Development of Bilingual ASR System for MediaParl Corpus

Petr Motlicek[1]*, David Imseng*[1]*, Milos Cernak*[1]*, Namhoon Kim*[2]

[1]Idiap Research Institute, Martigny, Switzerland
[2]Samsung Electronics Co. Ltd, Suwon, South Korea
{motlicek,dimseng,mcernak}@idiap.ch, namhoon.kim@samsung.com

Abstract

The development of an Automatic Speech Recognition (ASR) system for the bilingual MediaParl corpus is challenging for several reasons: (1) reverberant recordings, (2) accented speech, and (3) no prior information about the language. In that context, we employ frequency domain linear prediction-based (FDLP) features to reduce the effect of reverberation, exploit bilingual deep neural networks applied in Tandem and hybrid acoustic modeling approaches to significantly improve ASR for accented speech and develop a fully bilingual ASR system using entropy-based decoding-graph selection. Our experiments indicate that the proposed bilingual ASR system performs similar to a language-specific ASR system if approximately five seconds of speech are available.

Index Terms: multilingual automatic speech recognition, language identification, non-native speech

1. Introduction

Valais is a bilingual Swiss canton and its parliament debates in French and German. About two third of its members are French natives and one third are German natives. Some of the parliament members are that fluent in both languages that they switch languages, sometimes without even noticing. We refer to this kind of situations as code-switched speech [1].

Performing Automatic Speech Recognition (ASR) on the data from the Valais parliament is particularly challenging because it contains code-switched, accented and non-native speech. The debates take place in a large chamber (reverberate environment) and are recorded with a single distant microphone. To enable research on these challenging recordings, the MediaParl database [2] was released and is publicly available.

In literature, reverberation has been addressed through the application of some filtering or pre-processing (e.g., modulation filtering [3]), or the employment of a robust front-end more resistant to the reverberation thanks to temporal characteristics in critically-warped frequency sub-bands over relatively long windows (e.g., TRAP [4], RASTA post-processing [5] or FDLP [6]). In this paper, we employ Frequency Domain Linear Prediction (FDLP) feature extraction that has already been shown to improve recognition of reverberate speech [7].

The language mix in Valais leads to obvious difficulties with many people working in a non-native language and to a high variability in the speech recordings. The high variability of these accented recordings can be addressed by bilingual acoustic modeling techniques on different levels:
(a) Acoustic model: Niesler [8] for example studied sharing resources inspired by multilingual acoustic modeling techniques based on standard HMM/GMMs proposed by Schultz [9]. However, only marginal ASR performance gains were reported. More recently, Deep Neural Nets (DNNs) applied to multilingual modeling outperformed HMM/GMM based approaches [10, 11].
(b) Features: previous studies [12, 13, 14, 15] found that the relation between phonemes of different languages can be learned and exploited in multilingual acoustic model training. Posterior-based features, estimated by neural networks, are particularly well suited for such tasks.

In this paper, we first compare bilingual Bottleneck (BN) features estimated by a bilingual DNN (later denoted to as BN-HMM/GMM) and state-of-the-art HMM/DNN (hybrid) acoustic modeling also trained in a multilingual fashion. We show that bilingual BN-HMM/GMM as well as HMM/DNN outperform the corresponding monolingual approaches.

Then, we propose a fully bilingual ASR system performing entropy-based decoding graph-selection. In bilingual systems (i.e., no prior information about the language) some way of identifying input language is needful. This can be done either explicitly through a real Language Identification (LID) module before or after the monolingual decoding [2, 16] , or implicitly by running one bilingual ASR system [17]. In our approach, the language decision is made during decoding based on frame-level entropy information criteria estimated from word posterior probabilities.

The paper is organized as follows: Section 2 describes the MediaParl database. Monolingual experiments on accented speech are presented in Section 3 and the bilingual systems are compared in Section 4. Finally, conclusions are drawn in Section 5.

2. Database

Bilingual cantons of Switzerland case a high variability of regional (within language) and foreign (across language) accents. In this study, we focus on the two most spoken languages – French (FR) and German (GE) – and use the MediaParl database to evaluate bilingual, accented and non-native speech recognition [2].

The MediaParl corpus contains political debates of the parliament of Valais, a bilingual canton of Switzerland. The data was recorded with a single distant microphone in a reverberate environment (i.e., large chamber where political debates take place). This database therefore provides three major challenges for ASR research: bilingualism, accented speech and reverberation. The statistics of the database are given in Table 1.

