Enhancing ASR Systems for Under-Resourced Languages through a Novel Unsupervised Acoustic Model Training Technique

Horia CUCU[1], Andi BUZO[1], Laurent BESACIER[2], Corneliu BURILEANU[1]
[1] *Speech and Dialogue Research Laboratory, University "Politehnica" of Bucharest, Romania*
[2] *Laboratoire d'Informatique de Grenoble, University Joseph Fourier, Grenoble, France*
horia.cucu@upb.ro

***Abstract*—Statistical speech and language processing techniques, requiring large amounts of training data, are currently state-of-the-art in automatic speech recognition. For high-resourced, international languages this data is widely available, while for under-resourced languages the lack of data poses serious problems. Unsupervised acoustic modeling can offer a cost and time effective way of creating a solid acoustic model for any under-resourced language. This study describes a novel unsupervised acoustic model training method and evaluates it on speech data in an under-resourced language: Romanian. The key novel factor of the method is the usage of two complementary seed ASR systems to produce high quality transcriptions, with a Character Error Rate (ChER) < 5%, for initially untranscribed speech data. The methodology leads to a relative Word Error Rate (WER) improvement of more than 10% when 100 hours of untranscribed speech are used.**

***Index Terms*—speech recognition, under-resourced languages, unsupervised acoustic modeling, unsupervised training.**

I. INTRODUCTION

State-of-the-art Automatic Speech Recognition (ASR) systems for high-resourced languages use hundreds or even thousands of hours of *manually transcribed speech data* for training the acoustic model (AM) and corpora with billions of words to train the language model (LM). This is a critical issue in the development of a new ASR system, because the acquisition of such data is expensive and requires a lot of time. Under-resourced languages are characterized by lack of text corpora and annotated speech data, phonetic dictionaries, tools and language expertise.

Many acoustic and language adaptation techniques were proposed in the past decade to overcome this crucial issue in developing ASR systems for under-resourced languages [1]. However, these studies have addressed *the process of bootstrapping* to create a new, basic ASR system for a new language and did not focus on *what needs to be done further* to fill the performance gap between the ASRs for new languages and the ASRs for English, Mandarin, Arabic, which are trained on thousands of hours of speech.

For Romanian, a language for which five years ago there were no speech corpora available (neither for research, nor commercial usage), we started to create a low-cost ASR system using speech recorded in laboratory conditions [2]. We continued to improve the read-speech ASR system in a lightly-supervised training scenario, by using loose transcriptions of talkshows to adapt it to spontaneous speech also [3]. The main issue here is that loose transcriptions of audio-visual content are not widely available on the Internet (as opposed to the vast amounts of untranscribed audio-visual content). Consequently, the next logical step was to find a way to use this vast amount of untranscribed speech to further improve the ASR system. This scenario, in which untranscribed speech is used for acoustic modeling is called unsupervised acoustic model training. We recently introduced such a novel training method [4] and, in this paper, we thoroughly evaluate it in different scenarios and compare it with a basic unsupervised training technique.

The rest of the paper is organized as follows. In Section II we briefly explore the state-of-the-art in unsupervised acoustic model training and point out the main novelties of our study. In Section III we describe in detail the proposed method and the specific issues it addresses. In sections IV and V we present the experiments and in Section VI we draw some conclusions.

II. RELATED WORK AND KEY NOVEL FACTORS

The general procedure for unsupervised acoustic model training starts by using a seed acoustic model to transcribe a large amount of speech data. Afterwards, using confidence scoring and threshold optimization, a part of the transcribed data is selected for further retraining. The whole process can be repeated until the ASR system's performance saturates or until the amount of newly selected data is not significant anymore.

Unsupervised acoustic modeling is a relatively new research topic and there are only a few studies presenting different variations of the general procedure described above. The first tentative to train an acoustic model in an unsupervised fashion was presented in [5] and [6]. These studies use confidence scoring and threshold optimization to create acoustic models for Spanish and German. In [5], the authors used a Spanish ASR system, trained with a very small amount of data (3 hours of transcribed speech), to decode 25 hours of untranscribed speech. Afterwards, using confidence scoring and threshold optimization, they were able to select 2.7 hours of the ASR output for further retraining and obtained an improvement of 1.7% relative WER over the initial ASR system. In [6] the authors applied the unsupervised acoustic model training technique to create a German ASR system and they report much better results

This work has been funded by the Sectoral Operational Programme Human Resources Development 2007-2013 of the Ministry of European Funds through the Financial Agreements POSDRU /159 /1.5 /S/134398.

