MULTILINGUAL A-STABIL: A NEW CONFIDENCE SCORE FOR MULTILINGUAL UNSUPERVISED TRAINING

Ngoc Thang Vu, Franziska Kraus and Tanja Schultz

Cognitive Systems Lab
Karlsruhe Institute of Technology (KIT)
thang.vu@kit.edu, franziska.kraus@student.kit.edu, tanja.schultz@kit.edu

ABSTRACT

This paper presents our work in Automatic Speech Recognition (ASR) in the context of multilingual unsupervised training with application to Czech. Starting without any transcribed acoustic training data we built a Czech ASR by combining cross-language bootstrapping and confidence based unsupervised training. We present our new method called "multilingual A-stabil" to compute confidence scores and explore the relative effectiveness of acoustic models from more than one language such as Russian, Bulgarian, Polish and Croatian for unsupervised training. While conventional confidence measures such as gamma and A-stabil [1] [2] work well with well-trained acoustic models but have problems with poorly estimated acoustic models, our new method works well in both cases. We describe our multilingual unsupervised training framework which gives very promising results in our experiments. We were able to select 80.5% of the audio training data (18.5 hours) with a transcription WER of 14.5% when using a small amount of untranscribed data (only about 23 hours). The final best WER on Czech is 23.6% on the development set and 22.9% on the evaluation set by using cross-lingual boostrapping, which is very close to the performance of the Czech ASR trained with 23 hours audio data with manual transcriptions (23.1% on the development set and 22.3% on the evaluation set).

Index Terms— multilingual ASR, unsupervised training, confidence score

1. INTRODUCTION

With the distribution of speech technology products all over the world, the fast and efficient portability to new languages becomes a practical concern. One of the major time and cost factor for developing LVCSR systems for new languages is the need for large amounts of transcribed training data. Detailed transcriptions require about 20-40 times real-time, and even after manual verification the final transcriptions are not free of errors. As described in [3] rapid development of an automatic speech recognition system can greatly benefit from the use of unsupervised acoustic model training, i.e. the use of ASR hypotheses as transcriptions, generated by a proceeding iteration of the automatic speech recognition system on untranscribed data. Typically unsupervised training is used to improve an available ASR through the use of additional acoustic data. For the best performance, confidence measures [1] [2] [3] [4] [5] derived from the recognizer output are used to select or weight the contribution of the acoustic training data. In some cases there is no ASR system for a new language at all. If so, the cross-language transfer technique [8] should be used, where a system developed for one language is applied to recognize another language without using any training data of the new language. After that, an unsupervised training might be applied to improve the word error rate (WER) iteratively [6] [7]. In this paper, we show that confidence scores generated by acoustic models from cross-language transfer do not perform well. This leads to the problem that the amount of selected data for unsupervised training is very small and has a high WER.

Our key question is: Do we need transcribed acoustic training data for new languages, if we already have ASR systems for many other languages? In our case, we evaluate a scenario that only Russian, Bulgarian, Polish, and Croatian ASR and Czech training data without any transcriptions are available. We modify the cross-language transfer and use it in combination with unsupervised training to investigate the development of the Czech acoustic model under this condition, i.e we explore the relative effectiveness of using acoustic models from more than one language for cross-language transfer. After that we present our new method to compute word-based confidence scores with very high precision based on the agreement of the outputs of several acoustic models.

The reminder of this paper is organized as follows. In section 2 we describe the characteristics of Slavic languages, data resources and our baseline system. Section 3 presents the cross-language transfer, its use to the current task and the phoneme mapping. In section 4 we propose a new method to compute the confidence scores. Section 5 reports the experimental results on the Czech dataset. The study is concluded in section 6 with a summary and an outlook to future steps.

