Applying multi- and cross-lingual stochastic phone space transformations to non-native speech recognition

David Imseng, *Student Member, IEEE,* Hervé Bourlard, *Fellow, IEEE,* John Dines, *Member, IEEE,* Philip N. Garner, *Senior Member, IEEE,* and Mathew Magimai.-Doss, *Member, IEEE*

***Abstract*—In the context of hybrid HMM/MLP Automatic Speech Recognition (ASR), this paper describes an investigation into a new type of stochastic phone space transformation, which maps "source" phone (or phone HMM state) posterior probabilities (as obtained at the output of a Multilayer Perceptron/MLP) into "destination" phone (HMM phone state) posterior probabilities. The resulting stochastic matrix transformation can be used within the same language to automatically adapt to different phone formats (e.g., IPA) or across languages. Additionally, as shown here, it can also be applied successfully to non-native speech recognition. In the same spirit as MLLR adaptation, or MLP adaptation, the approach proposed here is directly mapping posterior distributions, and is trained by optimizing on a small amount of adaptation data a Kullback–Leibler based cost function, along a modified version of an iterative EM algorithm.**

On a non-native English database (HIWIRE), and comparing with multiple setups (monophone and triphone mapping, MLLR adaptation) we show that the resulting posterior mapping yields state-of-the-art results using very limited amounts of adaptation data in mono-, cross- and multi-lingual setups. We also show that "universal" phone posteriors, trained on a large amount of multilingual data, can be transformed to English phone posteriors, resulting in an ASR system that significantly outperforms a system trained on English data only. Finally, we demonstrate that the proposed approach outperforms alternative data-driven, as well as a knowledge-based, mapping techniques.

***Index Terms*—Non-native speech recognition, universal phone set, multilingual acoustic modeling**

I. Introduction

STATE-of-the-art Automatic Speech Recognition (ASR) systems typically use phonemes or phones as subword units. The set of all phonemes that are used to model speech in a given language is referred to as a *phoneme set*. The phoneme set 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 needs linguistic expertise and resources. However, statistical ASR systems usually focus on particular acoustic realizations of phonemes, with specific stationarity properties, which are then referred to as phones. Phones are then modeled by context-dependent or context-independent HMMs. As a consequence of this, it is often difficult to define a phonetic set that is unique to a specific language, and universally used across different ASR systems. Whilst most phonetic representations such as SAMPA [1] and ARPABET [2] can be represented using the International Phonetic Alphabet (IPA) [3], the underling phonetic lexicons do not necessarily use the same subset of IPA symbols.

Furthermore, even in the context of well defined phone sets, training phone HMM models remains 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" phone class, within and between languages, or in the case of accented speech (often borrowing phone realizations from two different languages). Several approaches have already been proposed to tackle this pronunciation variability, and to automatically adapt (in a supervised, loosely supervised, or unsupervised manner) phone HMM state emission probabilities from one (source) domain to a (destination) domain, possibly covering accented speech. As further discussed in Section II, those approaches include, among others, Probabilistic Phone Mapping (PPM) [4], Probabilistic Acoustic Mapping (PAM) [5], Maximum Likelihood Linear Regression (MLLR) [6], [7], Maximum A Posteriori adaptation (MAP) to non-native ASR [8], [9], linear MLP output or Linear Hidden Network (LHN) transformation [10].

In this paper, we propose an alternative approach, tackling some of the issues related to the acoustic modeling and multi/cross-lingual adaptation of phones, specifically crafted for HMM/MLP systems, and working directly with posterior distributions. More specifically, the phone variability problem is addressed here in the context of challenging non-native speech recognition tasks, where we also encounter phone set mismatch problems, as well as multi- and cross-lingual phone transformation requirements. The approach investigated here is indeed applied to non-native (English) speech recognition, adapting generic phone class sets initially trained on a large amount of English data, and adapted on a small amount of "destination" data to recognize accented speech (in different languages).

In such a context, and in addition to phone variability briefly discussed above, the approach proposed here can also handle (intra/inter-task, as well as intra/inter-lingual) phone set mismatches, and comparisons will be reported against conventional "manual" phone mapping (based on minimum linguistic expertise). Indeed, lexical resources that are distributed along with the databases can differ greatly, depending upon the defi-

D. Imseng, H. Bourlard are with Idiap Research Institute, Martigny, Switzerland and Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland

J. Dines, P. N. Garner and M. Magimai-Doss are with Idiap Research Institute, Martigny, Switzerland

nition and number of phones, as well as the notation adopted. One way to handle such problems is to select one notation and have a large lexicon that covers all the possible words. Since spoken language continuously evolves, new words need to be added regularly. An alternate solution is to perform a one-to-one mapping between the phone symbols. Such mappings are usually manually defined or derived in a data-driven way. However, a one-to-one mapping between different notations may not always exist. Also, as already reported by Sim [5] on an inter-lingual task and shown later in Section V for intra- and inter-lingual tasks, even if such a mapping exists, it could be detrimental to the system. The reason for this is partly related to acoustic modeling. We suppose that there exists an acoustic space that contains all acoustic observations that are involved in the human speech production process. During acoustic modeling, a specific phone set implicitly partitions this acoustic space into subspaces, each associated with a particular phone class. Of course, two different phone sets can partition the same acoustic space differently, which will not be taken into account during one-to-one mapping.

Finally, we decided to focus the present work, and evaluate the proposed approach, in the context of accented (English) speech recognition, and see how multilingual data can be most effectively exploited in this context. Indeed, cross-lingual ASR studies often mainly focus on the recognition of speech from native speakers, while effectively recognizing speech from both native and non-native speakers is still recognized as 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 phone 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 is the influence of the native language on the target language sound pronunciation [11].

In previous work [12], we boosted 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 created a universal set, by merging phones that share the same IPA symbol, and then trained 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 phone classes and the target phone set was learned by using a Kullback–Leibler divergence based HMM. The learned relation can be seen as a data-driven soft mapping between two sets that takes the acoustics into account. During recognition, the resulting stochastic mapping was then exploited to transform the conditional posterior probabilities of the universal phone classes into estimates of posterior probabilities of the phones belonging to the target database.

After a discussion of related work by others (Section II), we generalize the initial work [12] and cast it into a more rigorous theoretical framework in Section III. The training and recognition algorithms are derived in Section IV. The paper then explores mono-, cross- and multi-lingual stochastic transformations, and compares them with manual mapping and data-driven hard mapping (Sections V and VI). Results on a standard multi-lingual/accented database show that the proposed approach significantly outperforms manual and data-driven hard mappings. Furthermore, the experimental studies reveal that the transformed universal emission probabilities yield significant improvement compared to all other systems.

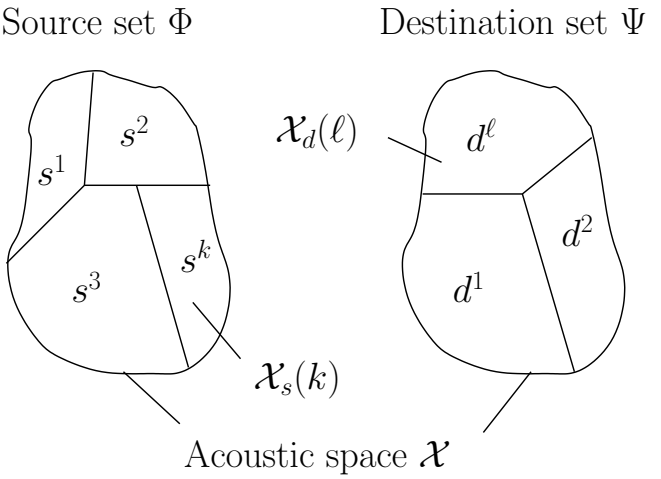


Fig. 1. Given a source language and a destination language, two different sets of phone classes cover the same acoustic space differently. $\mathcal{X}_s(k)$ and $\mathcal{X}_d(\ell)$ are acoustic subspaces associated with phone class s^k and d^ℓ respectively.

In Section VII, we then discuss the relationship between the proposed approach and related work such as PAM [5], MLLR [6], [7], LHN [10], semi-continuous hidden Markov models [13], or the estimation of language-independent acoustic models as presented by Schultz and Waibel [14] in more detail.

II. Related work

Standard ASR systems typically make use of phonemes as subword units. A phoneme is defined as the smallest sound unit of a language that discriminates between a minimal word pair [15, p. 310].

Humans are able to produce a large variety of acoustic sounds which linguists have categorized into segments called phones. Phones are not necessarily the smallest units to describe sounds but they represent a base set that can be used to describe most languages [15]. Those phone (phonetic) segments are also more stationary, hence more amenable to statistical modeling. We assume in this paper that all phones across speakers and languages, share a common acoustic space $\mathcal{X}$ (e.g., the acoustic space that could be "theoretically" covered by a human articulatory system). Of course, no single language makes use of all phones, and most languages only partially cover $\mathcal{X}$.

Therefore, as visualized in Figure 1, we assume here that for two different languages two different sets of phone classes partition the same acoustic space differently, and we define:

- A source set consisting of S phone classes $s^k, k = 1, \ldots, S$
- A target (destination) set consisting of D phone classes $d^\ell, \ell = 1, \ldots, D$

The phone classes can for example be phonemes as defined by linguists or context-dependent states that are usually used in state-of-the-art ASR systems. The assumption of a common acoustic space $\mathcal{X}$ is reasonable and usually underpins the approaches based on the pooling/adaptation of acoustic models from multiple languages [16].

Rottland and Rigoll [17] presented the tied posteriors approach, which considers the special case where the S source classes are context-independent monophones and the target classes are context-dependent triphones, both from the same language. In the present work, we focus on stochastic transformations in general, especially across languages. Furthermore, as we will describe later, we estimate the stochastic transformation matrix differently by directly using phone class posteriors instead of converting them to likelihoods and applying the maximum likelihood adaptation.