Digital Object Identifier 10.4316/AECE.2015.01009

(34% relative WER improvement over the initial ASR) using a different, lattice-based confidence score.

In [7] the authors explored the gains in ASR accuracy obtained for seed systems trained with different amounts of manually transcribed data. The conclusion was that unsupervised training cannot bring any accuracy improvements if the initial ASR system is trained on a large dataset. Moreover, the authors also investigated the gains in accuracy obtained if all the ASR hypotheses are used for retraining (i.e. no confidence measure is applied) versus the improvements obtained if the words posterior probability is used as confidence score for data selection. Finally, this study introduces for the first time the idea of iterative unsupervised training: the ASR system trained using the unsupervised training procedure is used to decode again the untranscribed speech data, which is further used in an unsupervised training procedure.

A somehow different variation of the iterative training procedure is introduced in [8]. In this study more and more untranscribed data is added progressively (the seed models are used to transcribe only a small part of the untranscribed data and generate better models. Going further, these models are used to transcribe a double amount of untranscribed data and generate better models and so on. The authors conclude that unsupervised training is almost as good as lightly-supervised training and that this procedure works with both high-quality and low-quality language models. The same research group also introduced a lattice-based unsupervised training method and reported even better results in [9].

Although the unsupervised acoustic modeling procedure was initially applied in the context of Maximum Likelihood (ML) training, several studies also investigated its usability for Maximum Mutual Information (MMI) training [10, 11] and Minimum Phone Error (MPE) training [10, 12]. In [12] the authors focus on the idea that, depending on the type of AM retraining (maximum likelihood or discriminative), the errors in the automatic generated transcriptions have different impacts on the final system performance. They argument that for discriminative AM retraining, it is desirable to select and transcribe manually some parts of the speech data, which are believed to be poorly recognized. These manual transcriptions are then used to supplement the fully automatic transcriptions.

As it was expected, the unsupervised AM training has been successfully applied to create or improve ASR systems for various new languages, such as Mandarin [10, 12], Arabic [11], Polish [13], Czech [14] and Vietnamese [15].

An innovative idea, recently introduced in [15], implies using several ASR systems, in six source European languages, to create transcriptions for speech data in a target language: Vietnamese. In this process the authors iteratively adapt the source ASR systems to the target language using unsupervised training based on the "multilingual A-stabil" confidence score [16]. Finally, they train a Vietnamese ASR system using the resulted transcriptions. In [14] the same authors use a similar multilingual unsupervised training procedure to develop a Czech ASR without any transcribed training data. They apply a combination of cross-language transfer and unsupervised training based on the same "multilingual A-stabil" confidence score.

In our study we explore the idea of using two complementary ASR systems for Romanian to transcribe new Romanian speech data, align and filter the ASR hypotheses and finally use the selected data to improve the main ASR system. The novelty of the proposed unsupervised training methodology involves *two key factors*: a) the unsupervised training process is based on two seed ASR systems and b) the data selection procedure does not involve confidence scoring and threshold optimization. A consequence of b) is that our method can also be used to transcribe raw speech data with mismatched acoustic conditions (e.g. dialects, elder speech), without the need to adapt any thresholds.

The usage of several seed ASR systems in unsupervised training was previously explored in [15], but in that study the seed models were for different languages.

The data selection procedure is totally new, different from the ideas reported in the literature: we do not use a confidence metric applied at state, word or sentence level on the output of a single ASR system, but instead we identify continuous, long sequences of identical words in the output of two ASR systems.

III. Method Description

The purpose of the proposed unsupervised training procedure is to improve an existing ASR system for a particular language (in our case Romanian). This system, further called *main ASR system*, was trained with a previously available annotated speech database and uses a language model created with a previously available text corpus. The proposed method has several steps and is illustrated in Figure 1.

The first step is to create two complementary ASR systems, called *seed ASR systems*, which will be used further to process untranscribed speech data. These seed ASR systems should reasonably make uncorrelated recognition errors and this fact can be exploited to select the (assumed) correct parts of the transcriptions (the aligned parts of the ASR hypotheses will be considered correct).

