Cheap Bootstrap of Multi-Lingual Hidden Markov Models

Daniele Falavigna and Roberto Gretter

HLT research unit, FBK-irst, 38100 Povo(TN), Italy

falavi@fbk.eu, gretter@fbk.eu

Abstract

In this work we investigate the usage of TV audio data for cross-language training of multi-lingual acoustic models. We intend to take advantage from the availability of a training speech corpus formed by parallel news uttered in different languages and transmitted over separated audio channels.

Spanish, French and Russian phone Hidden Markov Models (HMMs) are bootstrapped using an unsupervised training procedure starting from an Italian set of phone HMMs. The use of confidence measures in order to select the training audio data was also investigated and has proved to be effective. The usage of cross language information, i.e. exploiting the temporal alignment of news in different languages to build news-dependent Language Models (LMs), was also demonstrated to give benefits to the acoustic model training.

Index Terms: multi-lingual speech recognition, cross-language bootstrap, confidence measures, unsupervised training.

1. Introduction

The development of multi-lingual Automatic Speech Recognition (ASR) systems is a task that has been largely investigated in the past by many researchers, who proposed approaches based on the usage of: language specific [1], language universal [2] [3] and language adaptive [4] [3] acoustic models.

In general, the training of language specific acoustic models represents the best practice to adopt when a sufficient quantity (i.e. hundreds of hours) of audio recordings is available for the given language. On the contrary, when a reduced set of training data (i.e. tens of hours of audio recordings or less) is available for a language, two different approaches can be used:

- cross-language bootstrap [1] of the Acoustic Model (AM) of the target language starting from that of a well trained source language and subsequent training, or adaptation, e.g. applying linear transformations estimated with either Maximum Likelihood Linear Regression (MLLR) or with Maximum A-Posterior (MAP) probability approaches, using the available set of training data of the target language;

- training of a universal set of acoustic models using all of the available training data of all languages [2] [3] [4], possibly followed, also in this case, by a language dependent adaptation step.

Advantages and disadvantages of the two approaches have been deeply discussed in the literature [3] [4]. In particular the work reported in [3] shows that, if the phone context coverage of the target training data is good enough, "it is possible, in many cases, to build general-purpose language-universal and language-adaptive acoustic models that outperform language-specific ones".

In this work we started to investigate the usage of parallel audio data for the cross-language training of multi-lingual acoustic models. As will be seen in the next sections we intend to take advantage from the availability of a training speech corpus, named "Parallel News" (see Section 2), formed by multi-lingual parallel news, i.e. a same piece of news is uttered in several different languages and transmitted over separated audio channels.

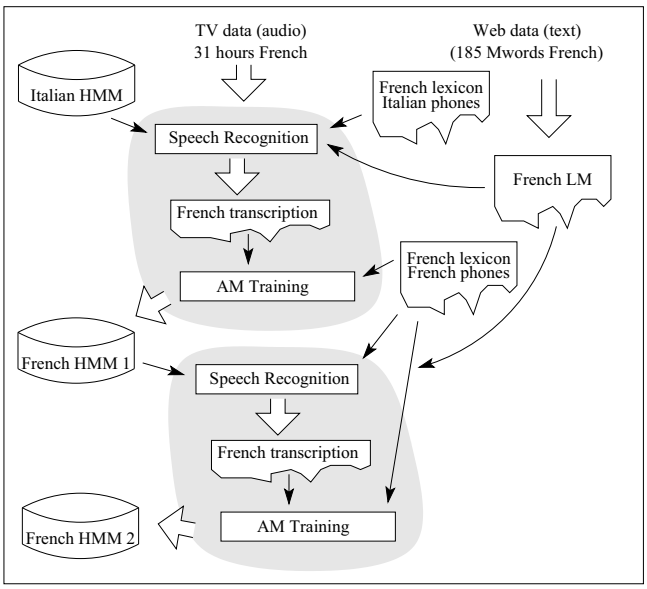

Figure 1: Block diagram of the cross-language bootstrap procedure (Italian to French case).

The proposed approach is similar, in principle, to that reported in [5], differing from this latter one on the way the parallel data are used for cross-language training.

Figure 1 shows the "unsupervised" training procedure used for bootstrapping the phone HMMs of a target language (French) starting from those of a "well trained" source language (Italian). With the set of Italian HMMs we automatically transcribe the French audio training data using a French Language Model (LM), estimated on text data collected from the Internet, and a lexicon expressed in terms of the Italian phones.

