Evaluating Intra- and Crosslingual Adaptation for Non-native Speech Recognition in a Bilingual Environment

György Szaszák, Philip N. Garner
Idiap Research Institute, Martigny, Switzerland
Email: gyorgy.szaszak@idiap.ch

***Abstract*—In this paper, use of intra- and crosslingual adaptation is addressed in cognitive infocommunication, for an ASR application in a bilingual environment. State-of-the-art linear regression based adaptation approaches are evaluated after a brief theoretical overview of the applied techniques. As expected, these contribute to significant improvement when used for intralingual adaptation. A simple phoneme mapping based approach is investigated for crosslingual adaptation between French and German in order to evaluate whether non-native data can help speech recognition. It is found that using native data for adapting a speech recognizer operating in the non-native language of the speaker, a modest improvement can be reached. However, when adaptation data is available from the speaker in its non-native language, it remains a better source of adaptation.**

I. Introduction

The MediaParl project aims the development of a bilingual speech database [1] and of an infocommunication application (an information retrieval system), available online [2], where video indexing of Parliament debates and interventions is provided. The indexing is based on audio transcription of the recorded material, coming from the Cantonal Parliament of the Canton of Valais in Switzerland. Audio transcripts are generated by using monolingual and/or bilingual Automatic Speech Recognition (ASR) systems and optionally a language identification module. A detailed description of the system is available in [2]. Since the MediaParl database was released by Idiap in 2012, the publicly available MediaParl infocummunication system[1] has been built.

Valais is a bilingual canton of Switzerland, with two official languages: French and German (approx. 1/3 of the population speaks German, and 2/3 speaks French as first language [1]). Whilst the French is relatively close to the standard French spoken in France, the German has a local dialect, which is often hard to understand even for German speaking people coming from another canton or state. In formal interactions, usually standard German (Hochdeutsch) is spoken, but even this is highly influenced by local accent. The automatic speech indexing system is developed for the Cantonal Parliament of Valais, where both languages are in active use. Some of the speakers are either bilingual or non-native. A bilingual speaker is regarded as speaking in both first and second languages at mother tongue level, whereas non-native speakers speak their second language with less or more characteristic accent.

[1] www.idiap.ch/webapps/webgrandconseil

This paper focuses on the ASR module used in the Mediaparl infocommunicaton system, with special attention paid to the language environment. When working in a bi- or multilingual environment, especially if the language environment is rich in dialects and accents like in Valais, speaker normalization and or/speaker adaptation are essential, as speech data shows very high variability. Several well-known baseline methods exist to perform speaker normalization or adaptation, many of them have already been successfully used in speech indexing or speech transcription systems, like Vocal Tract Length Normalization (VTLN) [3] or Cepstral Mean and/or Variance Normalization [4], [5]; for adaptation, Maximum Likelihood Linear Regression (MLLR) [6] and Constrained MLLR (CMLLR) [7], Maximum a Posteriori Adaptation (MAP) [8], and some other approaches modifying or combining these basic algorithms have been proposed. Structured MAP or Structured MAP Linear Regression (SMAPLR) [9], or its constrained form, CSMAPLR combine linear regression and the MAP expectation maximalization criterion, by exploiting some prior knowledge. These latter methods have benn investigated so far mostly for adaptation in speech synthesis [10], but they will be evaluated in ASR in this paper.

This paper does not focus on enhancing existing normalization and adaptation methods, but rather evaluating them within the current environment in order to find the best performing ones, with a special focus on speaker adaptive training [11]. The working hypothesis is that these standard adaptation approaches and their combination can lead to significant improvement in speech recognition performance, as available and above cited literature also implies this.

In the focus of the present study, crosslingual adaptation is also evaluated for non-native speech. Crosslingual adaptation is an important technique in text-to-speech (TTS) systems, especially in speech-to-speech translation, whereby TTS speech can be adapted to sound like the original speaker speaking in the target language [12], [13], even if the speaker does not speak the target language at all. Crosslingual adaptation involves the mapping of the phoneme sets, initially different for the source and target languages. This mapping can be rule-based or data-driven, like in [14], where regression trees were mapped from source to target language, allowing that each model (GMM) in the source language become mapped to the most similar target model or cluster of target models. Using crosslingual adaptation in ASR is less frequent, indeed, exploitability of such techniques is more restricted than in TTS.