2. SLAVIC LANGUAGES AND DATA RESOURCES

2.1. Peculiarities of Slavic Languages

The five languages Bulgarian, Croatian, Czech, Polish, and Russian investigated in this paper all belong to the Slavic branch of the Indo-European language family, which in total contains about 20 languages and dialects. Bulgarian and Croatian are South-Slavic languages, Czech and Polish are West-Slavic and Russian belongs to the East-Slavic branch. Russian has by far the largest speaker population (more than 165 Mio), Polish the second largest (about 56 Mio), while Czech and Bulgarian (both about 12 Mio) as well as Croatian (7 Mio) are significantly smaller. Slavic languages are well known for their rich morphology, caused by a high inflection rate of nouns using various cases and genders. With respect to the sound system, Slavic languages make use of a large number of palatal and palatalized consonants, which often are grouped with related non-palatalized consonants or form pairs of complex consonantal clusters. By contrast, the vowel inventory is very small for all languages. Polish has five basic vowels plus two nasal vowels, the other four languages only use the five basic vowels. Due to the rich morphology, word order is less important than in English and can thus be used as a mean of accentuation.

2.2. Data resources and baseline speech recognizer

GlobalPhone is a multilingual text and speech corpus that covers speech data from 20 languages, including Arabic, Bulgarian, Chinese (Mandarin and Shanghai), Croatian, Czech, English, French, German, Japanese, Korean, Polish, Portuguese, Russian, Spanish, Swedish, Tamil, Thai, Turkish, and Vietnamese [10]. The corpus contains more than 400 hours speech spoken by more than 1900 adult native speakers. GlobalPhone is available from ELRA, the European Language Resources Association. In each language about 100 native speakers read about 100 sentences each. The read texts were selected from national newspapers from the Internet. The read articles cover national and international political news as well as economic news from 1995-2009. The speech data is available in 16bit, 16kHz mono quality, recorded with a close-speaking microphone. Most transcriptions are internally validated and supplemented by special markers for spontaneous effects like stuttering, false starts, and non-verbal effects. For this work we selected five Eastern European languages from the GlobalPhone corpus, namely Bulgarian, Croatian, Czech, Polish, and Russian. Bulgarian was collected in 2003, the others in 1995 and 1999. Table 1 summarizes information about the speech data which was used for the experiments.

To build the language models we used our Rapid Language Adaptation Toolkit (RLAT) [9] to crawl for each language several websites with link depth 20 in up to twenty days [13]. We applied some automated normalization steps, (1) special characters were deleted, (2) digits, cardinal numbers, and dates were normalized, (3) punctuation was deleted, (4) all text data was converted to lowercase, and (5) a linear interpolation scheme was applied to optimize the language model on the development set. Table 2 summarizes the websites, days of crawling, and the performance of the resulting language models (based on the development set) for Czech.

Table 1. *GlobalPhone speech: Number of speakers (length of audio data in minutes) for five Eastern European languages*

Languages	Training set	Dev set	Eval set
Bulgarian	63 (1,027)	7 (149)	7 (143)
Croatian	72 (725)	10 (123)	10 (105)
Czech	82 (1,382)	10 (142)	10 (161)
Polish	79 (1,162)	10 (171)	10 (140)
Russian	95 (1,187)	10 (149)	10 (143)

Table 2. *Text corpus for Czech*

Websites	OOV	PPL	#vocab
www.lidovky.cz (20)	3.9%	2,148	195K
halonoviniy.cz (5)	5.2%	2,699	166K
respek.ihned.cz (5)	6.6%	3,468	173K
hn.ihned.cz (5)	5.2%	2,600	63K
aktualne.centrum (5)	9.5%	3,792	102K
Inter. LMs	3.8%	2,115	277K

For acoustic modeling, we applied the multilingual rapid bootstrapping approach which is based on a multilingual acoustic model inventory. This inventory was trained earlier from seven GlobalPhone languages [11]. To bootstrap a system in a new language, an initial state alignment is produced by selecting the closest matching acoustic models from the multilingual inventory as seeds. The closest match is derived from an IPA-based phone mapping. In this work, we did a phone mapping for each language and trained five different acoustic models. They used the standard front-end by applying a Hamming window of 16ms length with a window overlap of 10ms. Each feature vector has 143 dimensions by stacking 11 adjacent frames of 13 coefficient MFCC frames. A Linear Discriminant Analysis transformation reduces the feature vector size to 42 dimensions. The model uses a fully-continuous 3-state left-to-right HMM. The emission probabilities are modeled by Gaussian Mixtures with diagonal covariances. Table 3 gives a breakdown of the trigram perplexities, OOV rate, vocabulary size and WER for the five Eastern European languages.