Then, a first set of French HMMs (HMM-1 in Figure 1) is trained and used to re-transcribe the French audio training data; this second transcription step makes use of a French lexicon. A second set of French HMMs (HMM-2 in Figure 1) is then trained using the new resulting transcriptions. Note that the procedure shown in Figure 1 could be iterated several times.

A preliminary experiment, referred to the Russian language, makes use of confidence measures. In this case, after each automatic transcription step, a procedure is applied that selects a subset of audio segments corresponding to the most confident words. Details are given in Section 5.

In addition, a novel approach which exploits cross language information is described in Section 6. With this approach, referring to Figure 1, the decoding steps use a French LM derived from the automatic translations of the Italian news temporally aligned to the French ones.

2. The Parallel News Speech Database

International news are acquired from a satellite TV channel broadcasting news in different languages, potentially attractive as a source of parallel speech data. In one of these channels, news are transmitted in ten different languages: Arabic, English, French, German, Italian, Portuguese, Russian, Spanish, Turkish and Persian, each one over a different audio channel in the digital TV stream. As the video content is the same for all of the audio channels, the news can be considered time aligned.

Basically, one of these channel broadcasts a cyclic schema that lasts about 30 minutes, and roughly consists of: main news of the day (politics, current events); music & commercials; specialized services (stock, technology, history, nature); music & commercials. During the day, each piece of news has a life cycle: it can be repeated several times, according to its importance, it can be enlarged, shortened or updated, depending on other events that can occur.

From an ASR perspective, data cannot be considered really clean, in the sense that several phenomena take place: often, in case of interviews, but not only, some seconds of speech in the original language are played before the translation starts; commercials are often in English; there is the presence of music; sometimes a particular piece of news has not been translated yet in all of the languages, so that some channels may contain the original audio (in another language).

Concerning speech alignment, or for Machine Translation (MT) purposes, it has to be said that news in different languages are **not** exact translations one of each other. Sometimes the same piece of news is approached from different point of views, sometimes one language gives more details than the others.

Since May 2009 we digitally record every day one hour of audio stream: we first extract the list of audio channels and then, for each language, its audio track which is stored at 16kHz sampling rate.

Given the particular structure of the stream, it is quite easy to segment the different audio tracks into news, because the boundaries among news are characterized by background noise over all of the channels. So, from a practical point of view, a good and simple approach is to detect pauses on each separate channel, and consider as probable news boundaries the pauses that are common to all channels.

3. ASR System

The ASR system used for this work is the one developed in our labs [10], which makes use of state-tied, cross-word triphone HMMs, with output distributions modeled with mixtures of up to 16 diagonal covariance Gaussian densities.

The ASR system works with two decoding steps. The first step embeds cluster-based cepstral mean subtraction, variance normalization, and projection of acoustic features (i.e. mel frequency cepstral coefficients, frame log-energy and their corresponding first, second and third order time derivatives) based on heteroscedastic linear discriminant analysis.

The output of the first decoding step is used as a supervision for the second decoding step where, exploiting normalized (via constrained MLLR) acoustic features and adapted acoustic models, the best sequence of word hypotheses is generated.

Table 1: Statistics of the text data used to train the three LMs. All data are in millions.

language	# words	# 4-grams	# FSN states	# FSN trans
Spanish	296.1	31.4	45.0	118.0
French	242.9	21.5	30.4	84.2
Russian	11.5	5.4	22.5	52.3

3.1. Acoustic Models

At present, we have addressed three languages: Spanish, French and Russian. For each one of them we used, in the experiments reported below, 31 hours of untranscribed audio recordings for training corresponding HMMs. To this purpose we have adopted the speaker adaptive training procedure described in [6] starting, after each of the iterations shown in Figure 1, from scratch. To be more precise, the amount of speech data really used for acoustic training is less than 31 hours, due to the fact that some portions have been classified as music or noise, and thus excluded. For the three languages the retained speech measured in hours is: 23.4 for French, 23.9 for Spanish and 24.2 for Russian.

Worth note that the reason for having limited this work to Spanish, French and Russian target languages relies on the fact that manually transcribed test sets are available only for them (see Section 4 for the details). At present, we are extending the test sets in the Parallel News domain and, in a near future, we plan to cope with other languages as well.