978-1-4799-1546-0/13/$31.00 ©2013 IEEE

In [15], adaptation transforms for ASR models generated in the first language (L1) are used in the second language (L2) of bilingual speakers, yielding some improvement when L1 is Finnish and L2 is English, but not vice-versa.

The present work focuses on a special aspect of crosslingual adaptation, aiming to improve non-native speech recognition. The working hypothesis is that for non-native speakers, speech recognition performance can be improved by using mother tongue data adequately mapped to second language, supposing that accent perceived in the second language of a non-native speaker is governed mainly or in part by articulatory cues and patterns characteristic of the mother tongue, captured by the adaptation transform.

This paper is organized as follows: first, databases are briefly described and a short theoretical overview of adaptation methods is provided. Thereafter, linear regression based adaptations and speaker adaptive training are exhaustively evaluated in a French ASR task. Crosslingual adaptation experiments are presented next, and, finally, conclusions are drawn.

II. Data and baselines

A. The MediaParl Database

The MediaParl database is used to train ASR acoustic models. It consists of political debates recorded at the cantonal parliament of Valais. Debates take place always in the same room, they are recorded with distant talker microphones. The recordings mostly contain prepared speeches in both languages. Compared to similar multi- or bilingual databases, MediaParl stands out because of its size as it contains 20 hours of speech in both German and French.

A detailed description about the MediaParl database is given in [1], here only some basic characteristics of data are presented: audio recordings were formatted as MPEG ADTS, layer III, v1, 128 kbps, 44.1 kHz, mono, 16 bits, but then converted to WAVE audio, PCM, 16 bit, mono, 16 kHz. The database is split into train, development and test sets. The test set contains all bilingual speakers who actively used both languages (7 speakers, 2,446 utterances) to allow for bilingual evaluation. The remaining speakers are split into training set (180 speakers, 11,425 utts) and development set (17 speakers, 1,525 utts). In case of adaptation experiments, test set was further split into adaptation set (148 French, 156 German utts) and final test set (925 French, 1,521 German utts, approx. 31k ans 33k words respectively).

B. ASR Baselines

For the evaluation of speaker adaptation methods, a monolingual French ASR is trained. Feature extraction yields 39 Mel-Frequency Perceptual Linear Prediction (PLP) features (C0-C12+Δ+$\Delta\Delta$). Cepstral Mean and Variance Normalization (CMVN) is applied. Models are HMM tied-state triphones, with up to 3000 tied states, trained on the MediaParl database. State tying is based on the MDL criterion [16]. State emissions are modelled by 16 component GMMs. The phoneme set for French is composed of 37 phonemes.

For crosslingual adaptation experiments, a German ASR is also trained following the same setup and configuration described for the French ASR. The German phoneme set is composed of 56 phonemes.

C. Language Models

Language models are tri-gram models, trained on Parliament data transcriptions obtained from cantonal Parliaments throughout Switzerland, for French and German, respectively.

III. Theoretical Overview

This section provides a brief theoretical overview on state-of-the art adaptation methods. Readers familiar with the topic should directly jump to the next section.

In the easiest approach, only one adaptation transform is computed on all adaptation data. This is called a *global* transform. However, different models or even different mixtures of different states in a HMM/GMM system might benefit from different transformations to better fit to a given speaker. This means that multiple transforms should be created and a clustering of states or mixtures is needed to be specified in order to group parameters that share a transform. The number of clusters depends on available adaptation data: if more data is available for a given cluster, it can be split into further subclusters to allow for more specialized adaptation. In practice, this is usually done using a regression class tree [17], which allows for a flexible data-driven clustering. The regression class tree is constructed so as to group similar GMMs (that are close in acoustic space), this means that similar components will share a common transform. The tree can be grown until each leaf has sufficient adaptation data to estimate a transform. These transforms can be referred to as *tree-based* ones, which are expected to outperform global transforms.