Schultz and Waibel [14] proposed an HMM-based method to estimate language-independent acoustic models. In a conventional HMM/GMM framework, each state is modeled with a mixture of Gaussian distributions. If the IPA symbol set of two context dependent states from different languages is the same, the training data of all involved languages is then used for the estimation of the Gaussian components (means and variances). The mixture weights, however, are trained for each language individually. Hence, the approach involves a transformation in the sense that each (multilingual) universal phone class has a pool of S Gaussians. The universal phone class model is then transformed to a language specific model by estimating language dependent weights. Our work focuses on hybrid HMM/MLP systems and not on standard HMM/GMM systems, but we will show later that the proposed method is closely related to conventional Gaussian mixture based semi-continuous HMM systems.

Sim and Li [4] proposed (explicit) one-to-one Probabilistic Phone Mapping (PPM) that makes use of explicit phonetic reference transcriptions (in the form of target classes) and outputs of a phone recognizer that uses source classes. As a result, PPM maps each target class to the most similar source class.

Sim [5] extended PPM to Probabilistic Acoustic Mapping (PAM) for hybrid HMM/MLP ASR systems that allows implicit transformation of source posteriors into target posteriors. The approach proposed here is similar in spirit to PAM. Both approaches are based on posterior space transformations and we compare them in detail in Section VII. Our approach is crafted into a principled theoretical formalism, allowing for EM/Viterbi-like iterative training to optimize a global KL-criterion, which is shown to be more appropriate to posterior features on the investigated non-native English database (see Section VII).

Similarly to PAM, hidden feature transformation [10] can be used to improve non-native ASR. More specifically, in a hybrid HMM/MLP framework, a linear transformation is applied to the activation of an internal layer of the MLP. The transformation is performed with a linear hidden network (LHN) which is trained with the standard MLP error back-propagation algorithm. However, since a hidden layer is adapted, LHN is bound to a fixed phoneme set.

Various studies applied acoustic model transformations to non-native ASR in the form of conventional adaptation techniques such as MLLR [6], [7] or MAP [8], [9]. More recently, combining acoustic model transformation and pronunciation modeling for non-native ASR was also investigated [18]. For acoustic model transformation, MAP and model re-estimation were evaluated and combined with pronunciation modeling that was based on phonetic rule extraction. For each sentence of the a non-native database, the canonical transcription was compared to the transcription given by a phonetic recognizer. However, if the mother tongue of the (non-native) speaker was unknown, MAP and model re-estimation alone performed better than in combination with pronunciation modeling.

In this study, we investigate a new approach to map conditional phone class probabilities from a source set to a target set, given acoustic observations. In general, we consider the source and target phone class sets to be defined in different languages (although the similar idea of stochastic mapping could also be applied to two different sets of the same language). It is evident that sets of phone classes of foreign languages have a different coverage of the acoustic space $\mathcal{X}$. In the experimental section, we will see that two different phone class sets of the same language also provide different coverage of $\mathcal{X}$. We will also compare our method to standard adaptation techniques on the HIWIRE database (non-native English).

III. Stochastic phone space transformation

In the context of hybrid HMM/MLP recognizers, stochastic phone space transformation can be formulated as follows. Given an MLP (of parameters $\Theta_{\mathcal{S}}$) trained to estimate source phone class posterior probabilities conditioned on acoustic observations, we aim to perform ASR on a target database that makes use of a target phone class set. Of course, the "source" MLP $\Theta_{\mathcal{S}}$ could also have been trained on a mixture of languages to make it more amenable to cross-language adaptation/training.

Mapping source phone class posteriors into target phone class posteriors then requires the training of a stochastic matrix of parameters Θ_M instead of an MLP, which, together with the fixed $\Theta_{\mathcal{S}}$ will parameterize target phone class posterior distributions used as emission probabilities in the HMM/MLP recognizer. During the training of Θ_M, we thus assume to have access to a limited amount of target language training data $\boldsymbol{x} = \{x_1, \ldots, x_T\}$, which is not labeled in terms of source phone classes but only in terms of target phone classes. Furthermore, we assume that no target phone class segmentation is available (i.e., we can associate a target phone class sequence with $\boldsymbol{x}$, but we have no labeling for every x_t). The proposed approach will exploit a target HMM where the states (hidden variables) are associated with the target phone class sequence (with posterior distributions resulting from the stochastic mapping of the source posteriors).

A. Definitions

Given an acoustic sequence $\boldsymbol{x} = \{x_1, \ldots, x_t, \ldots, x_T\}$ drawn from the acoustic space, where x_t is an acoustic feature

vector containing for example perceptual linear prediction coefficients. We aim to estimate the target phone class posterior probabilities $P(d_t^\ell|x_t)$, given:

1) The source MLP posteriors $P(s_t^k|x_t,\Theta)$, simply estimated by presenting x_t (possibly together with some temporal context) at the input of the MLP $\Theta_\mathcal{S}$, and
2) the conditional target posterior $P(d_t^\ell|s_t^k,x_t,\Theta)$, conditional on the current input x_t and latent variable s_t^k denoting the specific (hidden) HMM source state s^k visited at time t.

Indeed, we can formulate the problem of estimating target phone class posteriors conditioned on the acoustic observation x_t at time t, the parameters Θ_M of the target HMM, and the parameters $\Theta_\mathcal{S}$ of the source MLP as follows:

$$P(d_t^\ell|x_t,\Theta) = \sum_{k=1}^{S} P(d_t^\ell|s_t^k,x_t,\Theta)P(s_t^k|x_t,\Theta) \quad (1)$$

$$= \sum_{k=1}^{S} P(d^\ell|s^k,\Theta_M)P(s_t^k|x_t,\Theta_\mathcal{S}) \quad (2)$$

where $\Theta = \{\Theta_\mathcal{S},\Theta_M\}$, and where we have made the following assumptions:

- The conditional probability $P(d_t^\ell|s_t^k,x_t,\Theta)$ can be seen as a similarity measure between a source class s^k and a target class d^ℓ. It can thus be assumed time invariant and independent of the acoustic observation x_t at time t.
- The source phone class posteriors $P(s_t^k|x_t,\Theta)$ 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 phone class posterior probability estimates are considered independent of Θ_M.

During recognition (see Section IV-B), the target phone class posterior estimates $P(d_t^\ell|x_t,\Theta)$ can be used to perform ASR on the target database.

Since the states of the target HMM (parameterized by Θ_M, in addition to the fixed $\Theta_\mathcal{S}$ coming from the source training) will be associated with the target phone class sequence, we can only estimate $P(d^\ell|s^k,\Theta_M)$ from the source posteriors $P(s^k|d^\ell,\Theta_M)$. Applying Bayes' rule to $P(d^\ell|s^k,\Theta_M)$ in (2) yields:

$$P(d_t^\ell|x_t,\Theta) = \sum_{k=1}^{S} \frac{P(s^k|d^\ell,\Theta_M)P(d^\ell|\Theta_M)}{\sum_{\ell=1}^{D} P(s^k|d^\ell,\Theta_M)P(d^\ell|\Theta_M)} P(s_t^k|x_t,\Theta_\mathcal{S}) \quad (3)$$

Given $P(s_t^k|x_t,\Theta_\mathcal{S})$, the estimation of $P(d_t^\ell|x_t,\Theta)$ thus requires us to estimate the conditional probability $P(s^k|d^\ell,\Theta_M)$ and the prior probability $P(d^\ell|\Theta_M)$.

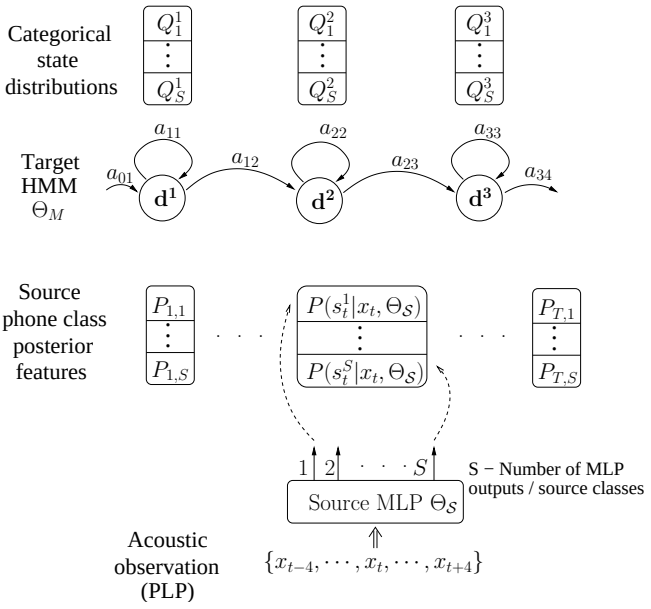


Fig. 2. The target HMM structure is "left-to-right" and obtained from the target phone class transcriptions. Each state is parametrized by a categorical distribution Q^ℓ of dimensionality S and emits posterior features P_t. The transition probabilities a_{ij} are also parameters of the HMM, but are fixed to constant values of 0.5 (except for $a_{01}=1$).

B. Estimation of the conditional probability $P(s^k|d^\ell,\Theta_M)$

Estimation of $P(s^k|d^\ell,\Theta_M)$ will be performed through an iterative Viterbi segmentation-optimization training procedure. As illustrated in Figure 2, this requires that we first forward pass all the training data $\boldsymbol{x}$ through the source MLP $\Theta_\mathcal{S}$ to obtain $P(\boldsymbol{s}|x_t,\Theta_\mathcal{S}),\forall t\in 1,\ldots,T$, as:

$$P_t = P(\boldsymbol{s}|x_t,\Theta_\mathcal{S}) = \begin{pmatrix} P(s_t^1|x_t,\Theta_\mathcal{S}) \\ \vdots \\ P(s_t^S|x_t,\Theta_\mathcal{S}) \end{pmatrix} = \begin{pmatrix} P_{t,1} \\ \vdots \\ P_{t,S} \end{pmatrix}$$

We then use P_t, with $t=1,\ldots,T$, as observed feature vectors for the target HMM, alongside with the target phone class transcriptions, to train the HMM parameters Θ_M.