To create the seed ASR systems, the initial training speech database is split into two parts based on the type of speech, acoustic environment, etc. Each part of the initial training speech database is used independently to create one of the seed acoustic models. The LM training text corpus is split into two parts also and each part is used independently to create a seed language model.

The second step in the procedure is the acquisition and diarization of raw, untranscribed speech data. Speech data acquisition is most easily done over the Internet, by capturing radio or television broadcast streams. Other sources of raw speech data are audio books, user-recorded data, etc. The segmentation and diarization of the speech data is mandatory because the raw speech data can contain non-speech parts (music, jingles, advertisements, etc.) that should be filtered out before speech recognition. The segmentation also helps the ASR hypotheses alignment process (aligning short sequences of words is less error-prone than aligning long sequences).

Next, the cleaned, untranscribed speech data is decoded using the two seed ASR systems. The resulted pair of ASR hypotheses is aligned using a Dynamic Time Warping (DTW) algorithm. The DTW alignment process aims to

identify *sequences of identical words* in the two transcription hypotheses. If these sequences of identical words:

- are continuous (the time difference between two consecutive words is less than 2 seconds) and
- are long enough (they contain more than 8 characters and the corresponding audio is longer than 1 second),

then they are considered to be transcribed correctly and they are selected, together with the corresponding audio data, to create a new annotated speech corpus. The thresholds used in the above selection process were determined empirically (some of them in [3], others in various experiments performed for this study).

This selection procedure increases the probability that the selected data is correct, because it ignores singular short words and even short sequences of short words, which can appear very often in the ASR hypotheses, and it assures that all words are part of the same utterance. After the selection, the border timestamps for every word sequence are used to cut the corresponding audio parts out of the initial speech files.

The proposed alignment and selection procedure produces utterances longer than 1 second, comprising at least a few words, as opposed to the single-word utterances produced by the selection procedures discussed in Section 2. This is very important because the selected data will be used to retrain the ASR system and the longer the audio clips the better for the training process.

Finally, the last step involves retraining the main ASR system using the existing transcribed speech database and the newly annotated speech database.

A. *Iterative unsupervised training*

The newly annotated speech database obtained as described above can also be used to enhance the initial seed ASR systems, which can be used further to decode again the same untranscribed speech data. In order to maintain the complementarity of the seed ASR systems the newly annotated speech database is split into two distinct parts and each part is used to augment the initial training data for one of the seed ASR systems.

This iterative process can be repeated until the gain in performance for the main ASR system is not significant anymore.

B. *Alternative unsupervised training method*

Throughout the next sections we will compare the proposed unsupervised training method with an alternative method, which involves the following steps:

1. acquisition and diarization of raw, untranscribed speech data (exactly as in the proposed method);
2. decode with the main ASR system (the best available);
3. retrain the main ASR system using the previously existing transcribed speech database and the raw transcriptions generated at step 2.

This alternative unsupervised training method does not involve any data selection process: in this case the raw ASR transcriptions are used directly to retrain the main ASR system. Of course, among these raw ASR transcriptions there will be many incorrect ones. Nonetheless, this is the simplest unsupervised training approach, it also brings ASR improvements (as shown in the following section) and can be used as a baseline for our system.