2.1. Dictionaries

The phonemes of the dictionaries are represented using the Speech Assessment Methods Phonetic Alphabet (SAMPA) [18] that supports multiple languages including French and German. For the French dictionary, we used BDLex [19] that uses 38

Table 1: *Statistics of the MediaParl dataset: number of words in the dictionary, the perplexity (PPL) of LM on the test set as well as amounts of training (TRN) and test (TST) data are shown for each language individually and for the bilingual setup.*

Language	**Dict.**	**LM**	**Data (h)**	
	# words	TST PPL	TRN	TST
French	11k	165	19.2	1.5
German	16k	213	17.8	2.1
Bilingual	27k	323	37.0	3.6

phonemes (including "sil"). The German dictionary is based on PhonoLex [20] using 55 phonemes (including "sil"). To account for the Swiss German peculiarities, we added three phonemic affricates, [p͡f], [ts], [tʃ] (considered to be native to German) and two nasal vowels [ã], [õ] (due to many French-specific words appearing in German transcriptions).

To compensate for many unseen words (abbreviations, names in both languages), we trained a grapheme-to-phoneme tool[1], namely Phonetisaurus [21], from existing dictionaries to derive finite state transducer based mapping of sequences of letters (graphemes) to their acoustic representation (phonemes).

For the bilingual dictionary, we simply merged the dictionaries of both languages (informative tests with tagged words did not improve recognition). The bilingual dictionary therefore employs a shared phoneme set, where phonemes that share the identical SAMPA symbol are merged.

2.2. Data partitioning

The bilingual speakers from the MediaParl dataset are equally represented in the training and the test sets. The training set contains 106 speakers and 75 speakers for French and German, respectively. The test set contains 12 speakers and 7 speakers for French and German, respectively. All the recordings are pre-segmented into utterances of about 10 seconds on average. Since the MediaParl database contains bilingual speakers, there is some code-switched speech, i.e., speakers switch between French and German. However, in most of the cases, the language switch happens at sentence boundaries and we can presume that there is no language-switch within the short utterances.

2.3. Language models

Given that the short utterance usually do not contain language switches, we have not investigated an impact of specific language modelling for code-switching speech, such as [22]. Rather, the conventional trigram Language Models (LMs) for French and German were built on the transcripts from the training data and text from the Swissparl corpus that contains Swiss Parliament proceedings[2]. The bilingual LM was built by simply interpolating the (equally weighted) individual LMs.

3. Monolingual ASR

To develop a bilingual ASR, we first build monolingual baselines for French and German. We found that a conventional HMM/GMM system that uses Mel-Frequency Cepstral Coefficients (MFCC) in combination with FDLP features yields significant improvement compared to a system that uses MFCC features only (i.e., by about 4% relative in word error rates as shown in Table 2). Therefore, we employ 78-dimensional MFCC and FDLP features (i.e., including Δ and $\Delta\Delta$) for all our experiments, done with the kaldi toolkit [23].

3.1. Monolingual acoustic models

We compare three different acoustic modelling techniques:
(a) HMM/GMM: conventional 3-state left-to-right context-dependent HMM/GMMs. Decision tree based clustering returns 3905 and 4032 tied states for French and German, respectively. In total, 50k Gaussians were used.
(b) BN-HMM/GMM: current state-of-the-art front-end using posterior-based features. The features are usually phone class posterior probabilities given the acoustics, and estimated with a DNN trained on large amount of data [24, 25]. The phone classes are often context-dependent triphones. In this setup, a language-specific 6-layer Bottleneck (BN) DNN is trained with following number of nodes in each layer: 702, 1000, 1000, 30, 1000, K, where K is given by the number of tied states in each language-specific HMM/GMM baseline. As an input, 9 consecutive MFCC+FDLP features and their derivatives are fed to the DNN (702-dimensional vector). The randomly initialized, fully connected language-specific DNNs are trained using the cross-entropy criterion. To prevent over-fitting, 10% of the training set is used for cross-validation. All activations of the nodes in the last layer are transformed using the softmax, whereas the sigmoid transfer function is applied in all other layers (except the BN layer that is linear). The linear output of the BN layer serves as BN features. We append derivatives and perform per-speaker normalization before using them for subsequent HMM/GMM training.
(c) HMM/DNN: (hybrid) HMM/DNN systems have been extensively studied and are nowadays considered as state-of-the-art speech recognizers. For our experiments, we employ RBM pre-training [26]. The DNN has 3-hidden layers (with the following number of nodes: 702, 2000, 2000, 2000, K), where K is the number of tied-states in each language.