In the simplest case, the target HMM uses one state per target class d^ℓ in a *left-to-right* structure, which is obtained from the destination phone class transcriptions. In Figure 2, e.g., we consider an utterance that can be transcribed as /d^1/ /d^2/ /d^3/. In this illustrative case, the associated HMM has five states including non-emitting start and end states. However, the presented algorithm is not limited to such simple HMM structures, but allows more complex ones such as using three states per phone class. For the ease of notation, and without loss of generality, we limit ourselves to the simplest case (one state per phone class) in the following derivations.

Each target HMM state d^ℓ, with $\ell\in\{1,\ldots,D\}$ (D being the number of HMM states), is thus parametrized by

[1]The MLP takes a temporal context of c preceding and following frames into account (we usually set $c=4$). However, for the ease of notation, we just write $P(s_t^k|x_t,\Theta)$.

a categorical distribution Q^ℓ.

$$Q^\ell = P(\boldsymbol{s}|d^\ell, \Theta_M) = \begin{pmatrix} P(s^1|d^\ell, \Theta_M) \\ \vdots \\ P(s^S|d^\ell, \Theta_M) \end{pmatrix} = \begin{pmatrix} Q_1^\ell \\ \vdots \\ Q_S^\ell \end{pmatrix}$$

A categorical distribution is a multinomial distribution where only one sample is drawn. The dimensionality of Q^ℓ is S, the total number of source classes.

Of course, transition probabilities a_{ij}, to go from state i to state j, should also be parameters of the target HMM, $\Theta_M = \{Q^\ell, a_{ij}\}$. However, we fixed them to constant values of 0.5 (except for $a_{01} = 1$), as is usually done in hybrid HMM/MLP systems.

Since the observed feature vectors P_t are posterior estimates and the HMM references Q^ℓ categorical distributions, it is reasonable to use the Kullback–Leibler divergence between those two as local scores, i.e.:

$$d(P_t, Q^\ell) = \sum_{k=1}^{S} P(s_t^k|x_t, \Theta_\mathcal{S}) \log \left[\frac{P(s_t^k|x_t, \Theta_\mathcal{S})}{P(s^k|d^\ell, \Theta_M)} \right] \quad (4)$$

The local score $d(P_t, Q^\ell)$ is not symmetric and we discuss the choice of $d(P_t, Q^\ell)$ as opposed to $d(Q^\ell, P_t)$ in Section VII. The resulting HMM is referred to as KL-HMM [19], which can be trained with a Viterbi optimization algorithm as presented in Section IV.

C. Estimation of the prior probability $P(d^\ell|\Theta_M)$

The trained HMM can be used to assign each x_t with an acoustic subspace $\mathcal{X}_d(\ell)$, as presented in Section IV-A1. Prior probabilities $P(d^\ell|\Theta_M)$ can thus be estimated as the relative count of acoustic vector observations x_i that are associated with $\mathcal{X}_d(\ell)$, i.e.:

$$P(d^\ell|\Theta_M) = \frac{|\{x_i|x_i \in \mathcal{X}_d(l)\}|}{\sum_{j=1}^{D} |\{x_i|x_i \in \mathcal{X}_d(j)\}|} \quad (5)$$

where the operator $|.|$ stands for the cardinality of a set.

IV. Implementation

The categorical distributions Q^ℓ can be learned using an iterative Viterbi segmentation-optimization scheme. The cost function can be defined by integrating the local score, given in (4), over time t and states ℓ, resulting in

$$\mathcal{F}(P, Q) = \sum_{t=1}^{T} \sum_{l=1}^{D} d(P_t, Q^\ell) \delta_t^\ell \quad (6)$$

where the Kronecker delta δ_t^ℓ, defined as:

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

with $\mathcal{X}_d(\ell)$ being the acoustic subspace that corresponds to a target class d^ℓ.

A. Training

As illustrated in Algorithm 1 below, the training consists of iteratively minimizing the cost function in (6) in the Q^ℓ space (optimization step) and δ_t^ℓ space (segmentation step) respectively. The segmentation is obtained by Viterbi forced alignment (Section IV-A1). We run the algorithm until convergence. Of course, convergence can easily be proved since at every segmentation and re-estimation step the same cost function is minimized, respectively in the Q^ℓ space (re-segmentation) and δ_t^ℓ (re-estimation).

Algorithm 1 HMM Training

Step 0: Initialization of Q_k^ℓ
for all $\ell \in \{1, \ldots, D\}$ and $k \in \{1, \ldots, S\}$ **do**

$$Q_k^\ell = \begin{cases} \frac{1}{S}, & \text{if } d^\ell \notin \text{Source set } \Phi \\ 1 - (S-1)\varepsilon, & \text{if } d^\ell \in \Phi \text{ and } s^k = d^\ell \\ \varepsilon, & \text{if } d^\ell \in \Phi \text{ but } s^k \neq d^\ell \end{cases}$$

ε being small, but positive
end for
Step 1: Segmentation:
Given $P_t \ \forall \ t$, perform forced alignment to assign each x_t to one $\mathcal{X}_d(\ell)$ such that $\mathcal{F}(P, Q)$ is minimized.
Step 2: Optimization:
for all $\ell \in \{1, \ldots, D\}$ **do**
Given $P_{t^*} \ \forall \ t^*$ such that $x_{t^*} \in \mathcal{X}_d(\ell)$, use (9) to estimate Q^ℓ.
end for
Iterate step 1 and 2 until convergence

1) Segmentation: To associate each x_t with one of the acoustic subspaces $\mathcal{X}_d(\ell)$, (and as a consequence P_t to Q^ℓ) the HMM aligns the source phone class posterior probability P_t with the states by minimizing $\mathcal{F}(P, Q)$, given in (6).

2) Optimization: Each P_t is used to update a particular categorical distribution Q^ℓ.

To minimize $\mathcal{F}(P, Q)$ subject to the constraint that $\sum_{k=1}^{S} Q_k^\ell = 1$, we introduce the Lagrange multiplier λ and take the partial derivative of the resulting function with respect to each variable Q_k^ℓ and set it to zero:

$$\frac{\partial}{\partial Q_k^\ell} \mathcal{F}(P_t, Q^\ell) + \lambda \left(\sum_{k=1}^{S} Q_k^\ell - 1 \right) = 0 \quad (7)$$

Solving (7) yields:

$$-\sum_{\forall t^*} \frac{P(s_{t^*}^k|x_{t^*}, \Theta_\mathcal{S})}{P(s^k|d^\ell, \Theta_M)} + \lambda = 0$$

where the sum extends over all t^* such that $x_{t^*} \in \mathcal{X}_d(\ell)$. Hence:

$$P(s^k|d^\ell, \Theta_M) = \frac{1}{\lambda} \sum_{\forall t^*} P(s_{t^*}^k|x_{t^*}, \Theta_\mathcal{S})$$

The sum to one constraint $\sum_{k=1}^{S} Q_k^\ell = 1$ guarantees:

$$\sum_{k=1}^{S} P(s^k|d^\ell, \Theta_M) = \sum_{k=1}^{S} \frac{1}{\lambda} \sum_{\forall t^*} P(s_{t^*}^k|x_{t^*}, \Theta_\mathcal{S}) = 1 \quad (8)$$

Solving (8) for λ yields:

$$\lambda = \sum_{\forall t^*} \sum_{k=1}^{S} P(s_{t^*}^k | x_{t^*}, \Theta_{\mathcal{S}}) = \sum_{\forall t^*} 1 = |\{x_i | x_i \in \mathcal{X}_d(\ell)\}|$$

We thus obtain:

$$P(s^k | d^\ell, \Theta_M) = \frac{1}{|\{x_i | x_i \in \mathcal{X}_d(\ell)\}|} \sum_{\forall t^*} P(s_{t^*}^k | x_{t^*}, \Theta_{\mathcal{S}}) \quad (9)$$

3) Initialization: For initialization, we may make use of prior knowledge as described below. However, experiments have shown that uniform initialization will usually yield similar results, although with slower convergence.

If the source and target classes are both phonemes and the IPA symbol of the destination class d^ℓ is not present in the source set, Q^ℓ is initialized uniformly. If the IPA symbol of d^ℓ and s^k are same however, all the components of Q^ℓ are set to a small positive value ε except for the corresponding component Q_k^ℓ which is set to $1 - (S-1)\varepsilon$. Since the cost function involves the computation of the KL divergence between P_t and Q^ℓ, given in (4), we need to ensure that Q^ℓ does not contain zeros.

B. Recognition

Given an acoustic test sequence $\boldsymbol{x} = \{x_1, \ldots, x_T\}$, we first use the source MLP $\Theta_{\mathcal{S}}$ to estimate the source posteriors $P_t = P(\boldsymbol{s}|x_t, \Theta_{\mathcal{S}})$. Then, we use the above described target HMM (see Figure 2) for decoding, using reference posterior vectors Q^ℓ as HMM parameters and KL-distances between P and Q as local scores. Such a KL-based HMM is often referred to as KL-HMM [19]. More specifically, this is equivalent to performing a standard Viterbi decoding with the following local distance, $\Delta_{t,\ell}$, between source posterior P_t and reference posterior Q^ℓ.

$$\Delta_{t,\ell} = \sum_{k=1}^{S} \log \left[\frac{P_{t,k}}{Q_k^\ell} \right] P_{t,k} \quad (10)$$

This recognition technique involves an implicit stochastic phone space transformation depending on Q^ℓ and P_t.

