UNSUPERVISED ACOUSTIC MODEL TRAINING USING MULTIPLE SEED ASR SYSTEMS

Horia Cucu, Andi Buzo and Corneliu Burileanu

Speech and Dialogue (SpeeD) Research Laboratory,
University “Politehnica” of Bucharest, Romania
{horia.cucu, andi.buzo, corneliu.burileanu}@upb.ro

ABSTRACT

Unsupervised acoustic modeling can offer a cost and time effective way of creating a solid acoustic model for any under-resourced language. This paper explores the novel idea of using two independent ASR systems to transcribe new speech data, align and filter the ASR hypotheses and use the presumably correct transcriptions to iteratively improve the two seed ASR systems. In parallel, the newly transcribed speech is used to retrain the mainstream ASR system. The methodology leads to WER relative improvements of 5.5% after the first iteration. The experiments are made with data in the Romanian language.

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

1. 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 recently proposed to overcome this crucial issue in developing ASR systems for under-resourced languages. Among them, lightly-supervised and unsupervised AM training seem to be the most successful methods in bootstrapping an acoustic model for a new language. These methods were initially proposed and are effective for high-resourced languages also, but their application to under-resourced languages is even more valuable given the more significant lack of annotated acoustic data for the latter type of languages.

The basic idea of both lightly-supervised and unsupervised training techniques is to use an initial, seed acoustic model (trained with manual transcribed speech data) to generate transcriptions for a large quantity of untranscribed speech data and then use this data for further retraining. This idea is based on the highly likely assumption that untranscribed speech data can be obtained much easier, regardless of the language. In lightly-supervised acoustic modeling this untranscribed speech data needs to be accompanied by approximate or loose transcriptions (which are also easier to obtain than correct, manual transcriptions).

In the case of lightly-supervised training, (a) the initial acoustic model generates transcriptions for the loosely-transcribed speech data, (b) these automatically generated transcriptions are aligned with the loose transcriptions, (c) a part of the transcriptions are selected to be correct based on some confidence score for the alignment and (d) the selected data is finally used in further AM retraining. The process can be repeated until no new data can be selected for further retraining.

The general procedure for unsupervised acoustic model training starts similarly by using a seed AM 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 AM retraining. Again 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. Note that in this case loose-transcriptions are not available, therefore the confidence scoring must be done on the ASR output alone.

While in previous works we focused on ASR domain adaptation [1] and lightly-supervised acoustic modeling [2] for under-resourced languages, in this paper we explore unsupervised acoustic modeling and introduce a novel method of selecting the correctly transcribed data. The idea is to use two complementary seed ASR systems to generate two sets of ASR hypotheses and then align these transcriptions with Dyanmic Time Warping (DTW) based algorithms. The method assumes that complementary ASR systems will make uncorrelated errors and the aligned parts of the transcriptions can be considered correct (the probability that two independent systems make the same identical errors is very small.). Going further, the selected (correct) data is split into two distinct parts and each part is used to further retrain one of the seed acoustic models. The process can be repeated until the amount of newly selected data is not significant anymore. In this study, we validate the proposed methodology on a Romanian broadcast news and conversations speech corpus, acquired automatically from the Internet. The method presents some clear advantages such as (i) the obtained annotated speech is accurate, (ii) no restriction is imposed to the audio data and (iii) new kind of speech can be obtained, e.g. elder speech, whisper, dialect, etc.

The rest of the paper is organized as follows. In Section 2 we discuss the state-of-the-art in unsupervised acoustic model training and point out the main novelties of our study. In Section 3 we describe in detail the proposed method and the specific issues encountered. In sections 4 and 5 we present the experimental setup and results and finally, in Section 6 we draw some conclusions.

2. RELATED WORK AND NOVEL KEY FACTORS

The first tentative to train an acoustic model in an unsupervised fashion was presented in [3]. In this study, the authors used a Spanish ASR system, trained with a very small amount of data (3 hrs of transcribed speech), to decode 25 hrs of untranscribed speech. Afterwards, using confidence scoring and threshold optimization, they were able to select 2.7 hrs of the ASR output for further retraining and obtained an improvement of 1.7% relative WER over the initial ASR system. A similar procedure using a different, lattice-based confidence score is presented in [4]. In this paper the authors applied the unsupervised acoustic model training technique to create a German ASR system and they report much better results: 34% relative WER improvement over the initial ASR.

In [5] the authors extended the series of experiments presented in [4] by exploring 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. In their study, the authors also investigate 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, the authors also explore 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.