3. MODIFIED CROSS-LANGUAGE TRANSFER

Cross language transfer refers to the technique where a system developed in one language is applied to recognize another language without using any training data of the new

Table 3. *Perplexities (PP), OOV rate, vocabulary size and WER for the five Eastern European languages*

Languages	PP	OOV	Vocabulary	WER
Bulgarian (BL)	500	1.0%	274k	20.3%
Croatian (HR)	813	3.6%	362k	28.9%
Czech (CZ)	1,886	3.7%	276k	23.1%
Polish (PL)	1,372	2.9%	243k	24.3%
Russian (RU)	1,675	3.4%	293k	36.6%

language. In these experiments Czech serves as the target language while BL, HR, PL and RU serve as source languages. [8] presented two principle ways of achieving a phoneme mapping: manual mapping using the IPA scheme or a mapping that was automatically derived from data using a target language phone recognizer. In our case, we evaluated the scenario that we do not have audio training data with transcriptions for developing a Czech ASR, so we cannot build a phoneme recognizer. Therefore, we decided to use a manual mapping although in [8] slightly better performance is presented using an automatically derived mapping. In total there are 41 phones in Czech, 20 phones of them can be found in all four languages. For the Czech phones which do not appear in the phone set of the target language we manually chose the phone with the most similar properties according to IPA. Table 4 describes the phone mapping for non-identical Czech phones to the other languages and the identical phones.

Table 4. *Phone mapping for non-identical Czech phones to the other languages*

Czech	Bulgarian	Croatian	Polish	Russian
c [ʦ]	[ʦ]	[ʦ]	[c]	[ʦ]
dj [ɟ]	[d^j]	[d]	[d]	[d]
h [ɦ]	[k]	[x]	[ɦ]	[h]
mg [ŋg]	[m]	[m]	[m]	[m]
nj [ɲ]	[n^j]	[n^j]	[n]	[n^j]
ng [ng]	[n]	[n^j]	[n]	[n]
rsh [r]	[r]	[r]	[r]	[r]
rzh [r]	[r]	[r]	[r]	[r]
sh [ʃ]	[ʃ]	[s^j]	[ʃ]	[ʃ]
tj [t]	[t^j]	[t]	[t]	[t^j]
x [x]	[x]	[s^j]	[s^j]	[x]
zh [ʒ]	[ʒ]	[z^j]	[ʒ]	[z^j]
aa [ʌː]	[ʌ]	[ʌ]	[ʌ]	[ʌ]
aw [au]	[ʌ]	[ʌ]	[ʌ]	[ʌ]
ee [ɛː]	[ɛ]	[ɛ]	[ɛ]	[ɛ]
ew [iw]	[ɛ]	[ɛ]	[ɛ]	[ɛ]
uu [uː]	[u]	[u]	[u]	[u]
ii [iː]	[i]	[i]	[i]	[i]
oo [oː]	[o]	[o]	[o]	[o]
ow [ou]	[o]	[o]	[o]	[o]

In contrast to the original approach of cross-language transfer [8] we did not modify the acoustic model of the source language, but the pronounciation dictionary of the target language, i.e. we modeled Czech words with phones of the other four languages. These mapped dictionaries allow the use of the acoustic model of the source language in combination with the Czech pronunciation dictionary and language model to decode the Czech training data in order to generate automatic transcriptions. Figure 1 shows the idea of modified cross-language transfer with Polish as source language and Czech as target language. Therefore, in contrast to [8] we can

Identical Phones - name [IPA]
ch [ʧ], p [p], b [b], d [d] , k [k], g [g], m [m], n [n], f [f], v [v], s[s], z[z], j [j], l [l], i [ɪ], u [ʊ], a [ʌ], e [ɛ], o [o]