V. Experimental setup

We study the proposed approach by applying it to non-native speech recognition. We start with the hypothesis that the stochastic phone space transformation is beneficial for non-native and accented speech because we can train the source MLP ($\Theta_{\mathcal{S}}$) with large amounts of (multilingual) data and then handle the variability in pronunciations with relatively small amounts of data by learning the transformation parameters Θ_M. Therefore, we estimate source phone class posteriors on databases that contain native speech of five different languages. We estimate language specific phone class posteriors as well as universal phone class posteriors that are trained on the data of all five languages. The non-native target database uses a different phonetic lexicon, thus the estimated phone class posteriors need to be transformed. We first describe the estimation of the source posteriors along with the databases that are used and then we describe the non-native target database as well as the phone class posterior transformation. For the sake of comparison, we also describe a system only trained on non-native speech at the end of the section.

System	Source set	S	TRN data	DEV acc.
MLP-EN	SAMPA English	45	12.4 h	58.8%
MLP-ES	SAMPA Spanish	32	11.5 h	73.2%
MLP-IT	SAMPA Italian	52	11.5 h	68.6%
MLP-SF	SAMPA French	42	13.5 h	65.5%
MLP-SZ	SAMPA German	59	14.1 h	60.4%
MLP-sUNI	Universal	117	12.7 h	52.0%
MLP-UNI	Universal	117	63.0 h	57.5%
MLP-AE	ARPABET English	38	2.4 h	58.2%

TABLE I
OVERVIEW OVER THE SEVEN DIFFERENT PHONE CLASS POSTERIOR ESTIMATORS. THE TOTAL AMOUNT OF TRAINING DATA, THE FRAME ACCURACY ON THE DEVELOPMENT DATA, AS WELL AS THE SOURCE SET INCLUDING THE NUMBER OF CLASSES (S) ARE GIVEN.

A. Source posteriors

We consider six different source sets, five monolingual phone sets and a universal phone class set. To estimate the source posteriors $P(s_t^k | x_t, \Theta_{\mathcal{S}})$, we investigate seven different MLP-based posterior estimators (trained with QuickNet software [20]), one for each monolingual phone set and, for the purpose of comparison, two for the universal phone class set.

To train the seven posterior estimators, we used recordings from SpeechDat(II) in five different languages. The 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 per speaker. For the MLP training, we split the databases into training (1500 speakers), development (150 speakers) and testing (350 speakers) sets, according to the procedure described in [21].

The training data were used to train five monolingual MLPs on British English (MLP EN), Spanish (MLP ES), Italian (MLP IT), Swiss French (MLP SF) and Swiss German (MLP SZ) respectively. Two MLPs were trained to estimate universal phone class posteriors. Since all the SpeechDat(II) dictionaries use SAMPA symbols, we merged phonemes that share the same SAMPA symbol across languages to build the universal phone class set. MLP UNI (universal MLP) was trained on all the data and MLP sUNI (small universal MLP) used only one fifth of the data (randomly chosen, to match the average amount of training data available to the monolingual MLPs).

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 [22], as input. The number of parameters in each MLP was set to 10% of the number of available training frames. Table I summarizes all systems and also shows the frame accuracies on the development data. MLP-AE will be presented in Section V-C.

B. Target posteriors

To study the proposed approach, we used the HIWIRE [8] 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 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 seven MLPs (MLP-EN, MLP-ES, MLP-IT, MLP-SF, MLP-SZ, MLP-sUNI and MLP-UNI) to estimate source posteriors $P(\boldsymbol{s}|x_t, \Theta_{\mathcal{S}})$.

To perform recognition on the HIWIRE adaptation set, we estimated $P(s^k|d^\ell, \Theta_M)$ on the adaptation data. For the ease of notation we limited ourselves to one state per phone class in Section III. For the experiments however, we used three states per class.

We tuned the word insertion penalty on the adaptation data and then used the test set for evaluation.

C. Training on non-native data

For the sake of comparison, system MLP-AE was trained on the HIWIRE adaptation set, i.e. an MLP was directly trained on the HIWIRE data set to estimate target phone class posteriors $P(d_t^\ell|x_t)$. Therefore system MLP-AE does not involve an HMM-based transformation. During MLP training, 90% of the adaptation data was used for training and the remaining 10% for validation.

The training of an MLP requires frame-based alignments. However, no alignments were available for HIWIRE. Therefore, we performed forced alignment. Since we did not have acoustic models for the target classes, we used the best performing transformed models (MLP-UNI) for the alignment. System MLP-AE is a standard hybrid system and therefore also uses the forced alignment to estimate prior probabilities.

VI. Results

We investigated all the systems described in Table I. For the significance tests, we used the bootstrap estimation method [23] and a confidence interval of 95%.

A. Native English training data

First, we considered phone space transformations within the target language and compared the performance of MLP-EN, trained on native English data, to studies from other researches on the HIWIRE database and to system MLP-AE which was trained on 2.4 hours of non-native English.

As shown in Table II, system *Seguera* [8], supplied with the database, is a standard HMM/GMM ASR system [24] that uses 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 systems *Gemello* [10] are hybrid systems, trained on TIMIT, WSJ0-1 and vehic1us-ch0 (Gemello 16), and LDC Macrophone and SpeechDat Mobile (Gemello 8). MLP-AE is a hybrid system, that only used the adaptation set, i.e. the MLP was trained on HIWIRE data. MLP-EN was trained on British English SpeechDat(II) data, recorded at 8 kHz, as described in Section V. All systems were evaluated on the same test set so results should be comparable. We hypothesized that system MLP-EN would perform best.

System	Source Database	Decoding	Adapt	TST
Seguera	TIMIT	GMM/HMM	no	91.4
Gemello 16	Microphone 16 kHz	Hybrid	no	90.5
Gemello 8	Telephone 8 kHz	Hybrid	no	88.4
MLP-AE	HIWIRE	Hybrid	no	92.8
MLP-EN	SpechDat(II) EN	KL-HMM	yes	**95.0**

TABLE II
WORD ACCURACIES ON ALL THE HIWIRE TEST DATA (TST). THE SYSTEM *Seguera* WAS PRESENTED IN [8] AND THE SYSTEMS *Gemmello* IN [10]. MLP-AE USED ONLY THE HIWIRE ADAPTATION SET DURING TRAINING AND MLP-EN USES CONVERTED PHONE CLASS POSTERIORS TRAINED ON ENGLISH SPEECHDAT(II) DATA.

System	Identical IPA symbols	Percentage
MLP-EN	31	82
MLP-ES	22	58
MLP-IT	24	63
MLP-SF	24	63
MLP-SZ	24	63

TABLE III
NUMBER AND PERCENTAGE OF ENGLISH HIWIRE PHONES THAT ARE COVERED BY THE PHONETIC LEXICONS DEFINED BY VARIOUS SPEECHDAT(II) DATABASES IN THE CASE WHERE AN ONE-TO-ONE IPA MAPPING IS PERFORMED.

Even though training and evaluating a system on 8 kHz data instead of 16 kHz data is penalizing the performance, as reported in [10] and also shown in Table II, system MLP-EN still performs best. This confirms that the proposed method can successfully exploit the adaptation data and outperform a system only trained on native English data, as well as a system only trained on low amount of non-native speech data.

B. Native non-English training data

We also explored cross-lingual phone space transformations and compared the performance of MLP-ES, MLP-IT, MLP-SF and MLP-SZ on the HIWIRE data. Intuitively, we hypothesized MLP-SF to perform better on French accented data, MLP-ES to perform better on Spanish accents and so on. Additionally, for the sake of comparison, we did not train a system on Greek data (to keep one unseen non-native accent data set for testing). However, we trained a system on Swiss German (MLP-SZ), a non-native accent that is not present in the test data. The resulting monolingual phone class posteriors, trained on different SpeechDat(II) databases, were then converted to 114 English states (38 phones used in the HIWIRE lexicon, each modeled with three states).

System	Decoding	FR	GR	IT	SP	TST
MLP-ES	KL-HMM	92.6	95.1	92.4	**93.6**	93.3
MLP-IT		*93.6*	**96.1**	**93.9**	*93.4*	**94.2**
MLP-SF		**93.8**	92.7	91.7	92.1	92.8
MLP-SZ		*93.6*	*95.2*	92.4	*92.9*	93.6

TABLE IV
WORD ACCURACIES OF THE CONVERTED PHONE CLASS POSTERIORS TRAINED ON SPEECHDAT(II) DATA FROM DIFFERENT LANGUAGES (SEE TABLE I). BEST RESULTS OF EACH COLUMN ARE MARKED BOLD; ITALIC NUMBERS POINT TO RESULTS THAT ARE NOT SIGNIFICANTLY WORSE.

System	Decoding	FR	GR	IT	SP	TST
MLP-EN	KL-HMM	94.9	96.2	93.8	95.0	95.0
MLP-sUNI		95.9	96.4	95.3	93.8	95.6
MLP-UNI		**97.0**	**97.9**	**96.1**	**96.4**	**96.9**

TABLE V
WORD ACCURACIES OF THE CONVERTED PHONE CLASS POSTERIORS TRAINED ON SPEECHDAT(II) DATA. MLP-EN WAS TRAINED ON ENGLISH DATA ONLY, MLP-SUNI ON A SIMILAR AMOUNT OF MULTILINGUAL DATA AND MLP-UNI ON FIVE TIMES MORE MULTILINGUAL DATA.

Table III shows the number of HIWIRE phone classes that are covered if identical IPA symbols are merged. On average, about 60% of the HIWIRE phone classes are covered by the foreign-languages (compared to 80% for the phonetical lexicon defined by the English SpeechDat(II) database). Therefore, we expected the cross-lingual systems to perform worse than MLP-EN.

Results are presented in Table IV. The best result of each accent is marked bold. Italic numbers point to results that are not significantly worse than the best result.

As expected, MLP-SF performs best on French non-native speech, MLP-IT performs best on Italian non-native speech and MLP-ES performs best on Spanish non-native speech. The Swiss German models do not perform best on any of the accents.

