ASR FOR LOW-RESOURCED LANGUAGES: BUILDING A PHONETICALLY BALANCED ROMANIAN SPEECH CORPUS

Miruna Stănescu (Paşca), Horia Cucu, Andi Buzo, Corneliu Burileanu

University "Politehnica" of Bucharest, Romania

stanescu.miruna@gmail.com

ABSTRACT

The construction of automatic speech recognition (ASR) systems is fundamentally dependent on the speech corpus used to train the acoustic models. The speech corpus should be phonetically balanced to assure that the acoustic models are properly trained. This paper presents the design and development of the first phonetically balanced Romanian speech corpus. It describes all the language processing steps taken in order to obtain a proper set of phrases, discusses some important aspects regarding Romanian phonetics and emphasizes the phrase selection mechanism.

Index Terms—ASR, corpora acquisition, corpora processing, diacritics restoration

1. INTRODUCTION

The phonetic characteristics of a speech corpus play a key role in the robustness of the future speech application. The construction of a speech corpus having particular phonetic characteristics must first focus on phrase selection and recording rather than speech labeling. The set of sentences to be recorded can be chosen to cover the phonetic events of a language either with an approximately uniform distribution or according to their frequency of occurrence in natural speech. In the first case the resulted corpus will be *phonetically rich*, while in the second case the resulted corpus will be *phonetically balanced* [1]. The first type of corpus is usually well suited for training text-to-speech (TTS) systems, while the second type is better adapted for the development of automatic speech recognition systems.

Obviously, all these issues regarding the distribution of the phonetic events within speech corpora are only important when such corpora are not already available and one needs to construct them from scratch. For languages such as English, French, German and many others this is a closed topic for quite a long time (more than 10 years). For these internationally-spoken languages, the resources needed to create robust ASR or TTS systems are widely available. On the other hand, for the so-called low-resourced languages the absence of large and standardized text and speech corpora is still a major obstacle in the development of robust speech applications. Speech corpora development was lately reported for languages such as Bangla [2], Bengali [3], Ukrainian [4] and Urdu [5]. All these papers highlight the various issues encountered and over-passed while creating speech corpora for under-resourced languages.

For the Romanian language there are only a few small continuous speech corpora, all created by research groups, among which only one [6] is freely available. The largest Romanian speech corpus was previously created by our research group and presented in [7][8]. Neither the other corpora, nor our previously developed speech corpus were designed to be phonetically balanced.

This work focuses on describing and detailing all the natural language processing (NLP) steps we made to select a small set of phrases which cover all the phones in Romanian and also maintain their real occurrence distribution. These phrases will be further recorded by several hundred speakers to create the first phonetically balanced Romanian speech corpus.

The rest of this paper is organized in five sections. Section 2 summarizes the corpus development procedure and Section 3 deals with various NLP issues regarding text corpus acquisition, normalization and phonetization. Section 4 uses the phonetically transcribed text to present and analyze the first statistics regarding the Romanian language phonetics. Section 5 deals with text selection methods and in the end, Section 6 draws some conclusions.

2. CORPUS CONSTRUCTION PROCEDURE

The construction of a phonetically balanced speech corpus consists of two major stages: a) selection of a proper set of phrases and b) recording this set of phrases using several hundred speakers. Just as illustrated in Figure 1, this work presents all the sub-steps we took to accomplish stage a).

Before the sentence selection is performed, we must have a large number of sentences from which to select. Consequently, the first step in the development procedure

consists in the acquisition of a large text corpus. The text corpus, which in our case was solely acquired over the Web, had to be normalized and pre-processed before it could serve as transcripts for a new speech corpus. As shown in Figure 1, this is the second step we made. The normalized (processed) text consisted of about 9.7 million phrases.

After the text corpus was normalized it has been subject to a first selection phase. This selection process aimed to remove the phrases considered to be difficult to read: phrases that contained unusual words, very short phrases and very long phrases. After this step the list of phrases was significantly reduced to 22,287 phrases.

Next, the selected phrases were phonetically transcribed using an existing phonetic dictionary and an automated graphemes-to-phonemes system. Consequently, after this step the text corpus was composed of a list of phonetically transcribed phrases.

Meanwhile, the normalized text was also phonetically transcribed (with the same tools) and this transcription was used to compute the occurrence distribution of the Romanian phones. Given that the phonetically transcribed text had about 850 million phones we take the resulted statistics to be representative for the Romanian language.

