EXPLOITING UN-TRANSCRIBED FOREIGN DATA FOR SPEECH RECOGNITION IN WELL-RESOURCED LANGUAGES

David Imseng, Blaise Potard, Petr Motlicek, Alexandre Nanchen, Hervé Bourlard

Idiap Research Institute, Martigny, Switzerland

{dimseng,bpotard,motlicek,ananchen,bourlard}@idiap.ch

ABSTRACT

Manual transcription of audio databases for automatic speech recognition (ASR) training is a costly and time-consuming process. State-of-the-art hybrid ASR systems that are based on deep neural networks (DNN) can exploit un-transcribed foreign data during unsupervised DNN pre-training or semi-supervised DNN training. We investigate the relevance of foreign data characteristics, in particular domain and language.

Using three different datasets of the MediaParl and Ester databases, our experiments suggest that domain and language are equally important. Foreign data recorded under matched conditions (language and domain) yields the most improvement. The resulting ASR system yields about 5% relative improvement compared to the baseline system only trained on transcribed data. Our studies also reveal that the amount of foreign data used for semi-supervised training can be significantly reduced without degrading the ASR performance if confidence measure based data selection is employed.

Index Terms— Semi-supervised learning, deep neural networks, confidence measures, speech recognition

1. INTRODUCTION

Automatic speech recognition (ASR) systems are based on statistical parametric methodologies and therefore require large amounts of data during training. Furthermore, the training data needs to be transcribed. Obtaining data transcripts is often expensive and usually involves human interaction.

However, for many languages in the world, there are only very small amounts of transcribed data available. Therefore, many studies addressed the exploitation of foreign out-of-domain and out-of-language data for the training of ASR systems [1–3]. It was shown that foreign data usually helps in low-resourced scenarios. On the other hand, in general, there is little (or no) performance gain if a large amount of target data is available [3].

In this paper, we investigate whether foreign data (out-of-domain or out-of-language) can improve the performance of an ASR system that has already been trained on 20 hours of matched data. In the context of hybrid HMM/DNN systems, where the emission probabilities of the hidden Markov model (HMM) states are estimated with deep neural networks (DNNs), we compare the benefits obtained with different kinds of speech data.

We study the performance of an ASR system built on 20 hours of French MediaParl data [4]. MediaParl contains German and French data recorded at the bilingual Valais parliament (Valais is a bilingual Swiss canton). In contrast to other studies, such as [5], the MediaParl database provides the unique opportunity of investigating how beneficial out-of-language data is when recorded in matched conditions. Since some speakers are bilingual, MediaParl also contains data recorded from the same speaker, the same domain and the same channel in French and German.

To investigate how important domain and language are, we use three un-transcribed foreign datasets, 50 hours of French MediaParl data (matched data), 50 hours of German MediaParl data (matched domain, unmatched language) and 50 hours of Ester data (unmatched domain, matched language). We use each dataset for DNN pre-training [6], as well as semi-supervised DNN training [7, 8], with and without data selection based on confidence measures. Our studies show that generative pre-training is language and domain independent, hereby confirming earlier studies [8, 9]. For semi-supervised training, experiments reveal that (1) automatically generated transcripts are beneficial, (2) language and domain seem to be equally important, and (3) confidence measure based data selection allows to significantly reduce the amount of foreign data without degrading the ASR performance.

The reminder of the paper is structured as follows: Section 2 gives more details about the used datasets and Section 3 briefly reviews the applied acoustic modeling techniques. Experimental setup and results are then presented in Sections 4 and 5, respectively.

2. DATABASES

We used data from two different databases: MediaParl [4] and Ester [10]. With regard to our target, MediaParl provides matched data as well as out-of-language data recorded under

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".

the same conditions (matched channel/domain) whereas Ester provides out-of-domain French data.

2.1. MediaParl

MediaParl [4] is a publicly available bilingual database recorded at the Valaisan Parliament. Valais is a bilingual French/German Swiss canton and the parliament members can therefore express themselves in either language.

The MediaParl speech corpus contains debates of the years 2006 and 2009-2012[1]. The parliament debates always take place in the same large, closed room. Each speaker intervention can last from about 10 seconds up to 15 minutes and the voice is recorded through a distant microphone, and played back simultaneously through loudspeakers.

MediaParl data is challenging for ASR because there is a large amount of background noise as well as reverberation - due to the room architecture and speech playback. Multiple speakers with a large variety of local accents, some of them non-native speakers, further increase the complexity of the database.

