "left-to-right" and obtained from the target phoneme transcriptions. Each state is parametrized by a multinomial distribution of dimensionality S. The transition probabilities are also parameters of the HMM.*

probabilities conditioned on acoustic observations, we would like to perform ASR on a target database that makes use of a target phoneme set. No source phoneme transcriptions are available for the target database. However, we assume that the target database can be divided into an adaptation and a testing set. For the testing set, $\hat{X} = \{\hat{x}_1, \ldots, \hat{x}_{\hat{T}}\}$, no transcriptions are available at all, but for the adaptation data $X = \{x_1, \ldots, x_T\}$, we assume access to target phoneme transcriptions, i.e. we assume that we can associate a sequence of target phonemes to X, but we are not able to associate a target phoneme to a particular x_t. Therefore, our approach makes use of an HMM where the states (hidden variables) will be associated with the target phoneme sequence.

Hence, we can formulate the problem of estimating target phoneme posteriors conditioned on the acoustic observation $\hat{x}_t$ at time t, the parameters θ_H of the HMM and the parameters θ_M of the MLP as follows:

$$P(d_t^l|\hat{x}_t, \theta) = \sum_{k=1}^{S} P(d_t^l|s_t^k, \hat{x}_t, \theta) P(s_t^k|\hat{x}_t, \theta) \quad (1)$$

$$= \sum_{k=1}^{S} P(d^l|s^k, \theta) P(s_t^k|\hat{x}_t, \theta_M) \quad (2)$$

where $\theta = \{\theta_H, \theta_M\}$. The target phoneme posterior estimates, $P(d_t^l|\hat{x}_t, \theta)$, can then be used to perform ASR on the target database.

Equation (2) was obtained by making the following conditional independence assumptions:

- The conditional probability $P(d_t^l|s_t^k, \hat{x}_t, \theta)$ can be seen as a similarity measure between a source phoneme s^k and a target phoneme d^l. It is assumed to be time invariant and independent of the acoustic observation $\hat{x}_t$ at time t.
- The source phoneme posteriors $P(s_t^k|\hat{x}_t, \theta)$ are obtained with the MLP[1] that was previously trained on an independent, frame-level labeled, database that may contain speech of the same language, a different language, or from multiple languages. Since frame-level labeling is available for the source database, the source phoneme posterior probability estimates are considered independent of θ_H.

[1]The deployed MLP takes a temporal context of four preceding and following frames into account. For the ease of notation, we just write $P(s_t^k|\hat{x}_t, \theta)$.

Since the states of the HMM will be associated with the target phoneme sequence, we have to estimate $P(s^k|d^l, \theta)$ rather than $P(d^l|s^k, \theta)$. Applying Bayes rule to $P(d^l|s^k, \theta)$, (2) becomes:

$$P(d_t^l|\hat{x}_t, \theta) = \sum_{k=1}^{S} \frac{P(s^k|d^l, \theta) P(d^l|\theta)}{\sum_{l=1}^{D} P(s^k|d^l, \theta) P(d^l|\theta)} P(s_t^k|\hat{x}_t, \theta_M) \quad (3)$$

where the sum in the denominator acts as a normalization factor. Given $P(s_t^k|\hat{x}_t, \theta_M)$, the estimation of $P(d_t^l|\hat{x}_t, \theta)$ thus requires us to estimate the conditional probability $P(s^k|d^l, \theta)$ and the prior probability $P(d^l|\theta)$.

2.1. Estimation of the conditional probability $P(s^k|d^l, \theta)$

To estimate $P(s^k|d^l, \theta)$, we perform a Viterbi (segmentation-maximization) training procedure. This requires that we first forward pass all the adaptation data X through the MLP to obtain $P(s_t^k|x_t, \theta_M)$. We then use $P(s_t^k|x_t, \theta_M)$ along with the target phoneme transcriptions, to train the HMM parameters θ_H. As illustrated in Figure 2, the deployed HMM uses one state per target phoneme d^l in a *left-to-right* structure that is obtained from the target phoneme transcriptions. In Figure 2 for example, we consider an utterance that can be transcribed as /d^3/ /d^1/ /d^2/. Thus, the associated HMM has five states q^3, q^1, q^2 including non-emitting start and end states. Each state q^l, where $l \in \{1, \ldots, D\}$, is parametrized by a multinomial distribution $y^l = \{y^l(1), \ldots, y^l(S)\}$. The dimensionality of y^l is S, the number of source phonemes. Each dimension k of the multinomial distribution y^l can serve as an estimate of the conditional probability of s^k, given the state d^l, the previously trained MLP and the HMM:

$$y^l(k) = P(s^k|d^l, \theta) \quad (4)$$

The transition probabilities a_{ij}, to go from state i to state j, are also parameters of the HMM, $\theta_H = \{y^l, a_{ij}\}$. We fixed them to 0.5 (except $a_{01} = 1$) to minimize their effect on decoding.