Speech recognition results in terms of word error rates (WERs) for French and German test sets are given in Table 2. HMM/DNN and BN-HMM/GMM perform similar and significantly better than the HMM/GMM baseline.

3.2. Bilingual acoustic models

For building a bilingual acoustic model, training data from both datasets (French and German) and a shared phoneme set (62 phonemes) were used. During decoding, we used language specific dictionaries and LMs and evaluated the bilingual acoustic models on each language (i.e., independently on each language-specific test set).

ASR results in terms of WERs are shown in Table 2 for the following acoustic models:
(a) HMM/GMM: the bilingual baseline is an HMM/GMM system with 4900 tied states and 75k Gaussians.
(b) BN-HMM/GMM: the bilingual 6-layer BN DNN is trained on spliced MFCC+FDLP features with the following number of nodes in each layer: 702, 2000, 2000, 30, 2000, K, (i.e., twice the width of the monolingual NN). K is given by the number of tied states in the bilingual HMM/GMM baseline (i.e., 4900). As for the monolingual acoustic models, the extracted BN features are enriched by derivatives and per-speaker normalization

[1] Available at `https://github.com/idiap/iss-dicts`.

[2] Internal Idiap text database that consists of publicly available parliament transcriptions throughout Switzerland (Basel, Bern, Fribourg, Vaud, Geneva, Neuchâtel Solothurn and Zürich).

Table 2: *Word error rates (WER) for different monolingual and bilingual acoustic models (AM), exploiting MFCC+FDLP features. Decoding is always performed with the language-specific (FR or GE) LM. We also present the baseline system built on MFCC only features.*

System	Monoling. AM		Biling. AM	
	FR	GE	FR	GE
HMM/GMM	24.2 %	21.5 %	25.2 %	22.0 %
– MFCC only	25.1 %	22.3 %	26.1 %	22.6 %
BN-HMM/GMM	22.4 %	16.9 %	22.0 %	16.7 %
HMM/DNN	21.6 %	17.1 %	**21.4** %	**16.6** %

is applied before HMM/GMM training.
(c) HMM/DNN: we use the same DNN architecture as for the monolingual case ($K = 3861$) and also apply RBM pre-training. Since decoding is performed with monolingual dictionaries and LMs, we adapt the DNN to the target language by randomly reinitializing the last layer and re-training the whole DNN using the data from target language only, as was done, for example, in [27].

Table 2 shows that bilingual acoustic modeling performs worse in case of the HMM/GMM system. For the DNN based systems on the other hand, the bilingual acoustic modeling yields improvement. Hence, we conclude that the DNN based acoustic modeling techniques, BN-HMM/GMM and HMM/DNN, are indeed able to exploit the bilingual data during acoustic model training and yield improvement over the monolingual acoustic models.

4. Bilingual ASR

The main objective of this paper is to build a fully bilingual ASR engine, i.e., without making use of a priori information about the language during recognition. To do so, we investigated three different approaches to perform bilingual ASR:
(a) Bilingual ASR baseline: bilingual acoustic models (i.e., HMM/GMM, BN-HMM/GMM and HMM/DNN) are employed with a bilingual LM during decoding. This system implicitly performs Language Identification (LID) through ASR and is also able to recognize a mixture of French and German words, i.e. code-switched speech.
(b) *LID* switch: This approach first performs LID to decide about the language. Then, the corresponding language-specific decoder is used during recognition. More specifically, we perform LID with the recently proposed hierarchical neural network based language identification [28] that makes use of DNNs. The first one estimates phone posteriors. The output of the first DNN is then used as input for the second DNN, which is trained to estimate language posteriors based on a longer temporal context. For this work, we use the DNN of the bilingual HMM/DNN system to estimate phone posteriors and then estimate language posteriors with an DNN that considers 30 frames of temporal context, sampled at 5 frames (i.e. the frames -15, -10, -5, 0, 5, 10,15). To decide the language, we simply average the frame-based language posteriors over the whole utterance.
(c) Parallel decoding (*ENT*): the decoding is commenced using combined (in parallel) language-specific graphs, and language-specific word recognition lattices are generated. We then measure an amount of uncertainty by using frame-based entropy information criteria computed from word posterior probabilities estimated from the lattices, similar to [29]. Frame-based entropies are eventually summed over time and the decision is taken (based on minimum entropy) to select the appropriate decoding graph.