We refer to the publicly available and manually transcribed database as MediaParl-core *(MP-core)*. Since the debates at the parliament are always recorded, it is easy to obtain un-transcribed data (not part of the distributed database). We refer to this un-transcribed data as extended MediaParl (*MP-ext*). The MP-ext data is automatically split into French and German parts by using a conventional automatic language identification (LID) system trained on MP-core. The accuracy of the LID system is approximately 98% [4].

2.1.1. MP-FR-core and MP-FR-ext

The majority – about two thirds – of the MediaParl database consist of standard Swiss French speech. The French part of MP-core is referred to as *MP-FR-core* and is split in training (19.2 h) and testing (1.5 h) data. MP-FR-core is considered as our target database, i.e., the one on which we would like to improve ASR performances.

The French part of MP-ext consists of more than 200 h of un-transcribed data. For the sake of comparison, about 50 h of speech were retained in the present study. This subset is referred to as *MP-FR-ext*. MP-FR-ext and MP-FR-core have matched conditions for both the language and the channel/domain.

2.1.2. MP-GE-core and MP-GE-ext

The remaining third of the MediaParl database consists of Swiss German speech. Swiss German in Valais encompasses a large variety of accents and local dialects. However, in the parliament, people almost exclusively speak in (accented) standard Swiss German.

The German part of MP-core is referred to as *MP-GE-core* and is split in training (17.8 h) and testing (2.1 h) data.

The German part of MP-ext consists of about 100 h of un-transcribed speech. For comparison purposes, about 50 h of speech were used to form the *MP-GE-ext* dataset. MP-GE-ext was recorded in the same conditions as our target MP-FR-core; the language is different, but channel and domain match the target database.

[1]For the recordings of 2009-2012, video streams are also available online: `http://www.canal9.ch/television-valaisanne/emissions/grand-conseil.html`.

2.2. Ester

Ester [10] is a database of standard French radio broadcast news. It comprises a large number of speakers in various recording conditions. In this study, we retained a subset of Ester consisting of native speakers in low noise conditions, it is 58.3 h long. Although there are some minor differences between standard French and Swiss French, we will consider them to be the same language. Channel and domain of Ester and MP-FR-core however, are very different.

3. ACOUSTIC MODELING

In this section, we describe the acoustic modeling approaches under investigation. Our basic system is a hybrid HMM/DNN [11], where the emission probabilities of the HMM states are estimated with a DNN. Details about the experimental setup are described later in Section 4. Here, we briefly review the applied techniques.

3.1. Unsupervised generative pre-training (RBM)

Hybrid HMM/DNN systems have been extensively studied over the last couple of years and are today's state-of-the-art speech recognizers. DNNs have several more layers than conventional Multilayer Perceptrons (MLPs) and therefore many more parameters. Since neural networks are usually trained with error back-propagation algorithms that may converge to local minima, the parameters are often initialized using pre-training algorithms [6].

We use a generative pre-training approach and train the network layer by layer. Each pair of layers is treated as a restricted Boltzmann machine (RBM) [6]. The first RBM uses Gaussian-Bernoulli units and the following RBMs then have Bernoulli-Bernoulli units. This pre-training approach is completely unsupervised and does not require transcriptions.

It was already shown that pre-training is language independent [8, 9]. However, it seems to be still unclear what makes some data suitable for unsupervised pre-training [9]. By using the datasets described in Section 2, we hope to better understand how to select suitable data for pre-training.

Amount of used data	10 %	20 %	30 %	40 %	50 %	60 %	70 %	80 %	90 %	100 %
Word error rate (WER)	15.0 %	14.9 %	14.9 %	14.6 %	14.6 %	14.5 %	14.3 %	14.4 %	14.4 %	14.2 %

Table 1. Word error rates on the test set of MP-FR-core for different amounts of automatically transcribed foreign MP-FR-ext data. We employed the *SEG* method and used only segments that have a confidence score above a certain threshold.

3.2. Semi-supervised pre-training

An alternative to unsupervised pre-training is semi-supervised pre-training that makes use of automatically generated transcripts. It consists of three steps: (1) transcript generation, (2) foreign DNN training, and (3) target DNN training.

3.2.1. Transcript generation

To exploit un-transcribed data, we first build a basic ASR system trained using manually transcribed MP-core data, which can then be used to generate transcripts for the foreign un-transcribed data.