The multinomial distributions $Y = \{y^1, \ldots, y^D\}$ can be optimized (maximization step) by using all the adaptation data X and minimizing a cost function $\mathcal{F}(X, Y)$, defined as follows:

$$\mathcal{F}(X, Y) = \sum_{t=1}^{T} \sum_{l=1}^{D} \mathcal{F}^l(x_t, y^l) \delta^l(x_t) \quad (5)$$

where $\mathcal{F}^l(x_t, y^l)$ is a cost function associated with state d^l and $\delta^l(x_t)$ is the Kronecker delta defined as:

$$\delta^l(x_t) = \begin{cases} 1, & \text{if } x_t \in \mathcal{X}_d(l) \\ 0, & \text{if } x_t \notin \mathcal{X}_d(l) \end{cases}$$

where $\mathcal{X}_d(l)$ is the acoustic subspace that corresponds to d^l. To associate each x_t with one of the acoustic subspaces $\mathcal{X}_d(l)$, the HMM aligns the source phoneme posterior probability vector $\mathcal{P} = \{P(s_t^1|x_t, \theta_M), \ldots, P(s_t^S|x_t, \theta_M)\}$ with the states by minimizing $\mathcal{F}(X, Y)$ (expectation step).

Since we estimate conditional probability distributions $P(s^k|d^l, \theta)$, given posterior distributions $P(s_t^k|x_t, \theta_M)$, it seems reasonable to use a Kullback-Leibler (KL) divergence based cost function for the optimization:

$$\mathcal{F}^l(x_t, y^l) = \sum_{k=1}^{S} P(s_t^k|x_t, \theta_M) \log \frac{P(s_t^k|x_t, \theta_M)}{y^l(k)} \quad (6)$$

Hence, this work makes use of a particular HMM structure which is referred to as KL-based HMM [3]. KL-based HMMs are particularly well suited to deal with posterior probabilities. Minimizing $\mathcal{F}(X, Y)$ subject to $\sum_{k=1}^{S} y^l(k) = 1$, yields [4]:

$$P^*(s^k|d^l, \theta) = \frac{1}{|\mathcal{X}_d(l)|} \sum_{\forall x_t^* \in \mathcal{X}_d(l)} P(s_t^k|x_t^*, \theta_M) \quad (7)$$

where $P^*(s^k|d^l, \theta)$ is the optimal estimate of $P(s^k|d^l, \theta)$ with respect to the cost function $\mathcal{F}(X, Y)$. The operator $|\cdot|$ stands for the cardinality of a set, and the sum extends over all the elements x_t^* associated with the acoustic subspace $\mathcal{X}_d(l)$.

The described HMM can be trained by applying an adapted version of the Viterbi algorithm, using (6) as local distances and re-estimating the multinomial distributions according to (7).

2.2. Estimation of the prior probability $P(d^l|\theta)$

As previously explained, the trained HMM can be used to assign each x_t to an acoustic subspace $\mathcal{X}_d(l)$. Prior probabilities $P(d^l|\theta)$, can thus be estimated as the relative count of acoustic vector observations x_t that are associated with $\mathcal{X}_d(l)$, i.e.:

$$P(d^l|\theta) = \frac{|\mathcal{X}_d(l)|}{\sum_{j=1}^{D} |\mathcal{X}_d(j)|} \quad (8)$$

3. Experimental setup and results

We hypothesize that the proposed approach can yield improvement on non-native ASR because universal phoneme posterior probabilities estimated by an MLP trained on multiple languages are more robust to pronunciation variability as observed in non-native speech. Furthermore, we suppose that the proposed stochastic phoneme space transformation is superior to manually derived phoneme set mappings.