System MLP-IT has the best average performance but, as hypothesized, the performance is significantly worse compared to system MLP-EN. Interestingly, Raab et al. [25] also evaluated native German, Italian, Spanish and French models on HIWIRE data. The performance they reported is lower than what we report here, but Italian still outperformed all other models.

Even though all the systems presented in Table IV perform significantly worse than system MLP-EN, the performance is satisfactory compared to the case when no adaptation data is used (*No adapt* in Table II) and when only adaptation data is used (*MLP-AE* in Table II).

C. Multilingual training data

In this section, we present results from multilingual phone space transformations, i.e. MLP-UNI and MLP-sUNI, and compare them to MLP-EN. MLP-EN and MLP-sUNI were trained on similar amounts of data. However, we expect MLP-sUNI to perform better than MLP-EN because it was trained on data from multiple languages. Furthermore, we hypothesize that MLP-UNI performs better than MLP-sUNI and MLP-EN because it was trained on large amounts of multilingual data.

Table V confirms both hypotheses and shows that the proposed approach can be used to transform robust universal phone class posteriors to monolingual phone class posteriors and improve ASR performance on non-native speech.

D. Many-to-one and one-to-one mapping

We hypothesize that one-to-one mappings between phonetical lexicons defined by different databases do not exist and expect the stochastic phone space transformation to outperform manual phone mappings as well as automatically determined one-to-one mappings. To investigate this hypothesis, we study the following transformations:

- Cross-lingual phone space transformations
- Multi-lingual phone space transformations

We define a cross-lingual phone transformation to be a transformation from a monolingual phone set (English, Italian, Spanish, Swiss French or Swiss German) to the target ARPABET English phone set. A Multi-lingual phone space transformation is a transformation where the source set consists of multilingual phone classes and the target set is the ARPABET English phone set. In both cases, we model each ARPABET phone with one HMM state (in the sections above we used three states per phone). In this section, we only use one state per phone because we will compare the approach to manual and hard mappings. Their application is not obvious if there are several states per phone.

We propose in (2) to estimate $P(d_t^l|x,\Theta)$ from the posterior estimates $P(s_t^k|x_t,\Theta_S)$ of all source phones s^k (soft decision). Alternatively, we also perform a one-to-one mapping and take a hard decision. i.e. just consider the most similar source phone. We assume that the optimal one-to-one mapping is a knowledge-driven manual mapping, i.e. mapping each target phone to the source class that shares the same IPA symbol. For each target phone without a matching source class, we manually selected the most similar one according to the IPA chart. For the sake of simplicity, we only applied the manual mapping strategy to MLP-EN, MLP-UNI and MLP-sUNI (most of the HIWIRE phones can be found in the English phone set and the universal phone class set). For information, the manual mapping is given at the end of the paper in Table XI. Bold entries highlight unmatched source and target symbols.

The results of the manual mapping experiments are given in Table VI. Compared to the soft decision, the performance loss is about 10% absolute. In all three cases, there are more source classes than target phones. Therefore, some source posteriors are just discarded. Obviously, that is a suboptimal solution and causes a degradation.

Similarly to PPM [4], we also applied a data-driven one-to-one mapping, yielding:

$$P(d_t^l|x_t,\Theta) = P(s_t^{k^*}|x_t,\Theta_S) \tag{11}$$

where the sum in (2) has been replaced by a max operator and where $k^* = \mathrm{argmax}_k\, P(d^l|s^k,\Theta_M)$. Consequently, if the number of source classes (S) and the number of target phones (D) are different, we can distinguish:

System	Soft	Hard	Manual
MLP-EN	93.3	82.1	83.2
MLP-ES	88.4	68.8	-
MLP-IT	90.7	60.7	-
MLP-SF	88.3	65.5	-
MLP-SZ	89.5	65.6	-
MLP-sUNI	94.3	69.4	81.2
MLP-UNI	96.0	61.7	87.2

TABLE VI
WORD ACCURACIES. COMPARISON OF SOFT AND HARD DECISION. HARD DECISION CAN BE SEEN AS A DATA-DRIVEN MAPPING. MANUAL (KNOWLEDGE-DRIVEN) MAPPING IS ALSO EVALUATED FOR SOME SYSTEMS (SEE TABLE XI FOR MORE DETAILS).

- $D < S$: some source posteriors are discarded
- $D > S$: multiple target phones are mapped to the same source class

Both scenarios are suboptimal for decoding. Table VI shows that the soft decision always performs substantially better than the data-driven hard mapping on the HIWIRE test set. The performance loss of the data-driven mapping is less important if the transformation is applied within the same language. If the transformation is cross- or multi-lingually applied, the performance loss is more important. Earlier studies on cross-lingual transformations also compared hard mapping (PPM) to soft mapping (PAM) and reported similar degradation (20% absolute increase in phone error rate) [5]. If the source and target set differ more, then the source posteriors estimated on the target data (MLP forward pass) tend to be ambiguous. Therefore, the distributions $P(d^l|s^k, \Theta_M)$ tend to be more ambiguous as well and miss-mappings can happen more easily. Furthermore, in case of hard mappings, all the probability mass assigned to under-represented classes other than the dominant one is lost.

Interestingly, we also note here that the performance of MLP-UNI is worse than the performance of MLP-sUNI. This may result from the fact that larger MLPs (like MLP-UNI) will be more "discriminant", yielding much lower probabilities to rare phone classes such as *nn, pp, bb, tt, dd* (see Table XI in the appendix). In those cases, the denominator of (3) tends to dominate the nominator. As a result, those rare phones will be more often used for the hard mapping. A comparison of the hard mappings of MLP-UNI and MLP-sUNI shown in Table XI confirms that the mappings mostly differ for consonants like *n, p, b, t, d*.

E. Small amount of training data

The number of parameters that need to be estimated for the stochastic transformation is relatively small. In our case, the size of the stochastic mapping matrix is $S \times D$, S being the number of source classes and D the number of target states, i.e. 117 x 114. Hence, we expect the proposed approach to perform well even for very small amounts of data. To confirm that hypothesis, we continuously decreased the amount of available data, by considering fewer utterances per speaker as seen in Table VII. For these experiments, we always used system MLP-UNI because it performed best in previous experiments.

Amount of data [min]	Considered Utterances
149	Utterances 1-50
90	Utterances 1-30
32	Utterances 1-10
16	Utterances 5-9
10	Utterances 3,5,7
3	Manually selected
2	Manually selected

TABLE VII
UTTERANCE CHOICE TO SIMULATE LOW AMOUNT OF DATA.

To have at least one acoustic sample for each target class, we could not consider all speakers anymore for datasets of less than ten minutes duration. For the creation of the dataset of 2 minutes and 40 seconds, we took the list of files of the 30-minutes dataset and manually selected the utterances required to cover the whole target set. A more sophisticated manual selection resulted in a dataset of 1 minute and 40 seconds.

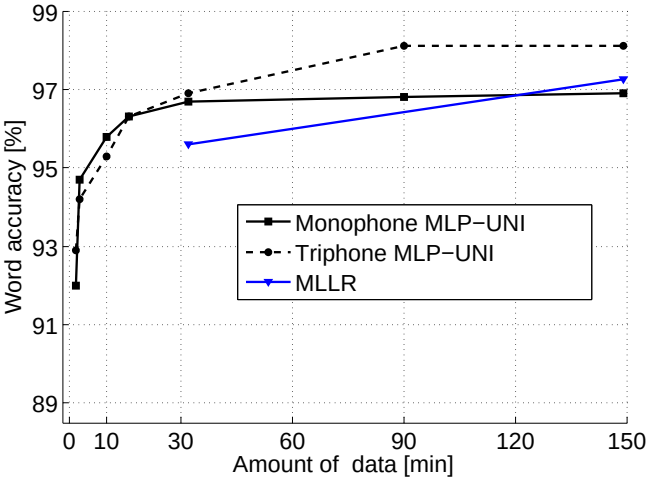


Fig. 3. Word accuracies for different amount of training data. MLP-UNI was implemented with context-independent (monophone) and context-dependent (triphone) phone mapping. As a reference point, speaker-dependent MLLR results, as reported in [8], are also given in the Figure. It is however important to keep in mind that they come from a different implementation.

We first evaluated monophone state mapping as we have done until now. Figure 3 demonstrates the efficiency of the proposed approach through excellent performance in the case of limited amounts of training data. However, it also shows that we are not able to take full advantage of the model in case of larger (typically more than 30 minutes) amounts of training data. Indeed, as already discussed at the beginning of this section, the investigated approach has a number of parameters equal to the size of the stochastic mapping matrix. However, it is always possible to increase the number of parameters by increasing the number of target states. As described in the next section, one possibility to increase the number of target states is to use context-dependent triphones instead of context-independent monophones.

F. Triphone Targets

The proposed approach is not limited to context-independent target phone classes, therefore we also investigate context-dependent target triphones. As seed models

for the context-dependent target states, we used the context-independent model of the center phone. For state tying, we applied a decision tree clustering [26] re-formulated as dictated by the KL criterion [27].

To avoid over-fitting, the stopping criterion is usually based on a combination of minimum cluster occupancy and minimum increase in log-likelihood threshold. To tune the thresholds, we took the first 30 utterances of each speaker as training set, and used utterances 31-50 of each speaker as development set. We then used the same thresholds for all our experiments.

As expected, Figure 3 shows that the MLP-UNI system benefits from the increased number of target states.

VII. Relation to similar approaches

We have shown that multilingually trained systems outperform systems trained on native English speech if we have to deal with non-native English speech during recognition. Furthermore, results revealed that very small amounts of data are sufficient to train the stochastic phone space transformation. In this section, we now discuss the local score given in (4) and show the relationship with PAM, LHN, MLLR and semi-continuous HMM systems. We also relate our approach to language-independent acoustic models presented by Schultz and Waibel [14].