In the end, we used the phonetic statistics to select a small group of only 200 phrases among the 22,287 phonetically transcribed phrases. For this we used the add-on phrase selection method [1] optimizing a phonetic balance score, as described in Section 6.

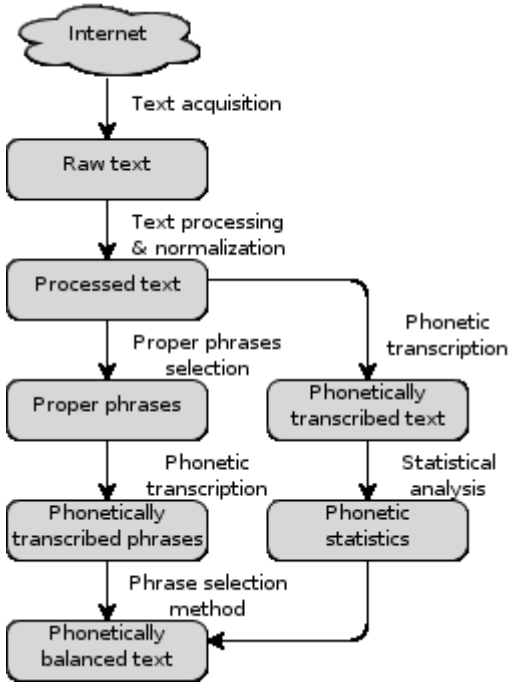

Figure 1. Corpus construction procedure

3. TEXT ACQUISITION, PROCESSING AND PHONETIZATION

The process of text corpora acquisition is at this moment dominated by the Web-as-resource or Web-as-Corpus (WaC) approach. Consequently, we have also used this approach: we have collected news from various Romanian online newspapers and we have also used the transcripts of the discussions in the European Parliament [9].

Given this acquisition method it is clear that the first processing operation to be accomplished was html-to-text conversion. Next, the abbreviations were expanded using a previously existing abbreviations list, the differently formatted numbers were programmatically converted to plain text, the punctuation marks and other special characters were cleaned out and the text was arranged in a one-phrase-per-line manner. In the end we used a Romanian lexicon to remove all the phrases that contained spelling errors or foreign words. All these processing operations were accomplished in order to clean the corpus of tokens that could eventually lead to pronunciation ambiguities during the recording stage.

One last NLP operation to be employed was the restoration of diacritics. This was necessary because all the news corpora which were acquired over the Web come without diacritics. For a news article, the diacritics are not very important since any reader has access to the paragraph-level context and ambiguities seldom appear. On the other hand, a short phrase with undiacriticized words is very often ambiguous to read. For this processing stage we used the diacritics analysis and restoration tool presented in [7]. This system was evaluated to have a character error rate of about 0.48% which is considered acceptable for our particular task.

After all these NLP and cleaning operations, we ended up with a normalized text of about 9.7 million phrases, this being the largest Romanian plain text corpus used for research purposes [7].

The resulted normalized text was afterwards phonetized because, as Figure 1 illustrates, the task of selecting a set of phonetically balanced phrases cannot be accomplished without having the phonetic transcription for every phrase in the normalized text corpus.

Initially we tried to transcribe the corpus using a previously existing phonetic dictionary, but we failed due to the fact that the dictionary was missing proper names which are very often encountered in news articles. The amount of words that were missing from the dictionary was very large (tens of thousands) so this issue couldn't have been manually approached.

To solve this problem we designed and implemented an automated graphemes-to-phonemes tool [7] and eventually used it to transcribe all the words that were missing from the phonetic dictionary. This machine

learning system was evaluated to have a phone error rate of 0.31% which is considered acceptable for our particular task. In fact, the transcription error for the whole text corpus is even lower because all the known words were transcribed using the phonetic dictionary, while the error-prone graphemes-to-phonemes tool was only used for the unknown words (approximately 1/4 of the words).

4. ROMANIAN PHONETIC ANALYSIS

The phonetically transcribed corpus is not sufficient to select a set of phonetically balanced phrases. There is one key resource missing: the Romanian phonetic statistics or, in other words, the occurrence distribution for the phones in the Romanian language. To the best of our knowledge, such a statistics for Romanian does not exist. Only one published work [10] presents some phonetic statistics, but these cannot be considered as representative for the Romanian language due to the small size of the corpus: 2500 phrases.