A. MLLR Adaptation of means

Maximum Likelihood Linear Regression (MLLR) is an often used and easy linear transformation for HMM/GMM ASR models. The means (μ) of GMMs are transformed to fit better the speaker [6]:

$$\hat{\mu} = A\mu + b = W\xi. \tag{1}$$

This means that the transformed mean vector ($\hat{\mu}$) is obtained by applying the W transform to the extended original mean vector (ξ), obtained by extending the original n-dimensional mean vector $\mu = [\mu_1, \mu_2, ..., \mu_n]$ as:

$$\xi = [1, \mu_1, \mu_2, ..., \mu_n]^T. \tag{2}$$

This transform is linear and it can be easily decomposed as:

$$W = [b, A]. \tag{3}$$

The transform can be computed by maximum likelihood (ML) estimation for the adaptation data from the speaker.

B. Variance adaptation

Variances of the speaker independent models can also be adapted, this is usually done in a two stage approach. In the first stage, mean adaptation is carried out, secondly, in a separate step, variances are also updated using already the adapted means (as a so-called parent transform) for the computation. The covariance matrices (Σ) can be updated based on the following formula [6]:

$$\hat{\Sigma} = B^T H B, \tag{4}$$

where H is the transform to be estimated, and B is the inverse of the Choleski factor of Σ^{-1}.

C. Constrained MLLR (CMLLR)

Constrained MLLR is a special transform for adapting means and variances together, with a shared transform [7]. This means that this transform has fewer parameters than the two stage variance adaptation described so far and hence, less adaptation data can be sufficient for effective estimation of the transform. The main advantage of CMLLR is that it can be performed in the feature space, which allows for using it as a parent transform. The mean μ and the variance Σ are transformed with CMLLR as follows:

$$\hat{\mu} = A\mu + b, \tag{5}$$

$$\hat{\Sigma} = A\Sigma A^T. \tag{6}$$

Transform parameters to be computed involve A and b.

D. Maximum a posteriori adaptation (MAP)

The basic idea of MAP adaptation is to try to establish speaker dependent models based on prior speaker independent knowledge. In this way, adapted system performance is expected to be close to that of a speaker dependent system, but the amount of necessary training data is much lower than in case of training a pure speaker dependent system. The adaptation formula for a mean vector is as follows:

$$\hat{\mu}_{jm} = \frac{\tau\mu^0_{jm} + N_{jm}\mu_{jm}}{N_{jm} + \tau} \tag{7}$$

that is, to update the mean vector of mixture component m associated with state j, original mean μ_{jm} and the mean calculated on the adaptation data μ^0_{jm} are recombined where τ is a weighting parameter for the a-priori knowledge, usually set in empirical way.

Comparing MLLR-like and MAP transforms, MLLR transforms usually work better than MAP with little adaptation data, however, MLLR performance often saturates when more training data is available. On the other hand, MAP has better asymptotic properties, that is, MAP exhibits convergence to real speaker dependent systems' performance, but this convergence is slow. MLLR-like and MAP transforms can be used simultaneously [18] to benefit from the advantages of both approaches.

E. Structured Maximum a Posteriori Linear Regression (SMAPLR)

Structured Maximum a Posteriori Linear Regression can be regarded as a combination of global and tree based linear transforms, enabled using a similar principle to pure MAP adaptation. It allows a smooth transition between a high order transform and a single global transform. The constrained form of SMAPLR (CSMAPLR) is a combination of SMAPLR and CMLLR, and performs adaptation of variances too (whilst SMAPLR is usually used for the means). CSMAPLR is described in detail by [10].

IV. Evaluating Adaptation Techniques

A. Linear Regression Based Approaches

In this section, we briefly evaluate linear regression based transfoms on monolingual French ASR. MAP adaptation is not investigated as data available from speakers is sparse.

Adaptation is performed in a supervised, static manner in two steps: first a global transform is computed (base transform), then a regression tree is generated, based on which tree-based transforms are also computed (tree based transform). The regression tree used in the experiment had maximum 32 leaves (for speakers with insufficient adaptation data this can be less as trees are pooled to ensure enough adaptation data for each terminal node).