3.2. Language Models

The collection of text data for training n-gram based LMs has been carried out through web crawling. Since May 2009 we have downloaded, every day, text data from various sources, mainly newspapers, in different languages. Due to the different number of web sources, the amount of texts we have collected varies across the languages. Some statistics related to them are given in Table 1: these text sets have been used for training corresponding 4-grams based LMs, and to build the Finite State Networks (FSNs) used for recognition. A crucial task for LM training from web data is text cleaning and normalization: several processing steps are applied to each html page to extract the relevant information, as reported in [7]: as a result, the text data are divided into articles. Finally, the size of the recognition dictionary was limited to 30k words for all languages.

3.3. Pronunciation Models

The multi-lingual ASR approach we are proposing makes use, for each of the selected languages, of a subset of the SAMPA phonetic alphabet. For Spanish, French and Russian we developed a phonetic transcriber, based on grapheme-to-phoneme rules. Note that the development of this latter module, together with procedures for processing numbers, are the only manual efforts required to build a new language.

Finally, for bootstrapping the phone HMMs of each target language, starting from those of Italian (the source language), we need to map each target phone into a source one. This mapping is of course questionable, especially for phones absent in the source language. As an example, Table 2 gives the phone conversion tables used in this work for French.

Table 2: Adopted SAMPA conversion table from French to Italian.

French ↓ Italian	2 → o E → e a∼ → a n @ → e Z → dZ	9 → u O → o e∼ → e n N → n g y → u	9∼ → u n R → r o∼ → o n h → sil	A → a H → sil y → e z → dz

4. Experiments and Results

To measure the performance of the proposed unsupervised HMM training approach, we manually transcribed some portions of the Parallel News recordings (not included in the training sets) in the different languages. This task was performed in our labs and is still ongoing. For test, we run the ASR using a fixed LM, changing only the acoustic models.

Some statistics related to the selected test data in the three languages are given in Table 3.

Table 3: Statistics of the test sets used in the ASR experiments.

language	duration	#words	PP	%OOV
Spanish	31 min 8 s	4987	130.5	0.46%
French	40 min 4 s	7087	116.5	1.62%
Russian	35 min 42 s	4131	1362.7	1.82%

The first two columns of Table 3 gives the overall durations of the test recordings and the total numbers of reference words, the last two columns report perplexities (PP) and Out-Of-Vocabulary (OOV) rates, computed with the LMs described above. Note the very high PP for Russian, due to the small amount of training data (see Table 1) and the fact that it is a highly inflected language. On the contrary, OOV rates exhibit reasonable values for all languages.

Figure 2 shows the Word Accuracies (WA) obtained on the test sets of all languages, together with the corresponding Phone Accuracies (PA). The latter measure is approximate and was computed by replacing each word, both in the reference and in the recognized word strings, with the first phonetic transcription found in the lexicon. In fact, we noticed that the most frequent errors correspond to substitutions of words having exactly the same phonetic transcription, or being slightly different due to morphological variations. For this reason, we believe that PA is better suited for measuring the effectiveness of the AMs.

Since for Spanish we also have at our disposal AMs trained on a quite large set (hundreds of hours) of manually transcribed speech data coming from a different domain (European Parliament Political Speeches, EPPS [12]), we decided to run a comparative experiment. This latter is indicated as *Spanish supervised* in Figure 2. Note that both PA and WA are better in the unsupervised case (PA: 92.6% against 92.5%, WA: 84.3% against 82.0%). In our opinion, this can be attributed to the acoustic mismatch between EPPS and Parallel News data.

From the Figures it also appears that, for Spanish, the bootstrap with the Italian phones performs quite well, and the convergence is pretty fast: the second and third iteration basically don't exhibit differences, while for the other languages the starting point is much worse and more iterations are needed to converge.

The absolute differences in both WA and PA among languages could be explained by basically observing two factors:

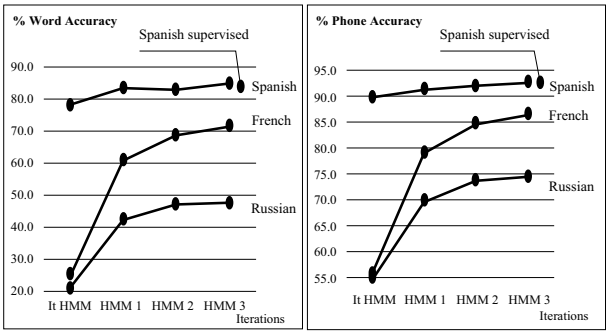

Figure 2: Word and Phone Accuracies obtained for each of the target languages and at each training iteration (see Figure 1).

first of all, the perplexities of the LMs are very different (by looking at them, results for Russian are even surprising); then, the acoustic similarity among languages is different. In particular, Italian is much more similar to Spanish than Russian and French.