A. Semi-continuous HMM

Kullback and Leibler introduced the term *discrimination information* [28], [29] which is often referred to as *Kullback-Leibler distance*[2], as defined by Cover and Thomas [30]. Since the Kullback-Leibler distance is not symmetric, Aradilla [31] used different variants of KL-based local scores for the KL-HMM framework. Given the posterior feature at time t, P_t, and the HMM state distribution of state ℓ, Q_ℓ, the following local scores have been proposed:

$$d_{KL} = d(Q^\ell, P_t) = \sum_{k=1}^{S} Q_k^\ell \log\left[\frac{Q_k^\ell}{P_{t,k}}\right] \tag{12}$$

$$d_{RKL} = d(P_t, Q^\ell) = \sum_{k=1}^{S} P_{t,k} \log\left[\frac{P_{t,k}}{Q_k^\ell}\right] \tag{13}$$

$$d_{SKL} = \frac{1}{2} d_{KL} + \frac{1}{2} d_{RKL} \tag{14}$$

where $P_{t,k} = P(s_t^k|x_t, \Theta_S)$ and $Q_k^\ell = P(s^k|d^\ell, \Theta_M)$. The local score d_{SKL} is the divergence of Kullback and Leibler [28]. Different local scores, result in different estimates for $P(s^k|d^\ell, \Theta_M)$ [31]:

$$P(s^k|d^\ell, \Theta_M) = \begin{cases} \frac{1}{C}\left[\prod_{\forall t^*} P(s_{t^*}^k|x_{t^*}, \Theta_S)\right]^{\frac{1}{\lambda}} & \text{for } d_{KL} \\ \frac{1}{\lambda}\sum_{\forall t^*} P(s_{t^*}^k|x_{t^*}, \Theta_S) & \text{for } d_{RKL} \\ \text{No closed form solution} & \text{for } d_{SKL} \end{cases}$$

where $\lambda = |\{x_i | x_i \in \mathcal{X}_d(\ell)\}|$ and C acts as a normalization constant.

[2]Although usually referred to as a divergence rather than a distance since it is not a metric.

The standard Viterbi algorithm for semi-continuous HMM systems [13] maximizes the likelihood $p(\boldsymbol{x}|\Omega)$, where $\boldsymbol{x}$ is a sequence of acoustic feature vectors and Ω are the parameters of the HMM. We consider a semi-continuous HMM similar to the HMM described in Section III build up from D state emission probability density functions d^ℓ, $\ell = \{1, \ldots, D\}$. Those distributions are usually assumed to be Gaussian mixtures and we assume to have a pool of S Gaussians s^k, $k = \{1, \ldots, S\}$. Each distribution is a linearly weighted combination of these S Gaussians. As shown in Appendix A, estimating the weights of the Gaussian mixtures along a maximum likelihood criterion is then equivalent to estimating $P(s^k|d^\ell, \Theta_M)$ if d_{RKL} is used.

B. Probabilistic Acoustic Mapping (PAM)

PAM, introduced by Sim [5, Section IV.C], estimates the target phone class probability $P(d_t^l|x_t)$ as follows:

$$P(d_t^\ell|x_t) = \frac{1}{Z} \exp\Big[\sum_{k=1}^{S} W(\ell, k) \log P(s_t^k|x_t) + b(\ell)\Big] \tag{15}$$

where Z acts as a normalization factor. W and b are the weight matrix and the bias vector of an MLP, respectively. $H(.)$ being the entropy and W^ℓ the weights associated with the l^{th} output, (15) can be rewritten:

$$P(d_t^\ell|x_t) = \frac{\exp\Big[-H(W^\ell, P_t) + b(\ell)\Big]}{\sum_{j=1}^{D} \exp\Big[-H(W^j, P_t) + b(j)\Big]} \tag{16}$$

If the MLP is trained with the cross entropy criterion, the local score d_{PAM} that is minimized can be written as:

$$d_{PAM} = -\log[P(d_t^\ell|x_t)] \propto \Big[H(W^\ell, P_t) - b(\ell)\Big] \tag{17}$$

We can rewrite (13) and (12) in terms of the entropy

$$d_{RKL} = H(Q^\ell, P_t) - H(P_t, P_t) \tag{18}$$

$$d_{KL} = H(Q^\ell, P_t) - H(Q^\ell, Q^\ell) \tag{19}$$

Hence, d_{KL} and d_{PAM} are closely related and $H(Q^\ell, Q^\ell)$ in d_{KL} acts as a target dependent bias. For d_{RKL} however, the bias is source dependent (P_t).

In the following, we summarize the differences between our approach, with d_{RKL} as local score, and PAM. Table VIII shows how these differences affect the word accuracy (WACC).

- Cost function: d_{RKL} performs better than d_{KL}, which performs similar to d_{PAM}.
- Embedded re-alignment: both, PAM and the proposed approach allow to benefit from re-alignment. In the case of PAM, a re-alignment requires the MLP to be retrained. As seen in Table VIII, PAM with re-alignment yields a better performance than PAM without re-alignment.
- Context-dependent models: in theory, both approaches can benefit from context-dependent models. In practice however, due to data sparsity, usually state tying is required. We developed an algorithm to perform state tying at the KL-HMM state level [27]. In the case of PAM, it is not obvious how to tie MLP outputs to train a context-dependent recognizer on limited amounts of data.

System	Score	Re-align	Linear	Context	WACC
KL-mono	d_{KL}	embedded	yes	no	96.7%
KL-tri	d_{KL}	embedded	yes	yes	97.6%
PAM	d_{PAM}	no	yes	no	96.2%
PAM	d_{PAM}	yes	yes	no	96.9%
PAM	d_{PAM}	no	no	no	97.1%
PAM	d_{PAM}	yes	no	no	97.4%
RKL-mono	d_{RKL}	embedded	yes	no	96.9%
RKL-tri	d_{RKL}	embedded	yes	yes	**98.1%**

TABLE VIII
WORD ACCURACIES (WACC) ON THE TEST DATA OF THE HIWIRE DATA SET. FOR ALL THE EXPERIMENTS ALL THE ADAPTATION DATA WAS USED FOR TRAINING. THE KL-HMM SYSTEMS USE MLP-UNI AS A FEATURE EXTRACTOR. LINEAR PAM CONSISTS OF A TWO-LAYER MLP AND NON-LINEAR PAM OF A THREE-LAYER MLP AS DESCRIBED IN [5].

Note that the optimal number of hidden units for the non-linear PAM approach was 800-900 in [5]. To evaluate whether more hidden units yield a better performance, we doubled the amount of hidden units and found a marginal improvement. Therefore, we report the performance of the latter configuration in Table VIII. We also investigated more than one re-alignment iteration for PAM, but did not observe further improvement.

C. Linear Hidden Network (LHN)

Another MLP-based adaptation approach performs a hidden feature transformation with an LHN [10]. The LHN is applied to the activations of the internal layer and can be trained using the standard back-propagation algorithm while keeping frozen the weights of the original network. Once the LHN is trained, it is combined with the original (unadapted) weights:

$$W_a = W_{LHN} \times W_{ORIG}$$
$$b_a = b_{LHN} \times W_{ORIG} + b_{ORIG}$$

where W_a and b_a are the weights and the bias of the adapted layer, W_{ORIG} and b_{ORIG} are the weight and bias of the layer following the LHN in the original unadapted network, and W_{LHN} and b_{LHN} are the weight and the biases of the linear hidden network.

Our approach differs from LHN in all the points already listed at the end of Section VII-B. Additionally, LHN is bound to a given and fixed phoneme set. Based on hidden layer adaptation, it is not obvious how to apply phone space transformations. To use an already trained *original* MLP, it needs to be trained from aligned data that makes use of the same phoneme set (targets) than the adaptation data.

Gemello et al. [10] used LHN to adapt an MLP, previously trained on native English, to the HIWIRE data. They investigated speaker-based adaptation (one LHN per speaker) and data-based adaptation (one LHN for all data). As shown in Table IX, the data-based LHN results in similar performance than the triphone MLP-UNI system (RKL-tri). For the speaker-based LHN adaptation, they adapted and tested for each speaker separately. Not every speaker pronounced each phone in the first 50 utterances (adaptation set). Therefore, we were not able to investigate the triphone MLP-UNI system on a per-speaker basis. However, a context-independent system (RKL-mono) can still be trained if there is no data for some target classes. RKL-mono outperforms the speaker-based LHN. Note that the results in Table IX are only given as a reference point since the proposed approach was trained on 8kHz multilingual data, and LHN on 16 kHz English data.

System	Adaptation	MLP trained on	WACC
LHN	Speaker-based	English 16 kHz	95.4 %
LHN	Data-based	English 16 kHz	98.2 %
RKL-mono	Speaker-based	Multilingual 8 kHz	96.1 %
RKL-tri	Data-based	Multilingual 8 kHz	98.1 %

TABLE IX
COMPARISON OF WORD ACCURACIES (WACC) ON THE TEST DATA OF THE HIWIRE DATA SET. AS AN ADDITIONAL REFERENCE POINT, WE SHOW THE LHN RESULTS REPORTED IN [10]. HOWEVER, THE RESULTS ARE ONLY CONDITIONALLY COMPARABLE SINCE THE PROPOSED APPROACH (RKL-MONO AND RKL-TRI) WAS TRAINED ON 8KHZ MULTILINGUAL DATA (MLP-UNI), AND THE LHN SYSTEMS ON 16 KHZ ENGLISH DATA.

D. Maximum Likelihood Linear Regression (MLLR)

MLLR has been widely used to perform acoustic model adaptation for HMM/GMM based recognizers. Seguera et al. [8] also applied conventional MLLR speaker adaptation with HTK to adapt models trained on TIMIT to the HIWIRE database.

To give another reference point, we applied the manual mappings given in Table XI to perform speaker-based MLLR with HTK as described in [8] and adapt the SpeechDat(II) English and multilingual seed models to HIWIRE.