Given this situation and the fact that we possessed the phonetically transcribed corpus (comprising about 850 million phones), we decided to create our own phonetic statistics. The resulted Romanian phones occurrence distribution is presented in Table 1. Detailed information regarding these statistics and a more in-depth analysis are given in [11] (paper submitted to ELMAR 2012).

For our discussion it is very interesting to note the highly unbalanced occurrence distribution for the Romanian phones: a) the most frequent phone occurs as often as the least frequent 18 phones altogether and b) the most frequent 6 phones cover 50% of all phone occurrences. This fact sustains the importance of a phonetically balanced corpus (a corpus in which the phones appear with their real frequency). For example in the case of an ASR system it is more desirable to better train the acoustic models which are found more often during recognition (the models for the frequent phones) and invest less effort in training the less frequent phones acoustic models.

In order to verify that these statistics are consistent over the entire text corpus we made the following experiment: a) we randomized the order of the phrases within the corpus, b) we split the corpus into three equal-sized sub-corpora and c) we computed the statistics on these three sub-corpora. The correlation coefficients computed between the three sub-corpora and the whole corpus and between pairs of sub-corpora were all very close to 1 (differed at the 7th decimal). The experiment and its result certify that the corpus on which the statistics were computed is large enough and that these statistics are representative for the Romanian language.

Phone (IPA)	Word Example	Freq [%]
e	mar**e** (sea/large)	11.20%
a	s**a**t (village)	9.77%
i	**i**ft (elevator)	7.97%
r	**r**isc (risk)	7.41%
t	**t**ot (all)	6.61%
n	**n**as (nose)	6.40%
u	ş**u**t (shot)	5.57%
l	**l**ac (lake)	4.69%
o	l**o**c (place)	4.48%
s	**s**are (salt)	4.10%
d	**d**ar (gift)	3.54%
k	a**c**um (now)	3.40%
p	**p**ar (pole)	3.36%
ə	gur**ă** (mouth)	2.88%
m	**m**ăr (apple)	2.87%
j	**fi**ară (wild animal)	2.21%
ʧ	**c**enuşă (ash)	1.83%
ɨ	**î**ntre (between)	1.31%
ʃ	co**ş** (basket)	1.30%
v	**v**apor (ship)	1.23%
f	**f**aţa (the face)	1.10%
z	**z**ar (dice)	1.09%
ʦ	**ţ**ăran (peasant)	1.04%
b	**b**ar (bar)	0.94%
ʲ	tar**i** (strong)	0.65%
e̯	**d**eal (hill)	0.64%
g	**g**alben (yellow)	0.63%
w	sa**u** (or)	0.61%
ʤ	**g**irafă (giraffe)	0.27%
o̯	**o**ase (bones)	0.24%
ʒ	a**j**utor (help)	0.22%
c	**ch**em (call)	0.21%
h	**h**arta (the map)	0.20%
ɟ	un**gh**i (angle)	0.03%

Table 1. Romanian phones occurrence distribution

5. TEXT SELECTION

Having now all the resources (the phonetically transcribed corpus and the phonetic statistics) at hand, we proceeded to selecting the set of phonetically balanced phrases. In fact, as Figure 1 shows, the text selection task was approached in two steps: a) we selected a subset of proper phrases from the processed text and b) from the subset of proper phrases we selected the set of phonetically balanced phrases.

Even though the text corpus collected from the Internet contains general Romanian phrases, some of them are not easy to read aloud. As our final goal was to create a set of phrases that would eventually serve as prompts in the recording stage, we decided to remove the phrases

considered to be difficult to read: phrases that contained unusual words, phrases with less than 5 words and phrases with more than 15 words. The unusual words were considered to be those words which are not part of the list of the most frequent 64k words in Romanian (as computed on the same text corpus). Moreover at this point we have also removed the duplicate phrases. After this first selection process the list of phrases was significantly reduced to 22,287 proper phrases.