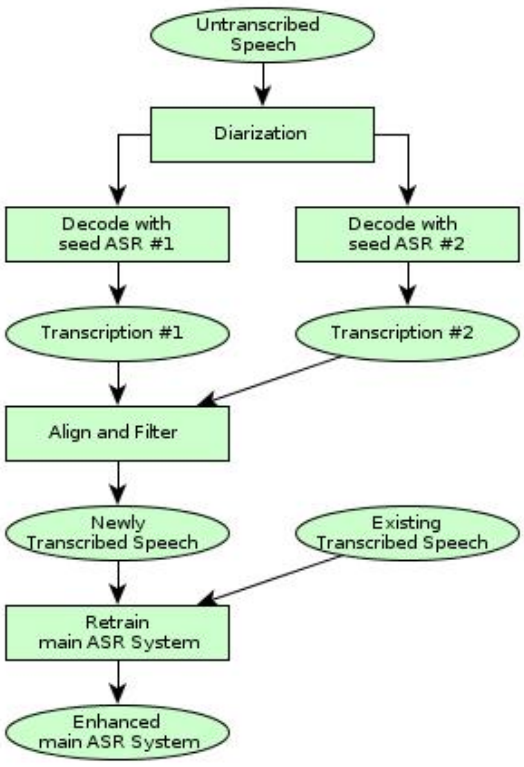

Figure 1. The block diagram of the proposed method

IV. EXPERIMENTAL SETUP

A. *Speech corpora*

ASR training and evaluation was done on several self-developed corpora (created by the Speech and Dialogue Research Group), because for the Romanian language there are no other speech corpora available.

The RSC-train (Read Speech Corpus) and RSC-eval corpora comprise Romanian read speech recorded by 165 speakers. The read speech corpora were obtained by recording various predefined texts, representing news articles and literature. The recordings were made in laboratory conditions, using an online recording application. More information regarding these corpora can be found in [2],[17].

The SSC-train (Spontaneous Speech Corpus) and SSC-eval corpora were created using a lightly-supervised acoustic modeling technique [3]. The originally loosely-transcribed speech data comprised broadcast conversational speech. A part of this speech data (SSC-eval) was manually annotated to create an error-free spontaneous speech corpus for evaluation only. This part consists of 3.5 hours of speech, among which 2.2 hours of clean speech. The remaining 1.3 hours of speech contains speech in degraded conditions (background noise, background music, telephone speech, etc.).

The SSC-untranscribed speech corpus was acquired over the Internet and contains broadcast news and conversational speech, without any transcriptions. SSC-untranscribed was segmented and diarized as described in the previous section,

using the LIUM Speaker Diarization Toolkit [20]. The segmentation and diarization processes aimed to filter-out all the non-speech parts of the corpus and to create single-speaker utterances.

For the initial supervised training we used a selection of 10 hours of read speech from the RSC-train corpus and a selection of 10 hours of conversational speech from the SSC-train corpus. For the unsupervised training procedure we used a selection of 100 hours of speech from the SSC-untranscribed corpus. For evaluation we used the evaluation part of the Read Speech Corpus (RSC-eval), with 6 hours of speech, and the evaluation part of the Spontaneous Speech Corpus (SSC-eval), with 3.5 hours of speech.

B. Acoustic models

All acoustic models used in this study are 5-state HMMs with output probabilities modeled with GMMs. As speech features we used the recently introduced noise robust features: Power Normalized Cepstral Coefficients (PNCCs) plus their first and second temporal derivates (13 PNCCs + deltas + double deltas). In all cases the 36 phonemes in Romanian were modeled contextually (context dependent phonemes) with 4000 HMM senones. The number of Gaussian mixtures per senone state was varied (8/16/32) in order to adapt the acoustic model setup to the size and variability of the training speech corpus. The acoustic models were created and optimized (using the CMU Sphinx Toolkit [18]) with the various training speech corpora mentioned above.

C. Language models

An online-newspaper text corpus (with 169M words) and a talkshows transcriptions corpus (with 40M words) were used independently to create two complementary language models. These two language models were used in the two seed ASR systems. A third language model obtained by interpolation was used in the main ASR system. The three language models have different sizes and vocabularies, as they were created with different text corpora. The models were developed with the SRI-LM Toolkit [19].

V. Experimental Results

Throughout the experimental results section the conditions of an under-resourced language were simulated by using only 10 hours of read speech and 10 hours of conversational speech for the initial training.

The baseline main ASR system was trained with both the two sets of training data (20 hrs) and uses the interpolated language model. The baseline seed ASR #1 was trained with read speech only (10 hrs) and uses the news LM (in which the language is closer to read-speech). The baseline seed ASR #2 was trained with conversational speech only (10 hrs) and uses the talkshows LM (in which the language is closer to conversational speech). The idea behind this was to create seed ASR systems with correlated models, so that each system is adapted to its type of speech (read or conversational). These baseline systems are evaluated in Table I.

