Recent Improvements of the SpeeD Romanian LVCSR System

Horia Cucu*, Andi Buzo, Lucian Petrică, Dragoş Burileanu and Corneliu Burileanu
University Politehnica of Bucharest, Speech and Dialogue Research Laboratory, Bucharest, Romania
*Corresponding author (e-mail: horia.cucu@upb.ro)

***Abstract*—This paper presents the main improvements brought recently to the SpeeD automatic speech recognition system. Several aspects, such as speech and text resources acquisition, noise-robust speech features and feature transforms are discussed. All the updates in our ASR system are accompanied by experimental results illustrating significant improvements: between 30% and 35% relative WER reductions for various case studies (read/spontaneous speech, noisy/clean speech). In the last part of the paper, our ASR system is also compared with Google's ASR system and a brief analysis of the results is presented.**

***Keywords*-ASR; LVCSR; noise-robust speech recognition; PNCC; LDA.**

I. INTRODUCTION

Large vocabulary continuous speech recognition (LVCSR) is still an unsolved topic for many languages. The reasons for this are (i) there is a lack of acoustic and linguistic resources needed for development (it is the case of so-called under-resourced languages) and (ii) the scientific research community is not stimulated by any national or international evaluation campaigns (as opposed to languages such as English, French or Chinese). The Romanian language is affected by both the aforementioned problems. In this context, the development of speech and language resources for automatic speech recognition (ASR) is a critical issue that must be addressed to push forward the research in this direction and create LVCSR systems comparable to those available for other languages. This is one of the main goals of the Speech and Dialogue (SpeeD) research group[1].

To the best of our knowledge, at the moment there are three LVCSR systems developed for the Romanian language. In 2011 we published the first LVCSR results for Romanian [1, 2], in August 2012 Google launched their online speech recognizer[2] for Android and Chrome and in December 2012 THINKTech Research Center[3] also published a paper [3] on broadcast news recognition for Romanian.

The goal of this paper is to present the main improvements brought to SpeeD's LVCSR system since 2011 and thoroughly evaluate the current version of the system. In Section II, we present the new ASR resources (speech and text corpora) that were recently collected and used in acoustic and language modeling. Section III deals with the novelties introduced in the ASR front-end (noise-robust speech features and feature transforms), which aim at providing a more robust phoneme-level modeling. In Section IV, we briefly describe the experimental setup for the large number of experimental results presented in Section V. Finally, Section VI is dedicated to the closing conclusions.

II. NEW ASR RESOURCES

For under-resourced languages, the most important obstacle in obtaining good ASR results is the lack of proper speech and language resources. Consequently, during the previous years, significant efforts were made to extend the size of the training read speech corpus and to create a spontaneous speech corpus. The text corpus was also extended by collecting and pre-processing new text data from the Internet.

A. Speech Corpus Extensions

In [1] we presented the (Romanian) continuous speech corpora available for training and evaluation in 2011. Since then we continued with the acquisition of acoustic data, extending the read speech corpus and creating a spontaneous speech corpus.

The read speech corpus was extended by recording various predefined texts representing news articles and literature. The recordings were made using an online recording application. The speakers were involved voluntarily in the speech corpus extension project and were mostly university students. Table I presents the read speech corpus (further called *RSC*), underlining its various components (the parts in bold were collected between 2011 and 2013).

As one can infer from Table I, the RSC corpus sums up to a total of 106 hours of speech and contains speech uttered by 165 different speakers. For the purpose of ASR development the RSC corpus was split into two parts: the *RSC-train* (to be used for ASR training) and the *RSC-eval* (to be used for ASR evaluation). The RSC-train corpus consists of 145k utterances (WORDS, 90% of CS_01 and CS_06), spoken by 157 different speakers, summing up to a total of 100 hours of speech. The RSC-eval corpus consists of 2,604 read utterances (10% of CS_01, CS_04 and CS_05), spoken by 22 different speakers, summing up to a total of 5.5 hours of speech. The RSC-eval corpus is freely available for download on SpeeD's website.