It can be seen in Table X that the multilingual data does not improve the word accuracy on HIWIRE if MLLR is used. We attribute the performance difference between MLLR on TIMIT and SpeechDat(II) English to the different nature of the data such as sampling frequency, microphone, and background noise. Recall that we have already seen a similar degradation if 8 kHz telephone data is used instead of 16 kHz microphone data (Table II and reported by Gemello et al. [10]).

System	Seed trained on	MLP	kHz	WACC
MLLR	TIMIT	-	16	97.3%
MLLR	SpeechDat(II) English	-	8	95.7%
MLLR	SpeechDat(II) multilingual	-	8	95.7%
RKL-tri	SpeechDat(II) English	MLP-EN	8	97.2%
RKL-tri	SpeechDat(II) multilingual	MLP-UNI	8	**98.1%**

TABLE X
WORD ACCURACIES (WACC) ON THE TEST DATA OF THE HIWIRE DATA SET. FOR ALL THE EXPERIMENTS ALL THE ADAPTATION DATA WAS USED FOR TRAINING. RESULTS ON TIMIT WERE REPORTED IN [8]. FOR THE KL-HMM SYSTEMS, WE ALSO LIST THE MLP THAT WAS USED AS A FEATURE EXTRACTOR.

E. Language-independent acoustic models

We can compare our work to the estimation of language-independent acoustic models, as presented by Schultz and Waibel [14]. In this HMM-based method, they propose to estimate language-independent acoustic models, the probability $p(x_t|s_i)$ to emit x_t in a context-dependent state s_i is modeled

by a mixture of Gaussians:

$$p(x_t|s_i) = \sum_{k=1}^{S} c_{s_i k} \mathcal{N}(x_t|\mu_{s_i,k}, \Sigma_{s_i,k})$$

Given two context dependent states s_i and s_j from different languages, the Gaussian components μ and Σ are shared across languages if the IPA symbol of s_i and s_j are the same, and the training data of all involved languages is used for the estimation of the Gaussian components. The mixture weights however, are trained for each language individually.

$$\begin{aligned} c_{s_i} \neq c_{s_j} \quad &, \quad \forall i \neq j \\ \mu_{s_i,k} = \mu_{s_j,k} \quad &, \quad \forall i,j : \text{ipa}(s_i) = \text{ipa}(s_j) \\ \Sigma_{s_i,k} = \Sigma_{s_j,k} \quad &, \quad \forall i,j : \text{ipa}(s_i) = \text{ipa}(s_j) \end{aligned}$$

Hence, the above approach uses a pool of S Gaussians for each universal phone class. In that case, language specific phone class models are then obtained by estimating language dependent weights (similarity measures between universal classes and mono-lingual phones).

To compare the language-independent acoustic modeling approach to our method, we can convert the universal phone class posteriors of system MLP-UNI to any language. Thus, the proposed system can be seen as a discriminative approach of estimating language-independent acoustic models.

VIII. Conclusion

In the specific context of accented speech recognition, involving high phone acoustic variability and phone set mismatches between (multilingual) phone sets, we proposed and evaluated an alternative posterior-based stochastic phone space transformation approach.

The proposed approach adapts a stochastic mapping matrix, the elements of which can be trained in the context of an EM or Viterbi like algorithm on small amounts of multi- and cross-lingual adaptation data. The resulting algorithm iteratively optimizes a principled KL-based function, which is believed to be more amenable to posterior distributions (and does not need to turn posteriors into scaled likelihood estimates).

The resulting system has been shown to be able to efficiently exploit multi- and cross-lingual adaptation data, using a parsimonious number of parameters while also being particularly well suited in the case of phone set mismatch. This conclusion is further supported by additional evidence and theoretical and experimental comparisons with similar approaches such as PAM, LHN and MLLR.

On the HIWIRE dataset, we successfully applied the phone space transformation in mono-, cross- and multi-lingual setups and demonstrated that the proposed approach fundamentally outperforms other data-driven transformations, as well as a knowledge-based mapping technique. Ten minutes of data along with word transcriptions were sufficient to successfully convert multilingual source phone class posterior probabilities given acoustic observations, to monolingual target phone class posterior probabilities. The multilingually trained system significantly outperforms a monolingual (English) system on non-native English ASR.

We have now started investigating how the proposed phone space transformation can exploit larger datasets, while also exploring its potential applications to improve ASR for under-resourced languages.

Appendix A

The standard Viterbi algorithm for semi-continuous HMM systems [13] maximizes the likelihood $p(\boldsymbol{x}|\Omega)$, where $\boldsymbol{x}$ is a sequence of acoustic feature vectors and Ω are the parameters of the HMM. To compare semi-continuous HMMs to the proposed approach, we consider a semi-continuous HMM similar to the HMM described in Section III build up from D state emission probability density functions d^ℓ, $\ell = \{1, \ldots, D\}$. In standard ASR systems, those distributions are usually assumed to be Gaussian mixtures. Here, we assume to have a pool of S Gaussians s^k, $k = \{1, \ldots, S\}$. Each distribution is a linearly weighted combination of these S Gaussians. We thus assume the following probabilistic model:

$$p(x_t|\Omega, d^\ell) = \sum_{k=1}^{S} c_k^\ell p_k(x_t|\Omega_k) \tag{20}$$

where $p(x_t|\Omega, d^\ell)$ stands for the likelihood of an acoustic observation x_t, given the state d^ℓ and the parameters $\Omega = \{c_k^\ell, \Omega_k\}$, where $\Omega_k = \{\mu_k, \Sigma_k\}$, with μ_k being the mean and Σ_k the variance of the Gaussian s^k. We assume that Ω_k is given $\forall k$ and only c_k^ℓ needs to be estimated. Thus, the maximum likelihood solution consists of:

- Segmentation: assigning acoustic observations x_t to one of the states modeled with the mixture distribution d^ℓ, i.e. assign x_t to an acoustic subspace $\mathcal{X}_d(\ell)$,
- Optimization: given a segmentation, optimize c_k^ℓ by maximizing (20) for each distribution d^ℓ

The well-known maximum likelihood solution for c_k^ℓ (see, e.g., Bilmes [32]) is given by:

$$c_k^\ell = \frac{1}{|\{x_i|x_i \in \mathcal{X}_d(\ell)\}|} \sum_{\forall x_{t^*} \in \mathcal{X}_d(\ell)} p(s^k|x_{t^*}, \Omega_k) \tag{21}$$

Exploiting (21) and (9), saying:

$$P(s^k|d^\ell, \Theta_M) = \frac{1}{|\{x_i|x_i \in \mathcal{X}_d(\ell)\}|} \sum_{\forall t^*} P(s_{t^*}^k|x_{t^*}, \Theta_S)$$

where the sum extends over all t^* such that $x_{t^*} \in \mathcal{X}_d(\ell)$, it thus follows that estimating c_k^ℓ along a maximum likelihood criterion is equivalent to estimating $P(s^k|d^\ell, \Theta_M)$ if the reversed KL-divergnce d_{RKL}, proposed in Section III, is used.

In that particular context, our approach is then similar to semi-continuous HMM systems [13]. In contrast to the generatively trained models, usually used in semi-continuous HMM systems, we use discriminatively trained MLPs to estimate phone class posterior probabilities.

Acknowledgment

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

APPENDIX B
PHONEME MAPPINGS

HIWIRE	UNI			EN	
	man	hard (UNI)	hard (sUNI)	man	hard (EN)
m	m	m	m	m	m
n	n	**nn**	n	n	n
ŋ	ŋ	ŋ	ŋ	ŋ	ŋ
p	p	**pp**	p	p	p
b	b	**bb**	b	b	b
t	t	**tt**	t	t	t
d	d	**dd**	d	d	d
k	k	k	k	k	k
g	g	g	g	g	g
f	f	f	f	f	f
v	v	v	v	v	v
θ	θ	**pf**	**pf**	θ	θ
ð	ð	ð	ð	ð	ð
s	s	**ss**	**ss**	s	s
z	z	**dz**	**ʒ**	z	z
ʃ	ʃ	**ʃʃ**	**ʃʃ**	ʃ	ʃ
h	h	h	h	h	h
ɹ	ɹ	ɹ	ɹ	ɹ	ɹ
j	j	**jj**	**ʎ**	j	j
l	l	**ll**	**ll**	l	l
w	w	w	w	w	w
tʃ	tʃ	tʃ	tʃ	tʃ	tʃ
dʒ	dʒ	dʒ	dʒ	dʒ	dʒ
i	i	**iː**	**iː**	**iː**	**iː**
u	u	**uː**	**uː**	**uː**	**uː**
ɪ	ɪ	ɪ	ɪ	ɪ	**iː**
ɛ	ɛ	**eə**	**eə**	**e**	**eə**
ɝ	**ɜː**	**œ**	**œ**	**ɜː**	**ɜː**
ʌ	ʌ	**ɐ̃**	**aːɐ**	ʌ	**ɑː**
ɔ	ɔ	**oːɐ**	**ɔɐ**	**ɔː**	**ɔː**
æ	æ	æ	æ	æ	æ
ɑ	ɑ	**aː**	**aː**	**ɑː**	**ɑː**
eɪ	eɪ	**eː**	**eː**	eɪ	eɪ
oʊ	**əʊ**	**oː**	**oː**	**əʊ**	**ɔː**
ɔɪ	ɔɪ	ɔɪ	ɔɪ	ɔɪ	ɔɪ
aʊ	aʊ	aʊ	aʊ	aʊ	aʊ
aɪ	aɪ	aɪ	aɪ	aɪ	aɪ

TABLE XI
KNOWLEDGE DRIVEN (MANUAL MAPPING) AND DATA-DRIVEN (HARD DECISION MAPPING) OF THE DESTINATION PHONEMES (HIWIRE) TO THE ENGLISH (EN) AND UNIVERSAL (UNI) SOURCE PHONEMES (SPEECHDAT). BOLD SYMBOLS ARE DIFFERENT FROM THE DESTINATION PHONEME SYMBOL. ALL SYMBOLS ARE IN IPA FORMAT.

the European Community's Seventh Framework Programme (FP7/2007-2013) grant agreement 213845 (the EMIME project: www.emime.org).