The goal of the first experiment was to evaluate the quality of the transcriptions generated by the proposed unsupervised training method. We used the SSC-eval corpus

TABLE I THE BASELINE MAIN ASR SYSTEM AND THE BASELINE SEED ASR SYSTEMS

ASR system	**Initial training corpus**	**WER [%]**	
		RSC	**SSC**
mainASR – baseline	read+conv (20 hrs)	26.5	42.2
seedASR #1 – baseline	read (10 hrs)	32.6	53.8
seedASR #2 – baseline	conv (10 hrs)	46.2	47.8

as if it was an untranscribed speech corpus, transcribed it with the two seed ASR systems and aligned the two hypotheses transcriptions. We used the timestamps of the aligned hypotheses to select the corresponding parts in the reference transcriptions and finally computed the WER and ChER for the aligned hypotheses. Although the WER was higher than expected (10.4%), we noticed that the word errors in the aligned hypotheses were usually substitutions of similarly pronounceable words. Most often, these word substitutions were due to single letter substitutions and therefore the ChER for the aligned hypotheses was much lower: 4.9%. The conclusion is that the seed ASR systems sometimes make correlated errors and these errors propagate in the aligned hypotheses. Consequently, the mainASR system will not be retrained with 100% correctly transcribed speech data, but with almost correctly transcribed data (ChER = 4.9%).

Next the unsupervised acoustic training method was applied on 20, 50, and finally 100 hours of untranscribed speech. The application of the method resulted in 5.2, 12 and respectively 22 hours of automatically transcribed speech. The main ASR systems retrained using this data are evaluated in

Table II. The conclusion arising from this experiment is that the WERs on both read and conversational speech decrease when more untranscribed speech is used, but this performance improvement is not linear (saturation will be reached at some point). For the maximum amount of untranscribed data used in this experiment (100 hrs) we obtained a relative WER improvement of 10.2% on conversational speech and 11.3% on read speech.

In order to asses the effectiveness of the proposed method we compared it with the alternative method described in Section III.B, which implies using all the raw transcriptions to retrain the baseline main ASR. The results are summarized in Table III. Here it is remarkable the fact that increasing the size of the untranscribed corpus does not necessarily lead to performance improvements. This happens because the raw transcriptions used for retraining

TABLE II. USING VARIOUS AMOUNTS OF UNTRANSCRIBED DATA FOR THE PROPOSED UNSUPERVISED TRAINING METHOD

ASR system	**Untranscribed corpus size**	**Retraining corpus**	**WER [%]**	
			RSC	**SSC**
enhanced mainASR (proposed method)	20 hrs -> 5.2 hrs	read+conv (20 hrs) + new mixed (5.2 hrs)	25.0	39.9
	50 hrs -> 12 hrs	read+conv (20 hrs) + new mixed (12 hrs)	24.3	38.8
	100hrs -> 22 hrs	read+conv (20 hrs) + new mixed (22 hrs)	23.5	37.9

TABLE III. USING VARIOUS AMOUNTS OF UNTRANSCRIBED DATA FOR THE ALTERNATIVE UNSUPERVISED TRAINING METHOD

ASR system	Untranscribed corpus size	Retraining corpus	WER [%]	
			RSC	SSC
enhanced mainASR (alternative method)	20 hrs -> 20 hrs	read+conv (20 hrs) + new mixed (20 hrs)	24.2	40.5
	50 hrs -> 50 hrs	read+conv (20 hrs) + new mixed (50 hrs)	25.2	40.6
	100hrs -> 100hrs	read+conv (20 hrs) + new mixed (100hrs)	24.6	40.2

are far from being correct. They were generated with the baseline main ASR (with a WER of 42.2% and a ChER of 23.3%).

For the maximum amount of untranscribed data used in this experiment (100 hrs) the relative WER improvement was 4.7% on conversational speech and 7.7% on read speech.

In the light of the above we can conclude that:

a) the proposed method is able to produce high-quality transcriptions (ChER < 5%) for about 22% to 25% of the initially untranscribed speech data,

b) the application of the proposed method leads to a more significant relative WER improvement than the alternative method and that

c) in the case of the proposed method, the main ASR system can still be improved if more than 100 hours of untranscribed speech can be acquired.