Starting from approximate transcriptions of broadcast news and talk shows collected over the Internet and using lightly supervised ASR training, we have also recently developed a spontaneous speech corpus. The methodology used in the acquisition is thoroughly described in [4]. The resulted spontaneous speech corpus consists of approximately 54k utterances and sums up to a total of 27.5 hours of speech. This corpus is further used for ASR training and is called *SSC-train*. A part of the speech data collected was manually annotated to create an error-free spontaneous speech corpus for evaluation only. This part is further called *SSC-eval* and 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 I. THE READ SPEECH CORPUS (NEWLY COLLECTED PARTS IN BOLD)

Corpus Name	Domain	Utterances	Hours of Speech	Speakers
WORDS	n/a	110k	42	17
CS_01	news, interviews	12.2k	22	12
CS_04		900	2.1	9
CS_05		**504**	**1.1**	**7**
CS_06	**literature**	**23.8k**	**38.1**	**155**

[1] Speech and Dialogue Research Group: http://speed.pub.ro
[2] Google ASR System: http://officialandroid.blogspot.ro/2012/08
[3] THINKTech Research Center: http://thinktech.hu

978-1-4799-2385-4/14/$31.00 ©2014 IEEE

B. Text Corpus Extensions

Regarding the acquisition of text corpora for language modeling, the text corpus available in 2011, called *europarl + 9am + hotnews* and described in [1] was extended by collecting and pre-processing new text data from the Internet. The *europarl + 9am + hotnews* corpus comprised online news, had about 169M words and required in-depth pre-processing (including diacritics restoration) before it could be used for language modeling. In 2012, we collected another text corpus composed of meeting and discussion transcriptions, with a size of about 40M words. In this new corpus, further called *meetings*, the diacritics do not need to be restored, as the text contains correct diacritical words.

III. New Speech Features

A. Noise Robust Speech Features

In recent years, the ASR noise robustness has been addressed by many speech research groups and by different approaches: speech enhancement, noise robust features, model compensation, etc. Among all these techniques, the noise robust features introduced in [5] and called Power Normalized Cepstral Coefficients (PNCCs) tend to bring the most important gains in accuracy. Major new distinctive attributes of PNCC processing include:

a) the use of a power-law nonlinearity that replaces the traditional log nonlinearity used in Mel Frequency Cepstral Coefficients (MFCC),
b) a noise-suppression algorithm based on asymmetric filtering that suppresses background excitation, and
c) a module that accomplishes temporal masking.

Moreover, PNCCs use medium-time power analysis, in which environmental parameters are estimated over a longer duration than is commonly used for speech, as well as frequency smoothing.

We experimented with these new features as replacements for the traditional MFCCs and obtained better results for all the various types of noisy speech. We therefore decided to use PNCCs by default in our LVCSR system. In 2013, a variation of these features was implemented in the CMU Sphinx[4] speech recognition toolkit, which is the core of our LVCSR system. Our research group also contributed to the integration of the features in the Sphinx4 Java decoder.

B. Feature Transforms (LDA+MLLT)

State-of-the-art LVCSR systems use cepstral features augmented with dynamic information from the adjacent speech frames (temporal derivates of the features). The standard MFCC + Δ + ΔΔ scheme, while performing relatively well in practice, has no real basis of existence from a discriminant analysis point of view. The same argument applies for the computation of the cepstral coefficients from the spectral features: it is not clear that the discrete cosine transform, among all linear transformations, has the best discriminatory properties even if its use is motivated by orthogonality considerations [6].

This was the argument for introducing the linear discriminant analysis (LDA) technique in the ASR front-end computation [7]. LDA is a linear projection transformation that maps the vector obtained by concatenating several feature vectors to a lower-dimensional space, with the goal of maximally separating the phonetic classes in the transformed space.

One problem with the LDA transformation is that it cannot cope with the diagonal modeling assumption (that is imposed on the Gaussian models in most ASR systems): if the dimensions of the projected subspace are highly correlated then a diagonal covariance modeling constraint will result in distributions with large overlap and low sample likelihood. In this case, a maximum likelihood linear transformation [8] (MLLT), which aims at minimizing the loss in likelihood between full and diagonal covariance models is known to be very effective.

The two feature transforms that were briefly presented in this section were recently introduced in our LVCSR system.

IV. Experimental Setup

A. Acoustic Models

For this study several 5-state HMM-based acoustic models (*AM001*, *AM002* and *AM005*) were developed. In all cases the 36 phonemes in Romanian were modeled contextually (context dependent phonemes) with 4,000 HMM senones. The number of Gaussian mixtures (GMs) per senone state was varied to adapt the acoustic model setup to the size and variability of the training speech database. The last column in Table II lists the maximum number of GMs that were trained and can be further used for each acoustic model. The acoustic models were created and optimized with the training speech corpora described in Section II.A: RSC-train and SSC-train (a total of 127 hours of speech). AM002 was trained using the traditional speech features plus temporal derivates (13 MFCC + Δ + ΔΔ) and AM001 was trained on the noise robust speech features (13 PNCCs + Δ + ΔΔ). The noise robust speech features without temporal derivates were used to train AM005. In this case the feature transforms described in Section III.B were also used. The features of the three acoustic models are presented in Table II.