Results are presented in Table I. Results (word accuracy and WER improvement compared to baseline) for global and tree-based transforms are presented separately. In Table I, MEAN adaptation refers to the adaptation of means, MEAN+VAR refers to a two-step adaptation of means first and variances next, while MEAN&VAR refers to a joint adaptation of means and variances (as done in CMLLR). As SMAPLR and CSMAPLR are structured MAP transforms, they can be regarded as tree-based ones for comparison.

TABLE I. Linear regression based adaptation results for French MediaParl

Adapt method	Global/ tree	Adapted parameters	Acc [%]	Rel. WER red. [%]
Baseline	-	-	75.0	0.0
MLLR	global	MEAN	75.8	3.2
MLLR	global	MEAN+VAR	76.1	4.4
CMLLR	global	MEAN&VAR	76.3	5.2
MLLR	tree	MEAN	77.6	10.4
MLLR	tree	MEAN+VAR	77.8	11.2
CMLLR	tree	MEAN&VAR	78.0	12.0
SMAPLR	tree	MEAN	78.0	12.0
CSMAPLR	tree	MEAN&VAR	78.1	12.4

Results show that any type of the examined adaptation methods yielded improvement in recognition performance. The highest improvement (12.4% relative WER reduction) is seen with CSMAPLR, although the difference between CMLLR, SMAPLR and CSMAPLR is minor. As expected, regression tree based (or structural) adaptation approaches perform significantly better.

B. Speaker Adaptive Training

In Speaker Adaptive Training (SAT), speaker specific transformations are used during the training process. These speaker specific transforms are used for "de-individualizing" the utterances during the model (re-)training. In this sense, SAT can be interpreted as a special form of normalization. The underlying assumption is that human speech is a product of two components [11]: the first component is the raw speech representing purely phonetic variations, while the second one is speaker specific and represents speaker variation (caused by physiological differences, age, gender or even the different acoustic environments and so on). Interpreting the second component as a filter, the final speech product is the convolution of the two. Thus, by modelling speaker variations, it is possible to normalize them and to create a model set independent of

speaker characteristics (which is then also further adaptable to each speaker separately). In this sense, speaker independent models are independent not in the sense of being all-speakers models trained on a large number of speakers, but are ideally free of any speaker specific influence.

If sufficient training data is available from the speakers, which is the case for current experiments, SAT can be done using speaker transforms obtained by regression tree-based clustering. Otherwise, global adaptation data could be used.

The contribution of SAT to relative WER reduction is shown in Table II in parallel with baseline ASR performance (word accuracy) for French ASR. Applying further adaptation after SAT (using SAT transforms as parent transform for test speakers and do another adaptation) leads to some more modest improvement.

TABLE II. *Baseline, SAT-normalized and SAT+MLLR adapted word accuracies and relative WER reduction in French ASR.*

System	Language	Acc [%]	Rel. WER red. [%]
Baseline	French	75.0	-
+ SAT	French	78.8	15.2
+ MLLR	French	79.1	16.4

V. Crosslingual Adaptation

As briefly explained in the introduction, the idea behind crosslingual adaptation is to use it for recognition of non-native speech, supposing that a speaker specific adaptation computed on L1 data can improve recognition in L2 for the same speaker. This involves a mapping of phonemes (or even the individual GMMs of the tied-state model set) from L1 to L2. The easiest way of mapping is a rule-based direct mapping between phonemes of L1 to phonemes of L2, this approach is used in present paper: each i^{th} state and j^{th} mixture component of source phoneme in L1 is mapped to the same i^{th} state and j^{th} mixture component of the target phoneme in L2. Target phonemes for given source phonemes are determined based on phonetic similarity as follows: phonemes sharing their SAMPA symbol and used in both languages are mapped to each other. Phonemes specific to one language are mapped to the acoustic-phonetically closest phoneme in the target language. For the French and German phoneme sets used, 31 phonemes are common and 31 are specific to only one of the languages. When mapping French to German, 6 phonemes have to be mapped, whereas a mapping of German to French involves 25 mappings to be defined.