TABLE IV. THE SEED ASR SYSTEMS AFTER THE FIRST UNSUPERVISED TRAINING ITERATION

ASR system	Retraining corpus	WER [%]	
		RSC	SSC
seedASR #1 – iteration #1	read (10 hrs) + 11hrs	26.9	42.4
seedASR #2 – iteration #1	conv (10 hrs) + 11hrs	37.9	42.9

TABLE V. THE PERFORMANCE FIGURES OF THE MAIN ASR SYSTEMS

ASR system	Retraining corpus	WER [%]	
		RSC	SSC
mainASR – baseline	read+conv (20 hrs)	26.5	42.2
mainASR – iteration #1	read+conv (20 hrs) + 22 hrs	23.5	37.9
mainASR – iteration #2	read+conv (20 hrs) + 40 hrs	23.6	37.9

A. *Iterative unsupervised training*

Our last experiment aimed at finding out if reiterating the unsupervised training procedure on the same untranscribed corpus can bring further improvements or not.

The baseline seed ASR systems were retrained using the initial training corpus (10 hrs of read speech and respectively 10 hrs of conversational speech) plus half of the newly transcribed speech (see Figure 1), in this case half of 22 hours of new, mixed speech. The new seed ASR systems were evaluated (see Table IV) and the results show that important improvements were obtained for both read and conversational speech (compare with lines 2 and 3 in Table I).

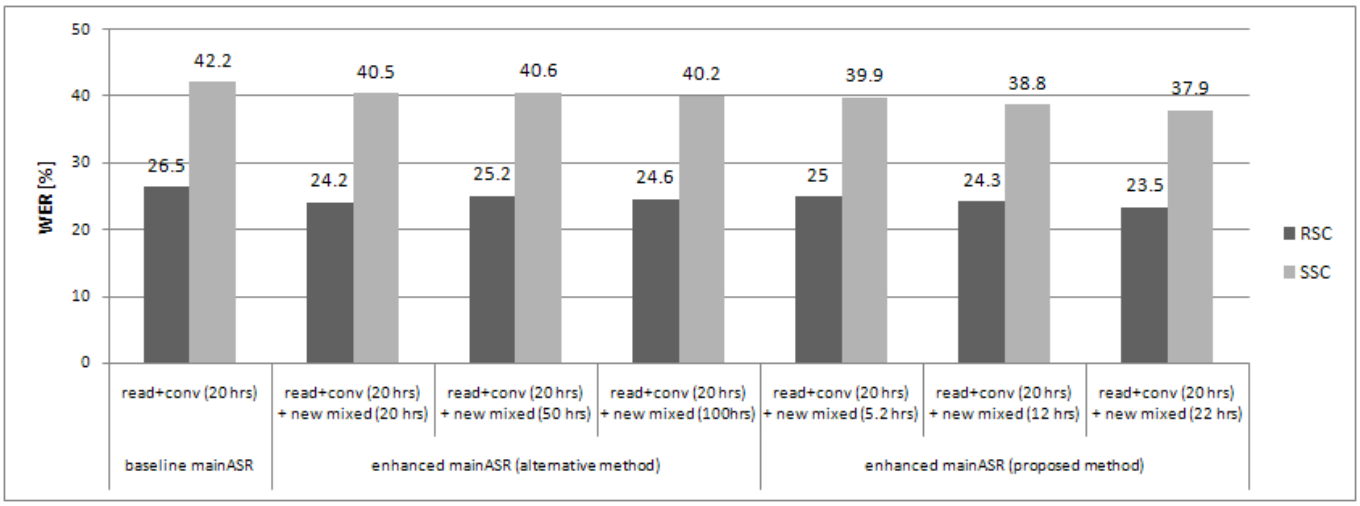

Figure 2. Comparison between baseline mainASR and enhanced mainASRs (alternative method vs proposed method)