3.1. Source phoneme posteriors

Source phoneme posteriors are estimated on British English, Italian, Spanish, Swiss French and Swiss German SpeechDat(II) databases. All SpeechDat(II) databases contain native speech and are gender-balanced, dialect-balanced according to the dialect distribution in a language region and age-balanced. The databases were recorded over the telephone at 8 kHz and are subdivided into different corpora. We only used *Corpus S*, that contains ten read sentences from each of the 2000 speakers per language.

We trained MLP-based posterior estimators with Quicknet[2] software, as explained in [1], for two different source phoneme sets in SAMPA[3] format.

- English phoneme set: we used only the British English data to train a monolingual MLP (MLP EN) to estimate English SAMPA phoneme posteriors.
- Universal phoneme set: since all the SpeechDat(II) dictionaries use SAMPA symbols, we merged phonemes that share the same symbol across languages to build a universal phoneme set. Two MLPs were trained to estimate universal phoneme posteriors; MLP UNI (universal MLP) with all available data and MLP sUNI (small universal MLP) with one fifth of the data randomly chosen, to match the amount of training data available to MLP-EN.

All the MLPs were trained from 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[4], as input. The number of parameters in each MLP was set to 10% of the number of available training frames. Table 1 summarizes all systems (MLP-AE is presented in Section 3.2).

Table 1: *Overview over all the phoneme posterior estimators. The total amount of training data as well as the phoneme set including the number of phonemes (S) are given.*

System	Phoneme set	S	Data (h)
MLP-EN	SAMPA English	45	12.4
MLP-sUNI	SAMPA universal	117	12.7
MLP-UNI	SAMPA universal	117	63.0
MLP-AE	ARPABET English	38	2.4

3.2. Target phoneme posteriors

To study the proposed approach, we used the HIWIRE [5] database. HIWIRE is a non-native English speech corpus that contains English utterances pronounced by natives of France (31 speakers), Greece (20 speakers), Italy (20 speakers) and Spain (10 speakers). The utterances contain spoken pilot orders made up of 133 words and the database also provides a grammar with a perplexity of 14.9. The dictionary is in CMU format and makes use of 38 ARPABET[5] phonemes. HIWIRE consists of 100 recordings per speaker, of which the first 50 utterances are commonly defined to serve as adaptation data and the second 50 utterances as testing data.

Since HIWIRE was recorded at 16 kHz, the recordings were down-sampled to 8 kHz to "match" the recording conditions of the SpeechDat(II) data. Then, the same MF-PLP feature analysis was applied and passed through each of the three MLPs (MLP-EN, MLP-sUNI and MLP-UNI) to estimate source phoneme posteriors. $P(s^k|d^l, \theta)$ and $P(d^l|\theta)$ were estimated on the adaptation data, as explained in Section 2. The testing set was used to estimate target phoneme posteriors, $P(d_t^l|\hat{x}_t, \theta)$, according to (3). The target phoneme posteriors were then divided by the priors $P(d_l|\theta)$ and a hybrid HMM/MLP system [6] was used to perform ASR.

[2] http://www.icsi.berkeley.edu/Speech/qn.html
[3] http://www.phon.ucl.ac.uk/home/sampa/
[4] http://htk.eng.cam.ac.uk/
[5] http://www.speech.cs.cmu.edu/cgi-bin/cmudict

For the sake of comparison, system MLP-AE was trained on the HIWIRE adaptation set. Target phoneme alignments were obtained with system MLP-UNI. During MLP training, 90% of the data was used for training and the remaining 10% for validation. System MLP-AE directly estimates target phoneme posteriors $P(d_t^l|\hat{x}_t)$ and does not involve an HMM-based phoneme space transformation. Thus, system MLP-AE has no access to $P(d^l|\theta)$ and makes use of the priors estimated by system MLP-UNI to perform hybrid ASR.

3.3. Results

We investigated all the systems described in Table 1 and compared them to the baseline that was reported in [5].

Table 2: *Word accuracies on the HIWIRE testing set. The baseline was reported in [5]. Systems MLP-AE, MLP-EN, MLP-sUNI and MLP-UNI are described in Table 1.*

base	MLP-AE	MLP-EN	MLP-sUNI	MLP-UNI
91.4	92.8	92.6	93.7	96.0

The baseline system used Mel-Frequency Cepstral Coefficients with Cepstral Mean Subtraction and was trained on the TIMIT database that contains read American English speech, recorded at 16 kHz. The baseline system did not use the adaptation set. System MLP-AE, yields a better performance than the baseline. The performance of system MLP-AE is not significantly different from the performance of system MLP-EN, that was trained on 12.4 hours of native English SpeechDat(II) data. For the significance test, we used the bootstrap estimation method [7] and a confidence interval of 95%. System MLP-sUNI was trained on 12.7 hours of multilingual data and significantly outperforms system MLP-EN. MLP-UNI was trained on five times more multilingual data than MLP-sUNI, which also yields significant improvement.