In [6] the authors investigate unsupervised AM training with different sized language models. The training procedure is a little different in the sense that more untranscribed data is added iteratively (the seed models are used to transcribe only a small part of the untranscribed data and generate better models; these are used to transcribe a double amount of untranscribed data and generate better models and so on). The conclusion is 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.

Although initially the unsupervised acoustic modeling procedure was applied in the context of Maximum Likelihood (ML) training, several studies also investigated its usability for Maximum Mutual Information (MMI) training [7, 8] and Minimum Phone Error (MPE) training [7, 9]. In [9] 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 new languages such as Mandarin [7, 9], Arabic [8], Polish [10], Czech [11] and Vietnamese [12].

An innovative idea, recently introduced in [12], 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 [13]. Finally, they train a Vietnamese ASR system using the resulted transcriptions. In [11] 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 go beyond the state of the art by exploring the idea of using *two complementary ASR systems* for Romanian to transcribe new Romanian speech data, align and filter the ASR hypotheses and use the matched words from the transcriptions alignment to *iteratively improve the two seed ASR systems*. The novelty of the proposed unsupervised training methodology comprises several key factors:

a) the unsupervised training process starts with two seed ASR systems;

b) ASR hypotheses filtering is done based on a phrase-level similarity confidence score;

c) the selected correct transcriptions are used to improve both the seed models.

d) there is no need to re-tune the system every time a new kind of speech recording is used, i.e. new recording conditions, elder speech, dialects, whispers, etc.

The usage of several seed ASR systems in unsupervised training was also explored in [12], but in that study the seed models were for different languages and the iterative adaptation procedure of these models also differs.

The ASR hypotheses filtering idea is totally 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 use a similarity confidence score applied at phrase (multi-word) level on the output of two ASR systems.

Finally, although we also adopt the iterative training procedure, the new data obtained after each iteration are split into two parts and each part is used to retrain a single seed ASR system, trying to preserve their complementarity.

3. 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). Generally, this ASR system, further called *main ASR system*, would have been trained with all the annotated speech data available and would use the best language model available. The method has five steps and it is illustrated in Figure 1.

The first step is to create two complementary ASR systems, which can be further used to process untranscribed speech data. These complementary ASR systems will 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 can be considered correct).

To create complementary ASR systems, the initial training speech database can be split into two parts based on the type of speech, acoustic environment, etc. Another aspect that could help in obtaining complementary ASR systems is language model selection. One could start with seed ASR systems using the same LM, or different LMs. Although the usage of different LMs seems like the most natural choice, the experiments showed that choosing the same, best language model for both seed ASR systems leads to better results (see Section 5).

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 (music, jingles, advertisements, etc.) parts 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 of words).

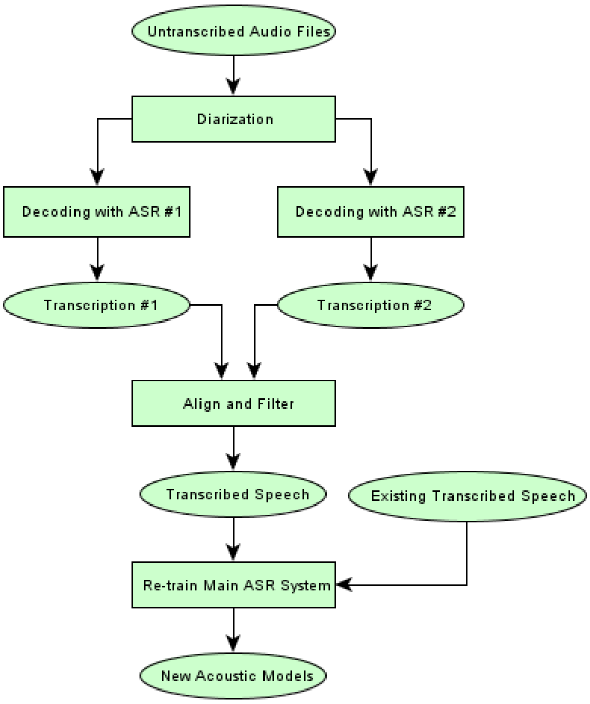

Figure 1. The block diagram of the proposed method

In step 3, the clean, un-transcribed speech data is decoded using the two seed ASR systems and the resulted sets of ASR hypotheses are aligned using a DTW algorithm.