3.2.2. Foreign DNN training

The automatically generated transcripts can then be used to train a DNN. However, automatically generated transcripts may contain errors and erroneous transcripts should ideally be excluded during DNN training. Inspired by a recent study [8], we apply confidence measures to pre-select confident transcripts. As a confidence, the amount of uncertainty measured by using frame-based entropy information criteria is used, similar to [12]. The confidence is computed from word posterior probabilities estimated from the decoding lattices using the forward-backward procedure. We evaluate three methods:

- The *straightforward* method (NONE) ignores confidence measure and uses all available transcripts.
- The *segment based* method (SEG) sums and normalizes the frame-based entropy scores on a per segment basis. Whole segments that have a confidence score below a certain threshold will be excluded from training.
- The *frame based* method (FRM) directly exploits frame-based confidence scores to include or exclude single frames, based on a given threshold.

3.2.3. Target DNN training

The foreign data may be from another language, leading to mismatched DNN outputs (different tied-state targets). We therefore adapt the DNN to the target data by randomly re-initializing the last layer of the DNN and re-training the whole DNN using the target data, as was done, for example, in [13].

4. EXPERIMENTAL SETUP

We used the Kaldi ASR toolkit [14] for our experiments. An overview over all evaluated systems is given in Table 2.

4.1. Baseline

Our baseline ASR was trained using MP-FR-core. As an acoustic model, we trained a DNN with 3 hidden layers. Each hidden layer consisted of 2,000 hidden units. At the input, we used nine consecutive frames (four preceding and four following frames) of 39-dimensional Mel-Frequency Cepstral Coefficients (MFCC) including deltas and double deltas. We trained the DNN to estimate posterior probabilities of 3,905 tied-state targets.

For the decoding, a trigram ARPA language model was trained from three different sources: transcripts from MP-FR-core, text from the Swissparl corpus containing Swiss Parliament proceedings, and text from Europarl – a multilingual corpus of European Parliament proceedings [15]. Europarl contains about 50 million words for each language and is used to overcome data sparsity of the MediaParl and Swissparl texts.

The standard HMM/DNN system yielded 16.2% word error rate (WER). After applying RBM based pre-training, using the MP-FR-core data, we obtained a WER of 15% (also shown in Table 2).

4.2. Confidence thresholds

As discussed in Section 3.2, we select confident audio data based on confidence measures. To ensure a fair comparison between the three foreign datasets and the two confidence measure methods, we set the thresholds such that a similar amount of foreign data was used in each case.

The amount of data to select was determined by studying how different amounts of high confidence MP-FR-ext data affect the WER on the MP-FR-core test set if the SEG method is used; the results are shown in Table 1. It seems that there is a noticeable drop in WER when 40% of the foreign data (about 20 h) is used. For the sake of comparison, we set all the confidence thresholds to retain the most confident 40% of the foreign datasets (for both SEG and FRM methods).

4.3. Systems

For each foreign dataset, we evaluated an unsupervised *RBM* system and semi-supervised *NONE*, *SEG* and *FRM* systems as described in Section 3.

	Amount of data	Lang.	Domain	Unsupervised RBM	Semi-supervised NONE (100%)	SEG (40%)	FRM (40%)
MP-FR-core	19.2 h	FR	PARL	15.0 %	Baseline (N/A)		
MP-FR-ext	51.2 h	FR	PARL	15.0 %	**14.2** %	14.6 %	14.6 %
Ester	58.1 h	FR	BN	14.8 %	14.6 %	14.9 %	14.6 %
MP-GE-ext	52.1 h	GE	PARL	15.0 %	14.7 %	15.0 %	14.7 %
MP-FR-ext & Ester	109.3 h	FR	MIXED	14.7 %	**14.2** %	14.6 %	14.3 %

Table 2. Word error rates (WERs) on the test set of MP-FR-core. Baseline without pre-training is 16.2% WER. Thresholds for confidence measure were set to retain 40% of the data. PARL stands for Parliament data, BN for broadcast news and MIXED uses PARL and BN data. The different systems (RBM, NONE, SEG and FRM) are described in Section 3.

To automatically generate the transcripts, we used a system trained on manually transcribed MP-FR-core data for the French systems (MP-FR-ext and Ester). To automatically transcribe the MP-GE-ext data, we used an ASR system trained on manually transcribed MP-GE-core data.

Since both French databases share a common phoneset, they can easily be combined during DNN training. We therefore also evaluate a system on all the un-transcribed French data (more than 100 h). For the confidence measure based systems, we also retained 40% of the data, i.e., about 40 h.

5. RESULTS

We first discuss the hypotheses under investigation and then present the experimental results.