5. Use of Confidence Measures

As mentioned in the introduction, for the Russian language we carried out a selection of the audio data used to train HMMs based on word confidence measures.

These latter ones are computed from word graphs generated in the second decoding ASR step using a method similar to the one reported in [11]. After each one of the recognition steps shown in Figure 1, we removed from the training data the segments corresponding to words exhibiting a confidence below a given threshold. This latter one was chosen in order to retain about 70% of the original training audio data, thus passing from about 24 hours of training data to about 17 hours.

From Table 4 we note that the usage of confidence measures allows to significantly increase both WA and PA. We expect further improvements from the optimization of the threshold used to define the percentage of audio data to retain for training HMMs (till now, no specific experiments have been carried out). Better improvements are also expected from the usage of LMs having lower perplexities that should allow to: achieve higher WAs, generate better WGs (i.e. with lower graph error rates) and, hence, to estimate more reliable word confidences.

Table 4: Comparison on Russian between standard recognition (identical to Figure 2), using Confidence Measures (CM) and Cross Language information (CL) (see Section 6).

	It HMM	HMM 1	HMM 2	HMM 3
Word Acc	20.8%	42.3%	47.2%	47.8%
Word Acc CM	20.8%	43.9%	50.3%	51.7%
Word Acc CL	20.8%	40.0%	48.6%	49.2%
Phone Acc	54.5%	69.6%	73.6%	74.5%
Phone Acc CM	54.5%	69.5%	75.0%	76.6%
Phone Acc CL	54.5%	66.4%	74.8%	75.7%

6. Exploiting Cross Language information

In this section we report initial experiments which exploit ASR in one language (Italian) to improve AM training on another language (French), taking advantage from the temporal alignment of news in the audio streams.

Table 5: Word Accuracy on the small news development set

AM \ LM	*base*	*10A*	*30A*	*70A*	*100A*
It HMM	25.5%	29.7%	44.0%	46.7%	45.7%
HMM 3	81.9%	53.1%	73.1%	72.3%	69.1%

As mentioned at the end of Section 2, by exploiting the fact that pauses common to all channels often identify news boundaries, we automatically segmented the audio tracks into separated aligned news, obtaining 2053 segments. Thus, on average each piece of news has a duration of about 54 seconds.

To build a LM focused on a given news in French, we proceed as follows. We process the Italian audio track using our state-of-the-art ASR, obtaining a transcription in Italian for each of the 2053 segments. Then, we apply a Machine Translation tool (Google Translate in this experiment) to obtain the corresponding *French text*. Then, we use this *French text* to score the articles used to train the French LM (see Section 3.2). We sort the articles according to this score and select the first N, which are in turn used to build a quite small 4-gram French LM.

The score used to select the articles is based on a count of common words between the *French text* and one article, excluding the most frequent words in the dictionary (mostly functional words that don't carry semantic meaning) [13]. The rationale is to select the articles more similar, in terms of lexical entries, to the *French text*. Future experiments will explore the possibility of simply using some dictionaries, instead of Google Translate or other MT engines, to get word-by-word translation.

We build in total 2053 small 4-gram French LMs, each one used to recognize the corresponding audio segment.

To tune this procedure we prepared a small 5 news development test, composed of manually transcribed Italian-French parallel news. We performed ASR on the 5 French news by using two AMs, respectively the first one (Italian HMM) and the last one (HMM3) of Figure 2. We compared 5 different LMs: the one used in Figure 2, here named *base*, and 4 LMs obtained by selecting the top 10, 30, 70 and 100 articles with the above described procedure. We name these LMs *10A*, *30A*, *70A*, *100A*. Results on the development set are reported in Table 5.

We noticed that, while we can obtain significant gain (from 25.5% to 46.7% WA) when using a weak AM (It HMM), this is no longer true with a better AM (HMM 3). Hence, we decided to use the focused LMs only in the first iteration of the training procedure, being confident that the whole process could benefit if we can start from a higher point in the curve. In the other iterations, the standard LM was used.