In step 4, the matched parts of the ASR hypotheses resulted from the alignment are selected, together with the corresponding audio data, to create a new annotated speech corpus. For this, we use a similar confidence score as the one introduced in [2], i.e. sequences of consecutive aligned words are considered to be correctly recognized if the number of characters forming these sequences exceed an empirically determined threshold. In addition, the aligned sequence duration has to be at least one second for the sequence to be taken into account. Moreover, the time difference between two consecutive words should not exceed two seconds (because this could mean that the audio file contains untranscribed non-speech fillers).

This selection procedure increases the likelihood that selected data is correct, because it ignores singular short words and even short sequences of short words, which can be very common in a language, and it assures that all words are part of the same utterance. After the selection of the correctly aligned sequences of words is done, their timestamps are used to cut the corresponding audio parts out of the initial speech files.

As described above, the confidence score used in this method is applicable at phrase-level and the resulted speech corpus will be composed of utterances of at least a few words, as opposed to single words as in the case of the selection procedures discussed in Section 2.

Step 5 involves retraining the seed acoustic models with the newly created annotated speech data. The complementarity of the retrained systems has to be preserved, because the ASR systems should be used in a new iteration of the whole process. Therefore, the newly created speech data has to be split into two distinct parts and each part should be used to augment the initial training data for one of the ASR systems. In other words, both seed ASR systems are retrained using the initial speech data plus a part of the newly created speech data. These systems will still be complementary, as each of them will be adapted to better recognize a part of the untranscribed speech data.

Using the enhanced ASR systems, the unsupervised training procedure can be restarted at step 3. Steps 3, 4 and 5 can be further reiterated until the amount of new data selected at step 4 does not significantly differ from the amount of data selected at the previous iteration.

4. EXPERIMENTAL SETUP

4.1. Speech corpora

The various Romanian speech corpora used in the experiments are listed in Table 1. All these corpora are created by the Speech and Dialogue (SpeeD) research group[1]. Note that the experiments presented in this paper are done on self-developed corpora because for the Romanian language there are no other speech corpora available for research.

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 [14; 15].

The SSC-train (Spontaneous Speech Corpus) and SSC-eval corpora were created using a lightly-supervised acoustic modeling technique [2]. 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.).

Table 1. Romanian speech corpora

Corpus name	Type of speech	Size	#Speakers
RSC-train	read speech	100 hrs	157
RSC-eval		6 hrs	22
SSC-train	conversational speech	27.5 hrs	unknown
SSC-eval		3.5 hrs	unknown
SSC-untranscribed	conversational + read speech	200 hrs	unknown

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 using the LIUM Speaker Diarization Toolkit [16]. The segmentation and diarization processes aimed to filter-out all the non-speech parts of the corpus and to create single-speaker utterances. The SSC-untranscribed corpus will be further used in the unsupervised training procedure.

4.2 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 (32/64/128) 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[2] Toolkit) with the various training speech corpora listed in Table 1.

4.3 Language models

Two tri-gram, closed-vocabulary, language models, previously created with the SRI-LM Toolkit[3], were used in this study: LM #1 and LM #2. LM #1 is a general language model trained with several online-newspaper text corpora (with a total 169M words). LM #2 is another general language model trained using the same online-newspaper text corpora, but biased towards broadcast conversational speech using a text corpus of 40M words. LM #2 was

[1] Speech and Dialogue Research Group: http://speed.pub.ro

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

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

obtained through language model interpolation, using a weight of 90% for the broadcast conversational speech. For both LMs the number of unigrams was limited to the most frequent 64k (this constraint was imposed by the Sphinx4 ASR decoder). The two language models have different sizes and vocabularies as they were created with different text corpora. The language models performance figures (perplexity –PPL and out-of-vocabulary rate – OOV rate) obtained on the transcriptions of the two evaluation speech corpora are listed in Table 2.

Table 2. LMs performance figures on transcriptions

	PPL		OOV rate [%]	
Language model	**RSC-eval**	**SSC-eval**	**RSC-eval**	**SSC-eval**
LM #1	216.4	176.9	2.08	3.13
LM #2	201.3	150.6	1.86	3.07

5. EXPERIMENTAL RESULTS

5.1 Baseline ASR and seed ASR systems

The baseline for all further experiments is the best Romanian ASR system developed so far in our research group (further called *mainASR*). This system uses an acoustic model trained on RSC-train + SSC-train and LM #2. Its word error rates (WERs) on the two evaluation corpora are listed in Table 3. As expected, the WER on read speech is much lower than the WER on conversational speech, not only because read speech is easier to recognize, but also because the RSC-train is larger than SSC-train.