The newly developed acoustic models were compared with a baseline acoustic model, called *AM000*. This baseline system is the one presented in [1] and was trained with 54 hours of read speech (a part of the WORDS corpus and a part of CS_01 corpus). The features of this acoustic model are also listed in Table II.

B. Language Models

Two tri-gram, closed-vocabulary, language models were created (using the SRI-LM Toolkit[5]) and compared in this study: *LM000* and *LM001*. LM000 is the baseline language model described in [1] and was trained using the *europarl + 9am + hotnews* text corpus (169M words). LM001 was obtained by the interpolation of LM000 with a language model trained on the *meetings* text corpus (40M words). The interpolation was done with the weights 0.1 for LM001 and 0.9 for the LM000. The interpolation weights tuning has not been considered for the moment. For both LMs the number of unigrams was limited to the most frequent 64k, due to an ASR decoder implementation limitation.

C. Speech Corpora for the Noise Robustness Experiments

The effect of using noise-robust features instead of the traditional MFCCs was assessed on variations of the RSC-eval and SSC-eval speech corpora. The reason is that the RSC-eval corpus does not contain any noise (it comprises clean recordings) and the SSC-eval

TABLE II. The Acoustic Models

Name	Training Corpus	Speech Features	# HMM Senones	# GMs
AM000	parts of WORDS and CS_01 (54 hours)	MFCCs + Δ + ΔΔ	4,000	16
AM001	RSC-train + SSC-train (127 hours)	PNCCs + Δ + ΔΔ	4,000	128
AM002	RSC-train + SSC-train (127 hours)	MFCCs + Δ + ΔΔ	4,000	128
AM005	RSC-train + SSC-train (127 hours)	PNCCs + LDA-MLLT	4,000	128

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

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

corpus contains both clean speech and speech in degraded conditions (background noise, background music, telephone speech, etc.).

In order to evaluate the ASR noise robustness in various noise conditions, several types of noise (street, babble and subway) were added digitally, at different signal-to-noise ratios (SNRs), over the clean speech in RSC-eval. Consequently, Section V.C presents the results obtained by the various ASR systems on the corpora: RSC-eval + street noise (5/10/15/20dB), RSC-eval + babble noise (5/10/15/20dB) and RSC-eval + subway noise (5/10/15/20dB).

The speech conditions in the SSC-eval corpus are annotated at phrase level and this allowed us to split the corpus into two parts: clean speech and noisy speech. In Section V.C, we assess the various ASR systems on the SSC-eval corpus as a whole, but also separately on its clean part and on its noisy part. This experiment aims to evaluate the ASR noise robustness in real world noise conditions.

V. Experimental Results

In all the subsequent experiments, the acoustic models were compared in terms of speech recognition word error rate (WER) on the two evaluation speech corpora: RSC-eval and SSC-eval.

A. Optimal Number of GMs for the New Acoustic Models

Whenever the training speech corpus is extended, the acoustic models core features (number of HMM senones and number of Gaussian mixtures per senone) must be re-evaluated. Based on the experience obtained in previous experiments, we decided to train acoustic models with 4,000 senones for all the experiments presented in this paper. The number of GMs per senone was varied in order to find the best setup. The experimental results for AM002 are presented in Table III. Note that for these experiments we used LM001.

As the results in Table III show, there is an optimal number of Gaussian densities per HMM senone. This was expected, because with a particular speech corpus (with its particular size and variability) there is an optimal number of AM parameters that can be efficiently trained. For our training speech corpus (RSC-train + SSC-train) the optimal number of GMMs per HMM state is 64. Similar results were obtained for AM001 and AM005 (all models which were trained on the same speech corpus).

B. The Effect of Extending the ASR Resources

The extension of ASR resources (both speech and text) brought an overall relative WER reduction of 30% on the read speech corpus and 33% on the spontaneous speech corpus (see line 1 vs. 4 in Table IV). The effect of extending the speech corpus is much more important (see line 1 vs. 3) than the effect of extending the text corpus (see lines 1 vs. 2 and 3 vs. 4). This can be partially explained by the fact that the training speech corpus was extended with more than 130%, while the text corpus was extended with less than 25%.

Another important conclusion that results from this experiment is that resource acquisition is still a critical direction which we should follow to significantly improve our system's accuracy (it is even more critical than implementing/using the state-of-the-art ASR techniques).