Table 3: *Word accuracies on the HIWIRE testing set if source phonemes are manually mapped to target phonemes.*

base	MLP-AE	MLP-EN	MLP-sUNI	MLP-UNI
-	-	83.2	83.5	88.8

Table 3 presents results for a manual mapping between source phonemes and target phonemes. We converted all involved phoneme sets to IPA[6] format and then mapped source and target phonemes that share the same IPA symbol. For each target phoneme without matching source phoneme, we manually selected the most similar source phoneme according to the IPA chart. For the complete manual mapping table, see [4].

The results from Tables 2 and 3 prove our hypothesis and confirm that the novel approach can be used to transform robust universal phoneme posteriors to monolingual phoneme posteriors and improve ASR performance on non-native speech. The huge performance gap between the proposed approach and a manual mapping shows that manually derived one-to-one mappings are detrimental to ASR systems and illustrates that target and source phoneme sets have significant differences in their coverage of $\mathcal{X}$.

3.4. Corollary

The HIWIRE database provides us with 144 minutes of adaptation data, enough to train a complete system. In Table 4, we show that the proposed approach yields equal performance with only ten minutes of adaptation data. If we use only one minute and 40 seconds of data (manually chosen to cover the whole target phoneme space), the system still yields significant improvement compared to systems MLP-AE and MLP-EN. Thus, the proposed approach has potential for fast adaptation of systems, to perform ASR for under-resourced languages.

Table 4: *Performance of system MLP-UNI with different amounts of adaptation data (in minutes).*

Data (in minutes)	144	32	10	2.7	1.7
Word accuracy	96.0	96.2	96.0	95.1	93.8

[6] http://www.langsci.ucl.ac.uk/ipa/

4. Conclusion

We proposed a stochastic phoneme space transformation approach and applied it to non-native ASR. The contribution of this paper is twofold. 1) We showed that different phoneme sets cover the same acoustic space differently and that manually derived phoneme mappings are detrimental to ASR systems. However, only ten minutes of data along with phoneme transcriptions are sufficient to transform multilingual phoneme posterior probabilities to monolingual English phoneme posterior probabilities. 2) We demonstrated that the transformed multilingual phoneme posteriors yield significant improvement on non-native ASR compared to native and non-native English systems.

In future, we intend to apply the proposed approach to ASR for under-resourced languages.

5. Acknowledgments

This research was supported by the Swiss NSF through the project Interactive Cognitive Systems (ICS) under contract number 200021_132619/1, 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).

6. References

[1] D. Imseng, H. Bourlard, M. Magimai.-Doss, and J. Dines, "Language dependent universal phoneme posterior estimation for mixed language speech recognition," in *Proc. of ICASSP*, 2011, pp. 5012–5015.

[2] T. Schultz, "Multilingual acoustic modeling," in *Multilingual Speech Processing*. Academic Press, 2006, pp. 71–122.

[3] G. Aradilla, "Acoustic models for posterior features in speech recognition," Ph.D. dissertation, Ecole Polytechnique Fédérale de Lausanne, 2008.

[4] D. Imseng *et al.*, "Improving non-native ASR through stochastic multilingual phoneme space transformations," Idiap, Tech. Rep. Idiap-RR-19-2011, June 2011.

[5] J. C. Segura *et al.*, "The HIWIRE database, a noisy and non-native English speech corpus for cockpit communication," 2007. [Online]. Available: http://cvsp.cs.ntua.gr/projects/pub/HIWIRE/WebHome/HIWIRE_db_description_paper.pdf

[6] N. Morgan and H. Bourlard, "Continuous speech recognition: an introduction to the hybrid HMM/connectionist approach," *IEEE Signal Processing Magazine*, vol. 12, no. 3, pp. 24–42, 1995.

[7] M. Bisani and H. Ney, "Bootstrap estimates for confidence intervals in ASR performance evaluation," in *Proc. of ICASSP*, vol. 1, 2004, pp. I–409–412.