The two seed ASR systems (created for the first decoding iteration) were trained on RSC-train and SSC-train, respectively. These seed ASR systems are also presented in Table 3. For seed ASR system #1 we created two versions: one uses LM #1 (the general LM trained on newspaper text) and the other uses LM #2 (the general LM biased towards broadcast conversational speech). Version #1 has the advantage of increased complementarity with seed #2 ASR, while version #2 benefits from a better language model and might recognize correctly more untranscribed speech. Section 5.2 will show which of the two versions of seed #1 will be further used.

One thing that is worth noting about the values presented in Table 3 is that the RSC-train corpus does not help too much in conversational speech recognition (the baseline system and seed #2 have similar WERs on SSC-eval). However, this corpus makes a huge difference when it comes to read speech recognition (the baseline system has a significantly lower WER than seed #2 on RSC-eval). This means that an ASR system cannot perform well with both RSC or SSC if it is trained only with one type of data.

Table 3. Baseline and seed ASR systems

ASR system	**Acoustic model (training corpus)**	**LM**	**WER [%]**	
			RSC-eval	SSC-eval
mainASR baseline	RSC-train + SSC-train (127 hrs)	LM #2	16.1	38.6
seed #1 v1	RSC-train (100 hrs)	LM #1	18.0	47.1
seed #1 v2	RSC-train (100 hrs)	LM #2	17.1	46.0
seed #2	SSC-train (27 hrs)	LM #2	36.1	39.9

5.2 Language model selection effect

The SSC-untranscribed corpus was processed in parallel with the three seed ASR systems (seed #1 v1, seed #1 v2 and seed #2). The resulted ASR hypotheses were aligned and the assumed correct speech data was selected exactly as described in Section 3. Out of the 200 hours of speech in the SSC-untranscribed corpus, we were able to select 60 hours by using version 1 of the seed #1 ASR, and 64 hours by using version 2 of the seed #1 ASR.

This new speech data was used to retrain the seed #2 ASR system. The performance of the resulted ASR systems is presented in Table 4. The results show us that using the better language model in the seed #1 ASR system generates more transcribed speech, which helps in improving an ASR system (for both read and conversational speech). Consequently, we used LM #2 in all the following experiments for both seed ASR systems.

Table 4. Language model selection effect

Acoustic model (training corpus)	**WER [%]**	
	RSC-eval	SSC-eval
SSC-train (27 hrs)	36.1	39.9
SSC-train (27 hrs) + 60 hrs*	27.9	37.1
SSC-train (27 hrs) + 64 hrs**	27.7	36.7

* the 60 hrs of speech were obtained by aligning the output of seed #1 v1 and seed #2;

** the 64 hrs of speech were obtained by aligning the output of seed #1 v2 and seed #2;

5.3 Iterative unsupervised training

In this experiment the seed ASR systems were iteratively used to decode the SSC-untranscribed corpus and retrained using the original training corpus and half of the automatically generated speech data. In Table 5 we present the first two iterations of the unsupervised training process because after the second iteration no further improvements were reported.

After the first retraining iteration, seed #1 ASR showed a significant improvement on conversational speech (16.5% relative WER). It is an expected result because this ASR system was initially trained only on read speech and in iteration #1 the read speech training corpus is augmented with 30 hrs of read + conversational speech.

Seed #2 ASR improvement is less significant on conversational speech (5.5% relative WER) because this system was initially trained on the SSC-train corpus. However, seed #2 ASR gets much better at recognizing read speech: 15.5% relative WER improvement.

The *mainASR* system exhibits divergent performance figures after iteration #1, when it is trained with the initial 127 hrs of speech (in RSC-train and SSC-train) plus the additional 64 hrs of speech obtained in an unsupervised fashion. The WER on read speech is slightly worse probably because the new acoustic model is less adapted to read speech. We do not consider this a worse result, but rather a more objective result, i.e. the system has gained in generality. On conversational speech the *mainASR* obtains an improvement of 5.5% relative WER.

Another iteration (#2) was performed in order to see if more annotated speech could be extracted from this corpus. The experiment showed that the amount of the extracted annotated speech is bigger (98 hours versus 64 hours obtained in the first iteration), however, the accuracy of the *mainASR* after iteration #2 was not improved. This means that in order to improve the performance either more speech needs to be acquired, or other type of speech (other than broadcast news) must be used.

The poor results obtained after iteration #2 might be explained by the following: even though more automatically annotated data is produced, its quality might be lower, because the ASR seeds are probably less complementary (after the first iteration they were trained with both read and conversational speech). This hypothesis will be verified in a future work.