TABLE III. THE OPTIMAL NUMBER OF GAUSSIAN DENSITIES PER SENONE

		WER [%]	
Acoustic Model	**# GMs**	**RSC-eval**	**SSC-eval**
AM002	16	17.6	41.4
	32	16.8	40.3
	64	**16.7**	**39.4**
	128	17.6	39.7

TABLE IV. ASR IMPROVEMENTS BROUGHT BY THE NEW SPEECH AND TEXT CORPORA

		WER [%]	
Acoustic Model	**Language Model**	**RSC-eval**	**SSC-eval**
AM000	LM000	23.8	58.3
	LM001	22.9	57.6
AM002	LM000	17.6	40.5
	LM001	16.7	39.4

C. The Effect of Using Noise Robust Features

Figure 1 illustrates the results obtained by the MFCC-based acoustic model (AM002) and the PNCC-based acoustic model (AM001) on the RSC-eval corpora with digitally added noise. The figure shows that regardless of the type of noise and regardless of the SNR, the PNCC-based acoustic model is better than the MFCC-based acoustic model. Even for clean speech the PNCC-based acoustic model obtains a slightly better WER. The relative WER reduction is, in average, 10% for an SNR of 20dB, 16% for an SNR of 15dB, 19% for an SNR of 10dB, and 13% for an SNR of 5dB.

Table V presents the results obtained by the two acoustic models on speech with real-world noise (line 1 vs. 2). The relative WER reduction on noisy speech only (SSC-eval noisy) is smaller (4%) than in the previous experiment. Another interesting fact is that for clean speech (SSC-eval clean) the noise-robust acoustic model (AM001) obtains a slightly worse WER than the MFCC-based acoustic model. Nevertheless, the PNCC-based acoustic model is overall better than the MFCC-based acoustic model on spontaneous speech (SSC-eval all).

D. The Effect of Using Feature Transforms

Table V also presents some preliminary results obtained with an acoustic model (AM005) trained with discriminative PNCC features (processed using the LDA-MLLT transforms). These experiments show that the feature transforms bring important ASR improvements on clean, read speech (8% relative WER reduction). However, on noisy speech the results are worse that those obtained with the noise-robust ASR system (AM001). The effect of these feature transforms needs to be analyzed more deeply before we can draw a solid conclusion.

E. Comparison with Google ASR

The multi-lingual ASR system created by Google can be accessed through the Internet, as a speech transcription web-service. Using an HTTP GET request (which comprises metadata, including the language, and an audio file) one can obtain the transcription of short audio speech files. In December 2013 we transcribed all the utterances in the evaluation speech corpora RSC-eval and SSC-eval using the Romanian version of this ASR system.

Google's ASR system for the Romanian language was probably developed as a data-driven system and does not take into account the language's specific issues such as diacritics, hyphenated compound words, etc. Therefore, aligning the raw Google output with the

TABLE V. ASR IMPROVEMENTS BROUGHT BY THE NOISE FEATURES + FEATURE TRANSFORMS

		WER [%]			
Acoustic Model	**Speech Features**	**RSC-eval**	**SSC-eval (all)**	**SSC-eval (clean)**	**SSC-eval (noisy)**
AM002	MFCCs	16.7	39.4	31.4	51.1
AM001	PNCCs	16.2	38.9	31.7	49.1
AM005	PNCCs + LDA-MLLT	14.8	39.1	31.2	50.7