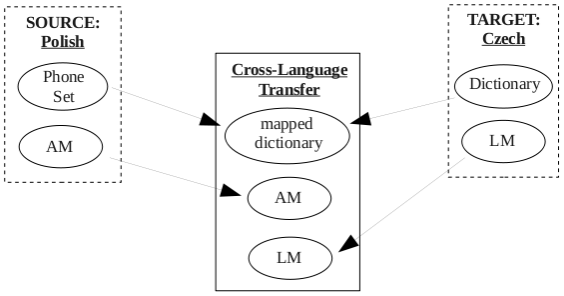

Fig. 1. Modified cross-language transfer with Polish as source and Czech as target language

use the context dependent system of the target language and thus benefit from the context similarity between languages. Furthermore, in our experiment the decoding took less time than it did by using the original method. The disadvantage of this method is that we adapted the phones of the source languages but did not train the Czech phones iteratively. For this reason, we used the modified cross-language transfer in this work only for generating automatic transcriptions but not for bootstrapping the final acoustic models.

4. CONFIDENCE SCORE BASED ON MULTILINGUAL ACOUSTIC MODELS

The basic idea of unsupervised training is to improve an acoustic model by iterative recognition of audio training data without manual transcriptions. Automatically generated transcriptions are used to retrain the acoustic model using this data. Typically this technique improves an available ASR through the use of additional acoustic data. For effective use of available acoustic data, it is important to utilize confidence measures to select or weight the contributions of the audio data so that only almost correct training data is used. In this paper, we use confidence scores as presented in [1]. For a given word lattice, the probability of any link can be computed in the same way as in the standard forward-backward algorithm for HMMs. The lattice node can be viewed as

HMM state and the links of the lattice give the possible transitions. As the nodes are associated with the words in the hypothesis, the emission probability of a node is the acoustical score of this time segment. The transition probability can be taken from the statistical language model which has been used in the decoding process. The result of the forward-backward algorithm is the confidence score, called gamma. [2] shows that this confidence score has a high correlation with the recognition error.

Another feature called acoustic stability (A-stabil) was also presented in [1] and also highly correlates with the recognition error. To compute this feature, a number of alternative hypotheses is generated. Each of these hypotheses is aligned against the reference output of the recognition, where the reference output is defined as the hypothesis with the best weighting between acoustic model and language model. For each word of the reference output the confidence score is defined as the number of times the same word occurs in the set of alternative hypotheses normalized by the number of alternative hypotheses. To evaluate gamma and A-stabil we plot the correlation between the selected confidence threshold and the recognition error. We used the CZ system to decode the development set and evaluated the WER of all words occurring in the specified confidence interval using steps of 0.1. Figure 2 compares gamma and A-stabil for two systems: a well-trained CZ system trained on 1,382 minutes of CZ training data and a CZ system resulting from cross-language transfer. It shows that gamma and acoustic stability work very well with well-trained acoustic models, but have problems with the initial acoustic models generated by cross-language transfer. Due to the poor performance of these confidence scores it is difficult to apply unsupervised acoustic model training. Hence, another more robust confidence score is required.

Based on the acoustic stability we propose a new method

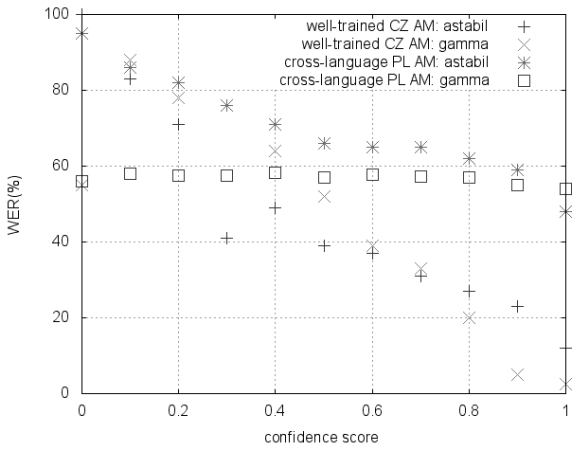

Fig. 2. The plot of recognition errors over gamma (and A-stabil) using a well-trained Czech acoustic model and an initial cross-language acoustic model (Polish).