While the WER reduction obtained by this method depends on the current development state of a given ASR system, by the amount of untranscribed speech data used, the characteristics of the test speech data, etc., it is unquestionable the benefit of easily obtaining annotated speech data from raw untranscribed speech. This method enables the acquisition of new type of speech data such as dialects, whispers, elder speech, etc., which are in general very difficult to acquire. The quantity of the annotated speech data obtained by this method is proportional to the quantity of the untranscribed speech collected.

5.4 Comparison with other unsupervised methods

An alternative unsupervised acoustic training scenario is the one in which the *mainASR* system is used to decode the SSC-untranscribed speech corpus and all the raw transcriptions are used to retrain this *mainASR* system.

The results obtained for this experiment showed that the "enhanced" *mainASR* (trained with RSC-train + SSC-train + raw transcriptions) does not obtain any WER reductions over the initial one. These results are in line with the findings reported in [5].

In the near future we also plan to compare our method with some of the confidence-based unsupervised training methods listed in Section 2.

6. CONCLUSIONS AND FUTURE WORK

We presented a novel method of unsupervised acoustic model training. The method aims at obtaining annotated speech data from raw untranscribed audio data. The transcription is obtained by aligning the output of two complementary ASR systems. The complementarity of the ASR systems assures that the errors produced by them are uncorrelated. Hence, the aligned parts of the transcriptions can be considered correct transcriptions. The newly obtained transcribed speech together with the existing annotated speech data are used to re-train a new acoustic model. This procedure is repeated until no improvement in ASR accuracy is obtained.

The experimental results have shown that the method is able to annotate 32% of the untranscribed speech data at the first iteration and 49% at the second iteration. After the first iteration a relative reduction of WER by 5% is obtained. The second iteration brought no WER reduction despite the amount of annotated data is higher meaning that more untranscribed speech needs to be used in order to further reduce WER.

The method leads to new opportunities for annotating new type of speech such as dialects, whispers, elder speech, etc., which are not easy to find. The amount of annotated speech is proportional to the quantity of the untranscribed speech data used.

Table 5. ASR improvements over several retraining iterations

ASR system	**Acoustic model (training corpus)**	**WER [%]**	
		RSC-eval	SSC-eval
mainASR baseline	RSC-train + SSC-train (127 hrs)	16.1	38.6
mainASR iteration #1	RSC-train + SSC-train (127 hrs) + 64 hrs	16.3	36.5
mainASR iteration #2	RSC-train + SSC-train (127 hrs) + 98 hrs	16.7	36.6
seed #1	RSC-train (100 hrs)	17.1	46.0
seed #1 iteration #1	RSC-train (100 hrs) + 30 hrs	15.8	38.4
seed #1 iteration #2	RSC-train (100 hrs) + 48 hrs	17.1	38.6
seed #2	SSC-train (27 hrs)	36.1	39.9
seed #2 iteration #1	SSC-train (27 hrs) + 34 hrs	30.5	37.7
seed #2 iteration #2	SSC-train (27 hrs) + 50 hrs	31.5	38.5

In the near future we plan to evaluate the data selection method (the quality of the automatically generated transcriptions) by using a reference, transcribed speech corpus as if it was an untranscribed speech corpus. Moreover, we plan to compare our method with some of the confidence-based unsupervised training methods listed in Section 2 and to explore the possibility of using several seed ASR systems for improved data selection accuracy.

ACKNOWLEDGEMENTS

This study has been partially developed with the financial support of the Romanian-American Foundation. The opinions, findings and conclusions or recommendations expressed herein are those of the authors and do not necessarily reflect those of the Romanian-American Foundation.

7. REFERENCES

[1] H. Cucu, A. Buzo, L. Besacier, C. Burileanu, "SMT-based ASR Domain Adaptation Methods for Under-Resourced Languages: Application to Romanian", in Speech Communication, Vol. 56 – Special Issue on Processing Under-Resourced Languages, pp. 195-212, 2014.

[2] 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.

[3] 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.

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

[5] 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.

[6] L. Lamel, J.-L. Gauvain, G. Adda, "Lightly Supervised and Unsupervised Acoustic Model Training", in Computer Speech & Language, vol. 16, pp. 115-129, 2002.

[7] 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.

[8] J. Ma, R. Schwartz., "Unsupervised training on a large amount of Arabic news broadcast data", in Proc. Int. Conf. Acoustics, Speech and Signal Processing (ICASSP), Honolulu, Hawaii, 2007, vol. II, pp. 349-352.

[9] 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.

[10] 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, pages 88-91.

[11] 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.

[12] 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.

[13] 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.

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

[15] 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 (submitted paper).

[16] 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.