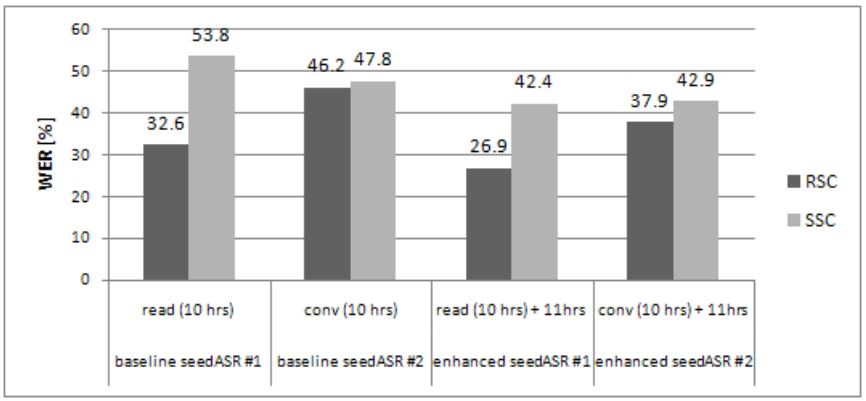

Figure 3. Comparison between baseline seedASRs and enhanced seedASRs

One could argue that after this first iteration the complementarity of the seeds is less obvious: the baseline seeds were trained on different types of speech (read vs. conversational), while the iteration #1 seeds are trained on read+mixed speech vs. conv+mixed speech. Consequently we rerun the first experiment (presented the beginning of this Section) to evaluate the quality of the transcriptions generated by the proposed method at the second iteration (using the iteration #1 seeds). There was no surprise to see that at iteration #2:

a) the seeds generate transcriptions for more speech data (because they are better than the baseline seeds);

b) the seeds generate transcriptions which contain more errors: WER is 14.6% and ChER is 7.1% (because they are less complementary than the baseline seeds).

In this context, when the 100 hrs untranscribed speech corpus was processed using the proposed method, at iteration #2, with the new seed ASR systems, we obtained 40 hrs of automatically transcribed speech (instead of 22 hrs at iteration #1) and a main ASR system with the performance figures listed in Table V. As one can see, after the second iteration, the main ASR performance figures (Table V) are very similar to those obtained after the first iteration (

Table II). We can conclude that the larger quantity of retraining data (40 hrs instead of 22 hrs) is compensated by its poorer quality (7.1% ChER instead of 4.9% ChER), leading to similar results. Consequently a second iteration on the same data seems useless.

For visualization convenience, the results in Tables I – IV are summarized in Fig. 2 and Fig. 3.

VI. Conclusions and Future Work

This study presented a method that can be successfully used to enhance a basic ASR system, initially created with very few acoustic resources for a new, under-resourced language. Provided that untranscribed speech data is widely available on the Internet, we showed that part of this data can be transcribed automatically and used to retrain the baseline ASR system, leading to significant improvements. We demonstrated that the proposed method creates transcriptions with a ChER of only 4.9% for about 25% of the untranscribed speed data. In this context, the experimental results illustrated that with 100 hours of untranscribed speech the baseline ASR system can be improved with over 10% relative WER. Higher improvements are expected if more untranscribed speech is used.

The unsupervised acoustic training method proposed in this paper was compared with an alternative method, which involves retraining the baseline ASR system using the raw transcriptions of all the speech data. In the near future we plan to compare our method with some of the confidence-based unsupervised training methods listed in Section II. In the near future we also plan to explore the possibility of using several seed ASR systems for improved data selection accuracy.

References

[1] L. Besacier, E. Barnard, A. Karpov, T. Schultz, "Automatic speech recognition for under-resourced languages: A survey.", in Speech Communication, Vol. 56 – Special Issue on Processing Under-Resourced Languages, pp. 85-100, http://dx.doi.org/10.1016/j.specom.2013.07.008

[2] H. Cucu, "Towards a speaker-independent, large-vocabulary continuous speech recognition system for Romanian", PhD Thesis, University "Politehnica" of Bucharest, 2011.

[3] A. Buzo, H. Cucu, C. Burileanu, "Text Spotting In Large Speech Databases For Under-Resourced Languages", in Proc. Int. Conf. Speech Technology and Human-Computer Dialogue (SpeD), Cluj-Napoca, Romania, 2013, pp. 77-82, http://dx.doi.org/10.1109/SpeD.2013.6682654

[4] H. Cucu, A. Buzo, C. Burileanu, "Unsupervised Acoustic Model Training using Multiple Seed ASR Systems", in Proc. Int. Workshop on Spoken Language Technologies for Under-resourced Languages (SLTU), St. Petersburg, Russia, 2014, pp. 124-130.