A more sophisticated approach is a data-driven mapping similar to the one used in [14], where a similarity measure is used to map GMMs in L1 to GMM clusters in L2, assuring also that each cluster has enough GMMs associated to compute adaptation transforms. This approach is in implementation phase, hence results are not yet available for the present paper.

In the crosslingual adaptation experiments, CMLLR transforms are used, which were found to operate close to the best performance but with low complexity. Adaptation transforms are generated in the source language (L1) against the target language (L2) ASR model set and used for recognition in the target language. We expect improvement for non-native speakers and some decrease in performance for native speakers.

As described in Section II, the test set consists of speakers speaking both the source and target languages. For adapting the ASR models, utterances in L1 are used, whereas recognition is done in L2. The decision whether a speaker is native or non-native in a given language is based on monolingual ASR performance for the given speaker as explained in [1]. If there is significant difference between monolingual ASR accuracy for a given speaker between German and French monolingual ASR performance, the speaker is regarded to be native in the better performing language and non-native in the other. If difference is slight (not significant), the speaker is regarded to be a real bilingual speaker, hence native in both languages. Based on these criteria, 3 speakers out of 7 in the test set were found to be native German and non-native French speakers, 3 speakers native in French and non-native in German and 1 real bilingual (native in both languages, with slight German preference). Results are presented in Table III, individually for each speaker, for the case where French speech recognition is evaluated with and without adaptation on German data. WER for non-native speakers improves by overall 3.8% relative (two speakers significantly, one not significantly). Scores of the bilingual speaker also improve. All native speakers suffered a performance decrease - as expected.

TABLE III. *Effect of German crosslingual adaptation on French ASR Word Error Rates (WER).*

Spkr ID	Native?	Baseline WER	Adapted WER
059	non-native	45.4	42.5
079	non-native	38.6	37.8
109	non-native	27.0	26.9
094	native	18.7	19.6
096	native	15.9	16.8
102	native	17.7	17.6
191	bilingual	40.9	37.2

Results for speech recognition with L1 French and L2 German are presented in Table IV. Non-native speakers benefit again from native adaptation data (one speaker had unfortunately insufficient data in German for reliable evaluation). The overall relative WER reduction seen for non-native speakers was over 4%. Scores of the bilingual speaker also improve slightly. There is no general improvement for native speakers (although one of them slightly improves).

The results confirm the initial working hypothesis, that recognition in non-native speech can be improved by adaptation in the native language, by using data from the same speaker. However, this improvement is inferior to the improvement seen when adaptation data is available in the same language the ASR is running (see Table I). Crosslingual adaptation improved scores of the bilingual speaker for both languages, however, with only one real bilingual speaker no further conclusions can be drawn.

VI. Conclusions

In this paper, intra- and crosslingual adaptation was investigated in a bilingual environment. Evaluation of state-of-the-art, linear regression based adaptation techniques and speaker adaptive training, or their combination improved recognition accuracy in a French ASR task, in accordance with available literature on using similar adaptation methods. Crosslingual adaptation was also investigated. Results show that there is

TABLE IV. *Effect of French crosslingual adaptation on German ASR Word Error Rates (WER).*

Spkr ID	Native?	Baseline WER	Adapted WER
094	non-native	32.1	30.7
096	non-native	24.1	22.7
102	non-native	41.4	40.0
059	native	19.1	20.1
079	native	26.0	26.3
109	native	21.5	22.4
191	biligual	31.3	30.9

an improvement after crosslingual adaptation for non-native speakers using their native language data for recognition in their non-native language, however, if speech is available in the non-native language from the same speaker, it is a better source for adaptation. Regarding future directions, modelling using MLP features could be better placed to analyse crosslingual adaptation capabilities because of MLP's higher flexibility. An interesting issue is us of bilingual acoustic models and use of native data to help non-native speech recognition.

ACKNOWLEDGMENTS

The authors would like to express their gratitude to the MediaParl project, funded by the Parliament Service of the State of Valais, Switzerland for their financial support and for providing access to the audio-video recordings.