to compute confidence scores using n monolingual acoustic models. In our case $n = 4$ and the acoustic models of Russian, Bulgarian, Croatian and Polish are applied. The implementation is almost the same as for acoustic stability. Using a set of alternative hypotheses derived from all four languages, we compute the frequency of each word of the reference output normalized by the number of alternative hypotheses. In order to generate the alternative hypotheses we build the word lattices first and use different weightings of acoustic model and language model of each language by rescoring. By applying this technique we got a multilingual arbiter which indicates the confidence for each word in the best hypothesis. Figure 3 illustrates the new method to compute word-based confidence score. We saw that the original

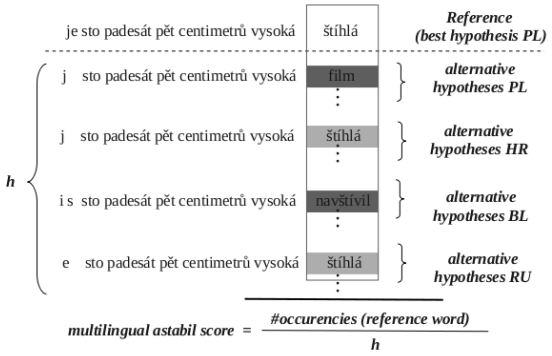

Fig. 3. Our new method to compute a word-based confidence score called "multilingual A-stabil"

definition of acoustic stability is a specialization of the new method selecting $n = 1$, that means monolingual. So we refer to it as "multilingual A-stabil". Figure 4 shows the recognition error over this score which presents a very high correlation between the feature and the recognition error for well-trained acoustic models and poorly estimated acoustic models. In contrast to gamma and A-stabil, multilingual A-stabil is much more robust.

5. EXPERIMENTAL FRAMEWORK AND RESULTS

5.1. Multilingual unsupervised training framework

In this section we present our multilingual unsupervised training framework, which consists of two main parts. The first part is an iterative process, in which we use more than one acoustic model to generate automatic transcriptions. We applied modified cross-language transfer to decode the untranscribed acoustic training and development data. Using the development set we evaluated "multilingual A-stabil" and estimated the optimal threshold. After that, all words that have a confidence score higher than this threshold were selected for adaptation in the next iteration. In our work a MAP adaptation was applied iteratively first to improve acoustic models

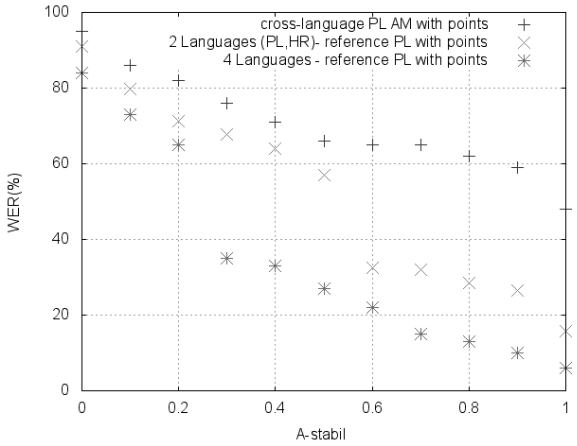

Fig. 4. The plot of recognition errors over A-stabil using an initial cross-language transfer acoustic model (Polish) and multilingual A-stabil with 2 (Polish and Croatian) and 4 (Bulgarian, Croatian, Polish and Russian) languages.

and thus increase the amount of data. This process terminates if the gain in amount of adaptation data from one iteration to the other is smaller than 5% relative. By using this process we could enlarge the amount of automatic transcriptions with a high precision on one side and select data from many different contexts due to the multilingual effect on the other side. In the second part, the original cross-language transfer was used to bootstrap the acoustic model of the target language with the selected data extracted in the first part. The final acoustic model is the one with the best performance on the development set. Figure 5 illustrates the multilingual unsupervised training framework.