5.1. Prior expectations

- Previous studies suggest that pre-training is language-independent [8, 9]. We hypothesize that pre-training is both domain and language independent.
- It was found that automatically transcribed data can improve the ASR performance [2, 8]. We hypothesize that automatically transcribed foreign data is beneficial, independently of language or domain.
- We hypothesize that the system with the most data (MP-FR-ext & Ester) yields the most improvement.
- We hypothesize that the frame based confidence measure outperforms the utterance based one because the frame based method is able to identify bad frames in a generally well recognized segment.

5.2. Results

The results are shown in Table 2 and confirm that pre-training is indeed language and domain independent. Experiments indicate that the amount of data available for pre-training is more important than the origin of the data.

Our study also shows that none of the systems that used foreign data with automatic transcription performed worse than the baseline system. The performances of the MP-GE-ext and the Ester systems are very similar, with a slight advantage for the Ester systems, hence it seems that hypothesis two is confirmed. However, the system that has matched domain and language (MP-FR-ext) performs considerably better than MP-GE-ext and Ester. Thus, matched data helps most.

The systems that used the combined dataset MP-FR-ext & Ester perform similarly to the MP-FR-ext systems. Therefore, we must reject the third hypothesis. If we have access to un-transcribed matched data, un-transcribed data from a different language or domain does not seem to be beneficial.

Overall, the frame-based confidence measure method seems to outperform the segment-based one (except for MP-FR-ext where they perform similarly). Thus, we accept hypothesis four. The frame-based confidence measure method allows to significantly reduce the amount of involved foreign data (down to 40%) without significantly degrading the ASR performance.

Finally, two systems performed best, with a 5% relative improvement over the baseline: MP-FR-ext and MP-FR-ext & Ester, with automatic transcription (no confidence).

6. CONCLUSION

We investigated how important domain and language are, if foreign datasets are exploited for unsupervised and semi-supervised DNN training. Our study indicates that the amount of data that is used for RBM pre-training is more important than the origin. For semi-supervised training however, matched data yields most improvement. If the data is not matched, it seems that domain and language are equally important. Furthermore, we also found that the amount of foreign data used for semi-supervised training can be significantly reduced without degrading the ASR performance if data selection relying on frame-based confidence measure is employed.

7. REFERENCES

[1] N. T. Vu, F. Kraus, and T. Schultz, "Multilingual a-stabil: A new confidence score for multilingual unsuper-

vised training," in *Proc. of the IEEE Workshop on Spoken Language Technology (SLT)*, 2010, pp. 183–188.

[2] S Thomas, M. L. Seltzer, Church K., and Hermansky H., "Deep neural network features and semi-supervised training for low resource speech recognition," in *Proc. of ICASSP*, 2013, pp. 6704–6708.

[3] D. Imseng, P. Motlicek, H. Bourlard, and P. N. Garner, "Using out-of-language data to improve an under-resourced speech recognizer," *Speech Communication*, 2013.

[4] 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.

[5] Y. Huang, D. Yu, Y. Gong, and C. Liu, "Semi-supervised GMM and DNN acoustic model training with multi-system combination and confidence re-calibration," in *Proc. of Interspeech*, 2013.

[6] D. Erhan, Y. Bengio, A. Courville, P.-A. Manzagol, P. Vincent, and S. Bengio, "Why does unsupervised pre-training help deep learning?," *Journal of Machine Learning Research*, pp. 625–660, 2010.

[7] S. Thomas, S. Ganapathy, A. Jansen, and H. Hermansky, "Data-driven posterior features for low resource speech recognition applications.," in *Proc. of Interspeech*, 2012.

[8] K. Veselý, M. Hannemann, and L. Burget, "Semi-supervised training of deep neural networks," in *Proc. of ASRU*, 2013.

[9] P. Swietojanski, A. Ghoshal, and S. Renals, "Unsupervised cross-lingual knowledge transfer in DNN-based LVCSR," in *Proc. of the IEEE Workshop on Spoken Language Technology (SLT)*, 2012, pp. 246–251.

[10] S. Galliano, E. Geoffrois, G. Gravier, J.-F. Bonastre, D. Mostefa, and K. Choukri, "Corpus description of the ester evaluation campaign for the rich transcription of french broadcast news," in *Proc. of the International Conference on Language Resources and Evaluation (LREC)*, May 2006.

[11] G. Hinton, Li Deng, Dong Yu, G.E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T.N. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

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

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

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

[15] P. Koehn, "Europarl: A parallel corpus for statistical machine translation," in *Proceedings of the 10th Machine Translation Summit*, 2005, pp. 79–86.