REFERENCES

[1] D. Imseng, H. Bourlard, H. Caesar, P. N. Garner, G. Lecorvé, and A. Nanchen, "Mediaparl: Bilingual mixed language accented speech database," in *Proc. of the 2012 IEEE Workshop on Spoken Language Technology*, 2012, pp. 263–268.

[2] G. Szaszak, M. Cernak, P. N. Garner, P. Motlicek, A. Nanchen, and F. Tarsetti, "Automatic Speech Indexing System of Bilingual Video Parliament Interventions," Idiap Research Institute, Tech. Rep. Idiap-RR-25-2013, Jul. 2013.

[3] D. Kim, S. Umesh, M. Gales, T. Hain, and P. Woodland, "Using VTLN for broadcast news transcription," in *Proc. International Conference on Speech and Language Processing (ICSLP)*, 2004, pp. 1953–1956.

[4] S. Furui, "Unsupervised intralingual and cross-lingual speaker adaptation for HMM-based speech synthesis using two-pass decision tree construction," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 29, pp. 254–272, 1981.

[5] O. Viikki and K. Laurila, "Noise Robust HMM-Based Speech Recognition using Segmental Cepstral Feature Vector Normalization," *Speech Communication*, vol. 25, pp. 133–147, 1998.

[6] M. Gales and P. Woodland, "Mean and Variance Adaptation within the MLLR Framework," *Computer Speech and Language*, vol. 10, pp. 249–264, 2006.

[7] V. Digalakis, D. Rtischev, and L. Neumeyer, "Speaker Adaptation Using Constrained Estimation of Gaussian Mixtures," *IEEE Transactions on Speech and Audio Processing*, vol. 3, pp. 357–366, 1995.

[8] J. L. Gauvain and C. Lee, "Maximum a Posteriori Estimation for Multivariate Gaussian Observations of Markov Chains," *IEEE Trans. SAP*, vol. 2, no. 2, pp. 291–298, 1994.

[9] O. Siohan, T. A. Myrvoll, and C.-H. Lee, "Structural maximum a posteriori linear regression for fast HMM adaptation," *Computer Speech and Language*, vol. 16, pp. 5–24, 2002.

[10] Y. Nakano, M. Tachibana, J. Yamagishi, and T. Kobayashi, "Constrained structural maximum a posteriori linear regression for average-voice-based speech synthesis," in *Proceedings of Interspeech 2006*, 2006, pp. 2286–2289.

[11] T. Anastasakos, J. McDonough, and J. Makhoul, "Speaker adaptive training: a maximum likelihood approach to speaker normalization," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 2, 1997, pp. 1043–1046.

[12] H. Liang, J. Dines, and L. Saheer, "A Comparison of Supervised and Unsupervised Cross-Lingual Speaker Adaptation Approaches for HMM-Based Speech Synthesis," Dallas, TX, USA, March 2010, pp. 4598–4601.

[13] M. Gibson and W. Byrne, "Unsupervised intralingual and cross-lingual speaker adaptation for HMM-based speech synthesis using two-pass decision tree construction," *IEEE Trans. ASLP*, vol. 19, no. 4, pp. 895–904, 2011.

[14] H. Liang and J. Dines, "An Analysis of Language Mismatch in HMM State Mapping-Based Cross-Lingual Speaker Adaptation," Makuhari, Japan, September 2010.

[15] R. Karhila, "Cross-lingual acoustic model adaptation for speaker-independent speech recognition," Ph.D. dissertation, Aalto University, Helsinki, 2010.

[16] K. Shinoda and T. Watanabe, "Acoustic modeling based on the MDL criterion for speech recognition," in *Proc. EuroSpeech97*, vol. 1, 1997, p. 99102.

[17] M. J. F. Gales, "The generation and use of regression class trees for MLLR adaptation," 1996, technical Report CUED/F-INFENG/TR263, Cambridge University.

[18] S. Goronzy and R. Kompe, "A combined MAP + MLLR approach for speaker adaptation," in *Proceedings of the the Sony Research Forum '99*, vol. 1, 1999, pp. 9–14.