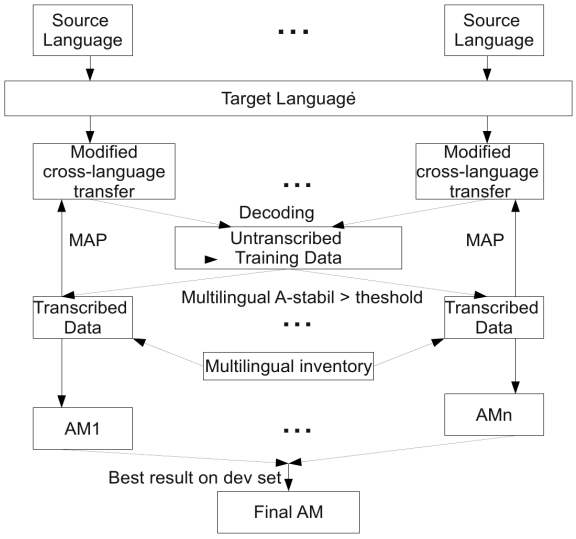

Fig. 5. Multilingual unsupervised training framework

5.2. Results

5.2.1. Generation of automatic transcripts for intial cross-language boostrapping

Starting with Russian, Bulgarian, Croatian and Polish we used the modified cross-language transfer without retraining to recognize the Czech development set. The word error rate is relatively high around 60%. Table 5 shows the performance of the initial cross-language models of each language on the Czech development set.

Table 5. *Performance of initial acoustic models using cross-language transfer on Czech development set*

	Bulgarian	Croatian	Polish	Russian
WER	61.04%	57.19%	55.83%	64.26%

With these initial models, we recognized the Czech training data and selected appropriate adaptation data using "multilingual A-stabil" confidence scores. Figure 4 shows the recognition error over this score on the development set for the first iteration. We observed that for the other iterations this curve has the same form and at a confidence score of 0.3, the WER decreases very strongly (for the first iteration from 67% to 35%). We assume that the reason lies in the multilingual effect. If the confidence score is less than 0.25, then all words in the alternative hypotheses, that are the same as the reference word could originate from one language, that means monolingual. So if this score is 0.3, then occurring words in alternative hypotheses must originate from more than one lanugage. [3] shows the trade-offs between accuracy and amount of retained data for a certain confidence threshold. For this reason we chose heuristically 0.3 as threshold to select data. The whole process terminated after 4 adaptation iterations. Table 6 shows the amount of selected data after each iteration in percentage of all untranscribed data and their quality in terms of WER.

Table 6. *Enlarging the amount of training data with automatic transriptions iteratively for instance in Polish*

Iteration	Amount of data	% of all data	Quality (WER in %)
1	5.5h	23.9	25.0%
2	14.3h	62.2	17.0%
3	15.9h	69.1	16.5%
4	16.4h	71.0	16.0%

5.2.2. Cross-language Boostrapping

After acoustic training data with high precision was selected, we used the boostrapping approach to train the Czech ASR by using the multilingual acoustic model inventory which was trained earlier from seven GlobalPhone languages [11]. To bootstrap the system, an initial state alignment is produced

by selecting the closest matching acoustic models from the multilingual inventory as seeds. The closest match is derived from an IPA-based phone mapping. After initialization the system is completely rebuilt using the selected data. We trained a quintphone system with 2000 contexts by applying merge&split and Viterbi training. Table 7 shows the performance of the four different systems which were trained with four different selected data sets on the Czech development set. The WER ranges from 24.1% to 24.6%. The best WER was achieved using the acoustic training data which was generated by modified cross-language transfer using Russian. To increase the amount of the acoustic training data, we recognized the training data using the acoustic model from the previous iteration and selected data with high confidence of "multilingual A-stabil". We obtained about 18.5 h (80%) of the training data with automatic transcriptions which have 14.5% WER. For the 2nd iteration we used the acoustic model from the 1st iteration to generate the state alignment and trained the system with the same parameter as in iteration 1 afterwards. Our best system has 23.6% WER on the development set and 22.9% WER on the evaluation set. The results show that

Table 7. *Performance of 2 iterations crosslingual bootstrapping on Czech development set*

Training	Bulgarian	Croatian	Polish	Russian	Data
1st iteration	24.5	24.6	24.4	24.1	16.4h
2nd iterarion	23.6	23.7	24.1	23.8	18.5h

the new confidence score "multilingual A-stabil" works well also with well-trained acoustic models. Using this confidence and iterative unsupervised training we can get more training data with highly accurate automatic transcriptions and also improve the acoustic model.