Table 3: *WERs for bilingual ASR. Recognition rates are given for the complete bilingual test set (ALL) and also split into test sets of the individual languages.* LID *stands for the system with the LID switch and* ENT *for the system that employs parallel decoding.*

System	Bilingual ASR		
	FR	GE	ALL
HMM/GMM	28.9 %	26.1 %	27.4 %
BN-HMM/GMM	23.7 %	19.4 %	21.5 %
+LID	22.0 %	20.3 %	21.1 %
+ENT	22.1 %	16.8 %	**19.3** %
HMM/DNN	24.1 %	20.1 %	22.0 %
+LID	21.4 %	20.9 %	21.1 %
+ENT	21.4 %	16.6 %	**18.9** %

Performance of the three investigated approaches is shown in Table 3. We can observe significant degradation in performance for the fully bilingual baseline (compared to Table 2). This is in line with previous studies [17] and can be attributed to the bilingual LM that induces inter-language confusion to the decoding. The parallel decoding system *ENT* performs best, but requires to run two decoders in parallel.

To reduce the computational load, it is feasible to start decoding using combined language-specific graphs and to decide the language during decoding. Figure 1 shows the performance of such the bilingual ASR with respect to the time interval used to collect frame-based word entropy estimates from the beginning of each decoded speech segment. It can be seen that by considering the first 5 seconds of an utterance, the language decision applied on the bilingual ASR yields performance of the monolingual ASR. Further, language decisions based on entropy estimates collected from the first second of the decoded output yield performance of fully bilingual ASR.

5. Conclusions

We developed a fully bilingual ASR system for French and German using the MediaParl corpus and successfully addressed its three major challenges: (1) reverberation, (2) accented speech and (3) bilingualism. We found that employing MFCC features in combination with FDLP features yields improvement (about 4% relative in WERs) compared to MFCC features only. The accented speech challenge and bilingualism are tackled by exploiting bilingual acoustic modeling techniques based on bilingually trained deep neural networks. The HMM/DNN system yields 15% and 25% relative improvements in terms of WERs, for French and German respectively, when compared to the bilingual HMM/GMMs.

Eventually, we presented a fully bilingual ASR system that makes use of confidence scores estimated from word recognition lattices over relatively short-period of time. With only 5 seconds of speech, the proposed system performs equally well as the best monolingual systems without knowing the language prior to decoding.

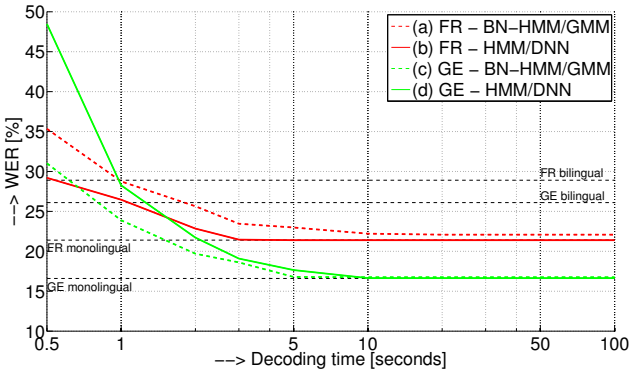

Figure 1: *Dependence of WER (for both French and German test sets) vs. time interval used to collect word entropy-based confidence scores from the beginning of each segment in the bilingual ASR (for BN-HMM/GMM and HMM/DNN acoustic models). Dashed horizontal lines determine WER baselines of bilingual HMM/GMMs when composed with either the bilingual decoder (i.e., bilingual ASR employing bilingual LM), or monolingual decoders.*

6. Acknowledgements

This work was supported by Eurostars Programme powered by Eureka and the European Community under the project "D-Box: A generic dialog box for multi-lingual conversational applications". P. Motlicek was also supported by Samsung Electronics Co. Ltd, South Korea, under the project "Multi-Lingual and Cross-Lingual Adaptation for Automatic Speech Recognition". We are grateful to the Parliament Service of the State of Valais for providing access to the parliament debate A/V recordings.

7. References

[1] P. Auer, *Code-Switching in Conversation: Language, Interaction and Identity*, P. Auer, Ed. Routledge, Aug. 1999.

[2] 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 IEEE Workshop on Spoken Language Technology (SLT)*, 2012, pp. 263–268.

[3] A. Kusumoto, T. Arai, K. Kinoshita, H. N., and N. Vaughan, "Modulation enhancement of speech by a pre-processing algorithm for improving intelligibility in reverberant environments," *Speech Communication*, vol. 45, pp. 101–113, 2005.

[4] H. Hermansky, "Trap-tandem: Data-driven extraction of temporal features from speech," in *Proc. of ASRU*, 2003, pp. 255–260.

[5] B. Kingsbury and N. Morgan, "Recognizing reverberant speech with rasta-plp," in *Proc. of ICASSP*, vol. 2, 1997, pp. 1259–1262.