REFERENCES

[1] *Speech Assessment Methods Phonetic Alphabet (SAMPA)*, http://www.phon.ucl.ac.uk/home/sampa/.

[2] *CMU Pronouncing Dictionary*, http://www.speech.cs.cmu.edu/cgi-bin/cmudict.

[3] *International Phonetic Alphabet*, http://www.langsci.ucl.ac.uk/ipa/.

[4] K. C. Sim and H. Li, "Robust phone set mapping using decision tree clustering for cross-lingual phone recognition," in *Proceedings of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, 2008, pp. 4309–4312.

[5] K. C. Sim, "Discriminative product-of-expert acoustic mapping for cross-lingual phone recognition," in *Proceedings of the Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2009, pp. 546–551.

[6] C. Leggetter and P. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hmms," *Computer Speech and Language*, vol. 9, pp. 171–186, 1995.

[7] M. Gales, "Maximum likelihood linear transformation for HMM-based speech recognition," *Cambridge University Engineering Department, Report CUED/F-INFENG/TR291*, 1997.

[8] 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

[9] Z. Wang, T. Schultz, and A. Waibel, "Comparison of acoustic model adaptation techniques on non-native speech," in *Proceedings of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, 2003, pp. 540–543.

[10] R. Gemello, F. Mana, and S. Scanzio, "Experiments on HIWIRE database using denoising and adaptation with a hybrid HMM-ANN model," in *Proceedings of Interspeech*, 2007, pp. 2429–2432.

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

[12] D. Imseng, H. Bourlard, J. Dines, P. N. Garner, and M. Magimai.-Doss, "Improving non-native ASR through stochastic multilingual phoneme space transformations," in *Proceedings of Interspeech*, 2011, pp. 537–540.

[13] X. D. Huang and M. A. Jack, "Semi-continuous hidden markov models for speech signals," *Computer Speech and Language*, vol. 3, no. 3, pp. 239–251, 1989.

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

[15] B. Gold and N. Morgan, *Speech and Audio Signal Processing: Processing and Perception of Speech and Music*. John Wiley & Sons, Inc, 2000.

[16] Burget *et al.*, "Multilingual acoustic modeling for speech recognition based on subspace gaussian mixture models," in *Proceedings of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, 2010, pp. 4334–4337.

[17] J. Rottland and G. Rigoll, "Tied posteriors: an approach for effective introduction of context dependency in hybrid NN/HMM LVCSR," in *Proceedings of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, vol. 3, 2000, pp. 1241–1244.

[18] G. Bouselmi, D. Fohr, and I. Illina, "Multilingual recognition of non-native speech using acoustic model transformation and pronunciation modeling," *International Journal of Speech Technology*, pp. 1–11, 2012.

[19] G. Aradilla, H. Bourlard, and M. Magimai.-Doss, "Posterior features applied to speech recognition tasks with user-defined vocabulary," in *Proceedings of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, 2009, pp. 3809–3812.

[20] *Quicknet*, http://www.icsi.berkeley.edu/Speech/qn.html.

[21] D. Imseng, H. Bourlard, and M. Magimai.-Doss, "Towards mixed language speech recognition systems," in *Proceedings of Interspeech*, 2010, pp. 278–281.

[22] *Hidden Markov Model Toolkit*, http://htk.eng.cam.ac.uk/.

[23] M. Bisani and H. Ney, "Bootstrap estimates for confidence intervals in ASR performance evaluation," in *Proceedings of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, vol. 1, 2004, pp. I–409–412.

[24] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[25] M. Raab, R. Gruhn, and E. 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.

[26] S. J. Young, J. J. Odell, and P. C. Woodland, "Tree-based state tying for high accuracy acoustic modelling," in *Proceedings of the workshop on Human Language Technology*, 1994, pp. 307–312.

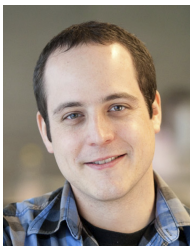
[27] D. Imseng, J. Dines, P. Motlicek, P. N. Garner, and H. Bourlard, "Comparing different acoustic modeling techniques for multilingual boosting," in *Proceedings of Interspeech*, 2012, p. to appear. [Online]. Available: http://publications.idiap.ch/downloads/papers/2012/Imseng_INTERSPEECH_2012.pdf

[28] S. Kullback and R. A. Leibler, "On information and sufficiency," *The Annals of Mathematical Statistics*, vol. 22, no. 1, pp. 79–86, 1951. [Online]. Available: http://projecteuclid.org/DPubS/Repository/1.0/Disseminate?view=body&id=pdf_1&handle=euclid.aoms/1177729694

[29] S. Kullback, "The Kullback-Leibler distance," *The American Statistician*, vol. 41, no. 4, pp. 340–341, November 1987, in Letters to the Editor.

[30] T. Cover and J. Thomas, *Elements of information theory*. New York: Wiley, 1991.

[31] G. Aradilla, "Acoustic models for posterior features in speech recognition," Ph.D. dissertation, Ecole Polytechnique Fédérale de Lausanne, 2008.
[32] J. Bilmes, "A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models," International Computer Science Institute, Berkeley, Tech. Rep. TR-97-021, April 1998.

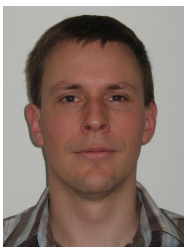


David Imseng received the B.Sc. and M.Sc. degree from Ecole Polytechnique Fédérale de Lausanne Switzerland, in 2006 and 2009, respectively. From September 2008 to April 2009, and from September 2012 to December 2012, he was a visiting scholar at the International Computer Science Institute (ICSI) Berkeley where he was working on speaker diarization and speech recognition, respectively. Since May 2009 he has been pursuing his Ph.D. at Idiap Research Institute, Martigny Switzerland, where he is currently working on multilingual speech recognition. His research interests include Speech and Speaker Recognition and Machine Learning.

Hervé Bourlard received the Electrical and Computer Science Engineering degree and the PhD degree in Applied Sciences both from "Faculté Polytechnique de Mons", Mons, Belgium. After having been a member of the Scientific Staff at the Philips Research Laboratory of Brussels and an R&D Manager at L&H SpeechProducts, he is now Director of the Idiap Research Institute, Full Professor at the Swiss Federal Institute of Technology at Lausanne (EPFL), and (Founding) Director of a Swiss NSF National Centre of Competence in Research on "Interactive Multimodal Information Management". Having spent (since 1988) several long-term and short-term visits (initially as a Guest Scientist) at the International Computer Science Institute (ICSI), Berkeley, CA, he is now a member of the ICSI Board of Trustees.

His main research interests mainly include statistical pattern classification, signal processing, multi-channel processing, artificial neural networks, and applied mathematics, with applications to a wide range of Information and Communication Technologies, including spoken language processing, speech and speaker recognition, language modeling, multimodal interaction, augmented multi-party interaction, and distant group collaborative environments.

H. Bourlard is the author/coauthor/editor of 6 books and over 300 reviewed papers (including one IEEE paper award) and book chapters. He is (or has been) a member of the program/scientific committees of numerous international conferences (e.g., General Chairman of IEEE Workshop on Neural Networks for Signal Processing 2002, Co-Technical Chairman of IEEE ICASSP 2002, General Chairman of Interspeech 2003) and on the Editorial Board of several journals (e.g., past co-Editor-in-Chief of "Speech Communication"). He is the recipient of several scientific and entrepreneurship awards.



John Dines graduated with first class honours in Electrical and Electronic Engineering from University of Southern Queensland in 1998 and received the Ph.D. degree from the Queensland University of Technology in 2003 with the thesis: "Model based trainable speech synthesis and its applications". Since 2003 he has been employed at the Idiap Research Institute, Switzerland, where he has been working mostly in the domain of meeting room speech recognition. A major focus of his current research is combining his background in speech recognition and speech synthesis to further advance technologies in both domains. He is a member of IEEE and a reviewer for IEEE Signal Processing Letters and IEEE Transactions on Audio, Speech and Language Processing.

Philip N. Garner received the degree of M.Eng. in Electronic Engineering from the University of Southampton, U.K., in 1991, and the degree of Ph.D. (by publication) from the University of East Anglia, U.K., in 2012. He first joined the Royal Signals and Radar Establishment in Malvern, Worcestershire working on pattern recognition and later speech processing. In 1998 he moved to Canon Research Centre Europe in Guildford, Surrey, where he designed speech recognition metadata for retrieval. In 2001, he was seconded (and subsequently transfered) to the speech group at Canon Inc. in Tokyo, Japan, to work on multilingual aspects of speech recognition and noise robustness. As of April 2007, he is a senior research scientist at Idiap Research Institute, Martigny, Switzerland, where he continues to work in research and development of speech recognition, synthesis and signal processing. He is a senior member of the IEEE, and has published internationally in conference proceedings, patent, journal and book form as well as serving as coordinating editor of ISO/IEC 15938-4 (MPEG-7 Audio).

Mathew Magimai Doss received the Bachelor of Engineering (B.E.) in Instrumentation and Control Engineering from the University of Madras, India in 1996; the Master of Science (M.S.) by Research in Computer Science and Engineering from the Indian Institute of Technology, Madras, India in 1999; the PreDoctoral diploma and the Docteur ès Sciences (Ph.D.) from Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland in 2000 and 2005, respectively. He was a postdoctoral fellow at International Computer Science Institute (ICSI), Berkeley, USA from April 2006 till March 2007. Since April 2007, he has been working as a Research Scientist at Idiap Research Institute, Martigny, Switzerland. His research interests include speech processing, automatic speech recognition, automatic speaker recognition, spoken language processing, signal processing, statistical pattern recognition and artificial neural networks.