6. SUMMARY

In this paper, we described our investigations on multilingual unsupervised training. For this purpose we developed a Czech ASR without any transcribed training data using Russian, Bulgarian, Polish, and Croatian acoustic models. A combination of modified cross-language transfer and unsupervised training was applied. We explored the relative effectiveness of using acoustic models from more than one language for modified cross-language transfer. After that we proposed a new method to compute word-based confidence scores based on acoustic stability of several acoustic models, called "multilingual A-stabil". This method performs well not only with well-trained acoustic models but also with poorly estimated acoustic models. The results are very promising achieving 18.5 hours which is 80.4% of all available audio training data with automatic transcriptions with about 14.5% WER. The best system has 23.6% on the development set and 22.9% on the evaluation set, which is very close to the performance of the Czech ASR trained with 23 hours audio data with manual transcriptions (23.1% on the development set 22.3% on the evaluation set). In the future we plan to do experiments with more source languages, especially languages that do not belong to the same language family as the target language. Furthermore, we will apply the multilingual unsupervised training framework on more realistic data sets such as broadcast news data.

7. ACKNOWLEDGMENT

The authors are grateful to Dr. Florian Metze for his support. This work was partly realized as part of Quaero Programme, funded by OSEO, French State agency for innovation.

8. REFERENCES

[1] T. Kemp and T. Schaaf. Estimating confidence using word lattices. In Proc. of European Conference on Speech Communication Technology, pp. 827-830 ,1997.

[2] T. Schaaf and T. Kemp. Confidence Measures for Spontaneous Speech Recognition. In Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing - Volume 2, 1997.

[3] G. Zavaliagkos and T. Colthurst. Utilizing untranscribed training data to improve performance, in DARPA Broadcast News Transcription and Unterstanding Workshop, Landsdowne, VA, USA, Feb. 1998.

[4] F. Wessel, K. Macherey and H. Ney. A comparision of wordgraph and N-best list based confidence measures. In European Conference in Speech Communication and Technology, Budapest, Hungary, 1999.

[5] G. Evermann and P. Woodland. Large vocabulary decoding and confidence estimation using word posterior probabilities. In Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, Instanbul, Turkey, 2000.

[6] J. Lööf, C. Gollan, and H. Ney. Cross-language Bootstrapping for Unsupervised Acoustic Model Training: Rapid Development of a Polish Speech Recognition System. In Interspeech, pages 88-91, Brighton, U.K., 2009.

[7] Lori Lamel, Jean-Luc Gauvain and Gilles Adda. Unsupervised acoustic modelling. In: Proc. ICASSP Orlando, USA, 2002.

[8] T. Schultz and A. Waibel. Experiments on cross-language acoustic modeling. In Proc. European Conf. on Speech Communication and Technology, Aalborg, Denmark, 2001.

[9] T. Schultz and A. Black. Rapid Language Adaptation Tools and Technologies for Multilingual Speech Processing. In: Proc. ICASSP Las Vegas, USA 2008.

[10] T. Schultz. GlobalPhone: A Multilingual Speech and Text Database developed at Karlsruhe University. In: Proc. ICSLP Denver, CO, 2002.

[11] T. Schultz and A. Waibel. Language Independent and Language Adaptive Acoustic Modeling for Speech Recognition. In Speech Communication August 2001., Volum 35, Issue 1-2, pp 31-51.

[12] A.Stolcke. SRILM - an extensible language modeling toolkit, in Proc. Int. Conf. on Spoken Language Processing, Denver, Colorado, 2002.

[13] N.T. Vu, T.Schlippe, F.Kraus, T.Schultz. Rapid Bootstrapping of five Eastern European Languages using the Rapid Language Adaptation Toolkit. In Interspeech, Makuhari, Japan, 2010.