[6] M. Athineos and D. Ellis, "Frequency-domain linear prediction for temporal features," in *Proc. of ASRU*, 2003, pp. 261–266.

[7] S. Ganapathy, S. Thomas, P. Motlicek, and H. H., "Applications of signal analysis using autoregressive models for amplitude modulation," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2009, pp. 341–344.

[8] T. Niesler, "Language-dependent state clustering for multilingual acoustic modelling," *Speech Communication*, vol. 49, pp. 453–463, 2007.

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

[10] G. Heigold, V. Vanhoucke, A. Senior, P. Nguyen, M. Ranzato, M. Devin, and J. Dean, "Multilingual acoustic models using distributed deep neural networks," in *Proc. of ICASSP*, 2013, pp. 8619–8623.

[11] J. Huang, J. Li, D. Yu, L. Deng, and Y. Gong, "Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers," in *Proc. of ICASSP*, 2013, pp. 7304–7308.

[12] D. Imseng, H. Bourlard, and P. N. Garner, "Using KL-divergence and multilingual information to improve ASR for under-resourced languages," in *Proc. of ICASSP*, 2012, pp. 4869–4872.

[13] L. Tòth, J. Frankel, G. Gosztolya, and S. King, "Cross-lingual portability of MLP-based tandem features - a case study for English and Hungarian." in *Proc. of Interspeech*, 2008, pp. 2695–2698.

[14] F. Grézl, M. Karafiát, and M. Janda, "Study of probabilistic and bottle-neck features in multilingual environment," in *Proc. of ASRU*, 2011, pp. 359–364.

[15] V. Do, X. Xiao, E. Chng, and H. Li, "Context-dependent phone mapping for lvcsr of under-resourced languages," in *Proc. of Interspeech*, 2013, pp. 500–504.

[16] C. Shia, Y. Chiu, J. Hsieh, and C. Wu, "Language boundary detection and identification of mixed-language speech based on map estimation," in *Proc. of ICASSP*, 2004, pp. I–381.

[17] Z. Wang, U. Topkara, T. Schultz, and A. Waibel, "Towards universal speech recognition," in *Proceedings of the 4th IEEE International Conference on Multimodal Interfaces*, ser. ICMI '02, 2002, pp. 247–252.

[18] J. Wells, "SAMPA computer readable phonetic alphabet," http://www.phon.ucl.ac.uk/home/sampa/, Feb 2013.

[19] G. Perennou, "B.D.L.E.X. : A data and cognition base of spoken French," in *Proc. of ICASSP*, vol. 11, 1986, pp. 325–328.

[20] F. Schiel, "Aussprache-lexikon PHONOLEX," http://www.phonetik.uni-muenchen.de/forschung/Bas/BasPHONOLEXeng.html, March 2013.

[21] J. Novak, N. Minematsu, K. Hirose, C. Hori, H. Kashioka, and P. Dixon, "Improving WFST-based G2P conversion with alignment constraints and RNNLM N-best rescoring," in *Proc. of Interspeech*, 2012.

[22] H. Adel, N. T. Vu, and T. Schultz, "Combination of Recurrent Neural Networks and Factored Language Models for Code-Switching Language Modeling," in *ACL (2)*. The Association for Computer Linguistics, 2013, pp. 206–211.

[23] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, and P. Schwarz, "The kaldi speech recognition toolkit," in *Proc. of ASRU*, 2011.

[24] F. Grezl, M. Karafiat, and L. Burget, "Investigation into bottleneck features for meeting speech recognition," in *Proc. of Interspeech*, no. 9, 2009, pp. 2947–2950.

[25] D. Yu and M. L. Seltzer, "Improved bottleneck features using pretrained deep neural networks," in *Proc. of Interspeech*, 2011, pp. 237–240.

[26] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle, "Greedy layer-wise training of deep networks," *Advances in neural information processing systems*, vol. 19, p. 153, 2007.

[27] D. Imseng, P. Motlicek, P. Garner, and H. Bourlard, "Impact of deep mlp architecture on different acoustic modeling techniques for under-resourced speech recognition," in *Proc. of ASRU*, 2013.

[28] D. Imseng, M. Magimai.-Doss, and H. Bourlard, "Hierarchical Multilayer Perceptron based language identification," in *Proceedings of Interspeech*, 2010, pp. 2722–2725.

[29] P. Motlicek, "Automatic out-of-language detection based on confidence measures derived from vcsr word and phone lattices," in *Proc. of Interspeech*, 2009.