Results are shown in Table 6, and confirm this initial idea: the CL experiment gives a sensible improvement, especially at the second iteration. We also tested the very same approach on Russian, see label CL in Table 4. Despite an inital drop, also in this case the improvement over the standard case is sensible.

Table 6: Comparison on French between standard recognition (identical to Figure 2) and using Cross Language information.

	It HMM	HMM 1	HMM 2	HMM 3
Word Acc	25.4%	58.6%	67.8%	71.2%
Word Acc CL	25.4%	61.5%	72.1%	72.7%
Phone Acc	56.1%	77.8%	84.1%	86.2%
Phone Acc CL	56.1%	78.5%	86.2%	87.1%

7. Discussion and Future Work

In this paper we have presented a cheap method to bootstrap phone HMMs for three languages: Spanish, French and Russian. Text data for training LMs come from the web, acoustic data come from TV transmission. Phone HMMs are built in a complete unsupervised manner starting from existing Italian phone HMMs. Comparative experiments show that the bootstrap procedure can benefit both from the usage of confidence measures, and from the usage of cross language information, by exploiting ASR in the source language to "predict" a news specific LM in the target language.

A speech corpus, *Parallel News*, is still under acquisition, where there is a temporal alignment of news in ten different languages. In the next future we plan to use about 300 hours of parallel speech recordings (instead of 31) and to extend the experiments to other languages of the Parallel News database.

Several factors will be investigated to exploit cross lingual information, among them we will focus on: joint use of confidence measures and cross lingual information; composition (e.g. using ROVER) of multiple ASR output; better selection of articles; use of multiple source languages (e.g. Italian plus English) to predict the news specific LM.

8. References

[1] T. Schultz and A. Waibel, "Fast Bootstrapping of LVCSR Systems with Multilingual Phoneme sets," in *Proc. of EUROSPEECH*, Rhodes, September 1997, pp. 371–374.

[2] J. Kohler, "Multilingual Phoneme Recognition Exploiting Acoustic-Phonetic Similarities of Sounds," in *Proc. of ICSLP*, Philadelphia, Oct. 1996, pp. 2195–2198.

[3] H. Lin, L. Deng, D. Yu, Yi fan Gong, A. Acero, and C.H. Lee, "A Study on Multilingual Acoustic Modeling for Large Vocabulary ASR," in *Proc. of ICASSP*, Taipei, April 2009, pp. IV–4333–4336.

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

[5] M. Paulik, S. Stuker, C. Fugen, T. Schultz, T. Schaaf and A. Waibel, "Speech Translation Enhanced Automatic Speech Recognition," in *Proc of ASRU*, San Juan, Puerto Rico, Dec. 2005.

[6] D. Giuliani, M. Gerosa, and F. Brugnara, "Improved automatic speech recognition through speaker normalization," *Computer Speech and Language*, vol. 20, pp. 107–123, 2006.

[7] C. Girardi, "HtmlCleaner: Extracting Relevant Text from Web Pages," in *Proc. of WAC3*, Louvain-la-Neuve, Belgium, September 2007, pp. 15–16.

[8] M. Bacchiani and B. Roark, "Unsupervised Language Model Adaptation," in *Proc. of ICASSP*, Hong-Kong, 2003, pp. 224–227.

[9] D. Falavigna, M. Gerosa, R. Gretter, and D. Giuliani, "Phone-To-Word Decoding Through Statistical Machine Translation and Complementary System Combination," in *ASRU Workshop 2009*, Merano, Italy, December 2009, pp. 519–524.

[10] F. Brugnara, M. Cettolo, M. Federico, and D. Giuliani, "A Baseline for the Transcription of Italian Broadcast News," in *Proc. of ICASSP*, Istanbul, Turkey, June 2000, pp. 1667–1670.

[11] G. Evermann and P. .C. Woodland, "Large Vocabulary Decoding and Confidence Estimation Using Word Posterior Probabilities," in *Proc. of ICASSP*, Istanbul, Turkey, June 2000, pp. 2366–2369.

[12] D. Giuliani and F. Brugnara, "Experiments on Cross-System Acoustic Model Adaptation," in *ASRU Workshop 2007*, Kyoto, Japan, Dec. 2007.

[13] D. M. Blei, A. Y. Ng and M. I. Jordan, "Latent Dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993-1022, 2003