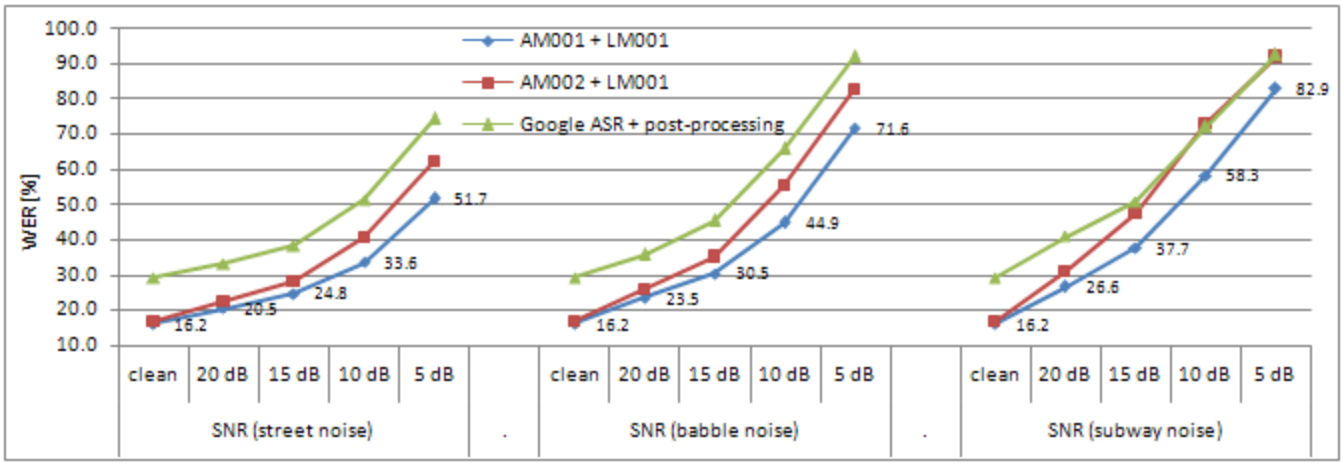

Figure 1. ASR improvements brought by the noise robust features + comparison with Google ASR

TABLE VI. SPEED'S ASR SYSTEM VS. GOOGLE'S ASR SYSTEM

	WER [%]	
ASR System	**RSC-eval**	**SSC-eval**
AM001 + LM001	16.2	38.9
Google ASR	43.3	60.0
Google ASR + post-processing	29.2	51.4

reference transcriptions in our evaluation speech corpora leads to very poor WERs (see line 2 in Table VI). In order to do a "fair" comparison we restored the diacritics on Google ASR's output using our diacritics restoration system [1] and used these post-processed transcriptions for evaluation (see line 3 in Table VI).

Comparative results between our best ASR system and Google's ASR system are presented in Table VI and Figure 1. Table VI shows the results on clean read speech and spontaneous speech and Figure 1 illustrates the results on noisy read speech. On clean read speech the absolute WER obtained by our ASR system is 13% lower, while on spontaneous speech (clean and noisy) it is 18.6% lower. On noisy read speech the absolute WERs of our system are between 13.1% and 17.6% lower, depending on the SNR. Regardless of the type of speech our ASR system is at the moment much better than Google's ASR system for the Romanian language.

VI. CONCLUSIONS

This paper presented the developments made to SpeeD's ASR system for the Romanian language since 2011. We briefly described the recently collected speech and text resources and presented the significant gains in ASR accuracy obtained by incorporating them in the acoustic and language models. Furthermore, we discussed the noise robustness approach in our ASR system and showed how the newly introduced acoustic features and feature-transforms improve the accuracy. Overall, the relative WER reductions obtained by the 2013 system are: 32.0% for clean read speech and 33.2% for spontaneous speech.

Another important conclusion is that the most significant gain in performance (on clean speech) was obtained thanks to the acquisition of new training speech corpora. This means that our ASR system is still in need of more training data and that, at this particular moment, this direction might be even more important than implementing and using the state-of-the-art ASR techniques.

Finally we compared our ASR system with the Google ASR web-service and showed that we outperform Google on all types of speech (read/spontaneous, clean/noisy).

ACKNOWLEDGMENT

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 author(s) and do not necessarily reflect those of the Romanian-American Foundation.

REFERENCES

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

[2] H. Cucu, L. Besacier, C. Burileanu, A. Buzo, "Enhancing Automatic Speech Recognition for Romanian by Using Machine Translated and Web-based Text Corpora," in Proc. Int. Conf. Speech and Computer (SPECOM), Kazan, Russia, 2011, pp. 81-88.

[3] B. Tarjan, T. Mozsolics, A. Balog, D. Halmos, T. Fegyo, P. Mihajlik, "Broadcast news transcription in Central-East European languages," in Proc. Int. Conf. Cognitive Infocommunications (CogInfoCom), Kosice, Slovakia, pp. 59-64.

[4] 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.

[5] C. Kim, R.M. Stern, "Power-Normalized Cepstral Coefficients (PNCC) for robust speech recognition," in Proc. Int. Conf. Acoustics, Speech and Signal Processing (ICASSP), Kyoto, Japan, 2012, pp. 4101-4104.

[6] G. Saon, M. Padmanabhan, R. Gopinath, S. Chen, "Maximum likelihood discriminant feature spaces," in Proc. Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP), Istanbul, Turkey, 2000, pp. 1129-1132.

[7] N. Kumar and A.G. Andreou, "Heteroscedastic discriminant analysis and reduced rank HMMs for improved speech recognition," in Speech Communication, vol. 25, no. 4, pp. 283–297, 1998.

[8] M.J.F. Gales, "Semi-tied covariance matrices for hidden Markov models," in IEEE Transactions on Speech and Audio Processing, vol. 7, no. 3, pp. 272–281, 1999.