[5] G. Zavaliagkos, T. Colthurst, "Utilizing Untranscribed Training Data to Improve Performance", in DARPA Broadcast News Transcription and Understanding Workshop, Lansdowne, USA, 1998, pp. 301-305, http://dx.doi.org/10.1.1.27.5882

[6] T. Kemp and A. Waibel, "Unsupervised Training of a Speech Recognizer: Recent Experiments", in Proc. Eurospeech, Budapest, Hungary, 1999, pp. 2725–2728.

[7] F. Wessel and H. Ney, "Unsupervised training of acoustic models for large vocabulary continuous speech recognition", in Proc. Automatic Speech Recognition and Understanding Workshop (ASRU), Trento, Italy, 2001, pp. 307-310, http://dx.doi.org/10.1109/TSA.2004.838537

[8] L. Lamel, J.-L. Gauvain, G. Adda, "Lightly Supervised and Unsupervised Acoustic Model Training", in Computer Speech & Language, vol. 16, pp. 115-129, 2002. Available: http://dx.doi.org/10.1006/csla.2001.0186

[9] T. Fraga-Silva, J.-L. Gauvain, L. Lamel, "Lattice-based Unsupervised Acoustic Model Training", in Proc. Int. Conf. Acoustics, Speech and Signal Processing (ICASSP), Prague, Czech Republic, 2011, pp. 4656-4659, http://dx.doi.org/10.1109/ICASSP.2011.5947393

[10] L. Wang, M.J.F. Gales and P.C. Woodland, "Unsupervised training for mandarin broadcast news and conversational transcription", in Proc. Int. Conf. Acoustics, Speech and Signal Processing (ICASSP), Honolulu, Hawaii, 2007, vol. IV, pp. 353-356, http://dx.doi.org/10.1109/ICASSP.2007.366922

[11] J. Ma, S. Matsoukas., "Unsupervised training on a large amount of Arabic news broadcast data", in Proc. Int. Conf. Acoustics, Speech and Signal Processing (ICASSP), Hawaii, 2007, vol. II, pp. 349-352, http://dx.doi.org/10.1109/ICASSP.2007.366244

[12] K. Yu, M.J.F. Gales, L. Wang and P.C. Woodland, "Unsupervised training and directed manual transcription for LVCSR", in Speech Communication, Vol. 52, pp. 652–663, 2010. Available: http://dx.doi.org/10.1016/j.specom.2010.02.014

[13] J. Loof, C. Gollan, and H. Ney, "Cross-language Bootstrapping for Unsupervised Acoustic Model Training: Rapid Development of a Polish Speech Recognition System", in Proc. INTERSPEECH, Brighton, U.K., 2009, pp. 88-91.

[14] N.T. Vu, F. Kraus and T. Schultz, "Cross-language bootstrapping based on completely unsupervised training using multilingual A-stabil", in Proc. Int. Conf. Acoustics, Speech and Signal Processing (ICASSP), Prague, Czech Republic, 2011, pp. 5000-5003, http://dx.doi.org/10.1109/ICASSP.2011.5947479

[15] N.T. Vu, F. Kraus and T. Schultz, "Rapid building of an ASR system for Under-Resourced Languages based on Multilingual Unsupervised Training", In Proc. INTERSPEECH, Florence, Italy, 2011, pp. 3145-3148.

[16] N.T. Vu, F. Kraus and T. Schultz, "Multilingual A-stabil: A new confidence score for multilingual unsupervised training", in Spoken Language Technology Workshop (SLT), Berkeley, California, USA, 2010, pp. 183-188, http://dx.doi.org/10.1109/SLT.2010.5700848

[17] H. Cucu, A. Buzo, L. Petrică, D. Burileanu and C. Burileanu, "Recent Improvements of the SpeeD Romanian LVCSR System", in Proc. Int. Conf. on Communications (COMM), Bucharest, Romania, 2014, pp. 111-114, http://dx.doi.org/10.1109/ICComm.2014.6866659

[18] CMU Sphinx Toolkit: http://cmusphinx.sourceforge.net

[19] SRI-LM Toolkit: http://www-speech.sri.com/projects/srilm

[20] M. Rouvier, G. Dupuy, P. Gay, E. Khoury, T. Merlin, S. Meignier, "An Open-source State-of-the-art Toolbox for Broadcast News Diarization," in Proc. INTERSPEECH, Lyon, France, 2013.