Next we approached the task of selecting a very small set of phonetically balanced phrases (about 200 phrases). Due to time and money constrains, we cannot ask the speakers which will participate in the recording sessions to speak more than an hour. This is why the set of phrases to be selected should not exceed 200 phrases. Given this, the task of having a phonetically balanced set became more difficult. To asses its difficulty we performed one more experiment. We have randomly split the set of proper phrases into groups of 100, 200, 500, 1k, 2k and 5k phrases and evaluated the phonetic balance of these groups. The phonetic balance for every phrase group was evaluated as the correlation coefficient between the phones distribution within the phrase group and the phones distribution in Romanian (as computed in Section 4). The correlation coefficient was chosen out of the need for a quantitative measure that would allow us to compare the two distributions. Qualitatively, they could be compared using a plot like in Figure 3. Here, if we try to assess the differences between the occurrence distribution in Romanian and the occurrence distribution in a randomly selected set of 200 phrases, we see that differences as high as 3% may appear. For the "rare' phones (phones with low number of occurrences), a one percent difference can actually represent half the occurrences for that particular phone. So, in order for the distributions to be aligned for all phones, we are using Pearson's correlation coefficient as a measure for this alignment.

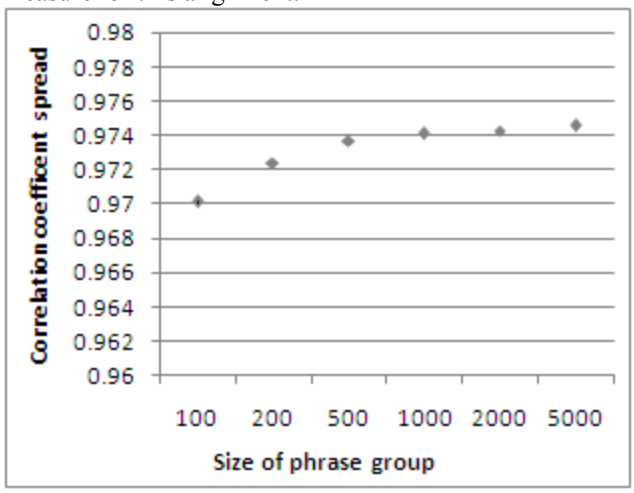

Figure 2. The phonetic balance of randomly selected phrase sets

The results are presented in Figure 2 and show that the average correlation coefficient varies between 97% and 97.5%. This is a very low correlation for our task. Moreover, we can see that even if the size of the set of phrases is increased by 2, 5, 10, 20 or even 50 times, the correlation coefficient does not vary very much.

The conclusion we can draw based on the results in Figure 2 is that that a randomly selected set of 200 phrases is very unlikely to have a phonetic distribution that is highly correlated with the real phonetic distribution in the Romanian language. This conclusion sustains all the efforts presented in this paper, confirming that if the final target is a small set of phonetically balanced phrases, than this set of phrases cannot be chosen arbitrarily.

Given the set of 22k proper phrases the only way to select the best set of 200 phrases would have been to form all the possible set of 200 phrases and evaluate their phonetic balance. Obviously this method is way too time consuming. Because of this we decided to use the add-on procedure similar to [1]. This method starts with an empty set of selected phrases consists in the following steps:

1) For each phrase within the set of proper phases, compute the number of distinct phones that do not appear in the set of selected phrases;

2) Select the proper phrase with the most distinct phones that do not appear in the set of selected phrases, and move it to the set of selected phrases;

3) Repeat steps 1) and 2) until the set of selected phrases contains all the phones (this preselecting procedure assures that the rare phones are not missing from the selected phrases);

4) Compute the phonetic balance score for the set of selected phrases. Obviously, at first, when the set contains only the preselected phrases, this score will be very bad;

5) For each phrase within the set of proper phrases compute the improvement in the phonetic balance score brought by this phrase;

6) Select the proper phrase that brings the biggest improvement in the phonetic balance score and move it into the set of selected phrases;

7) Repeat steps 5) and 6) until the set of selected phrases reaches the desired size (in our case 200 phrases).

Just as before, we used the Pearson's correlation coefficient between the phones distribution in the set of selected phrases and the real phones distribution in Romanian as the phonetic balance score.

Using the add-on procedure we managed to select a set of phrases with a phonetic balance score of 99.8%. This means that its phones distribution is highly correlated with the real Romanian phones distribution. This result is also illustrated in Figure 3 which clearly shows that the two distributions are very similar. For comparison, in Figure 3 we have also plotted the phones occurrence distribution for a randomly selected set of 200 phrases.

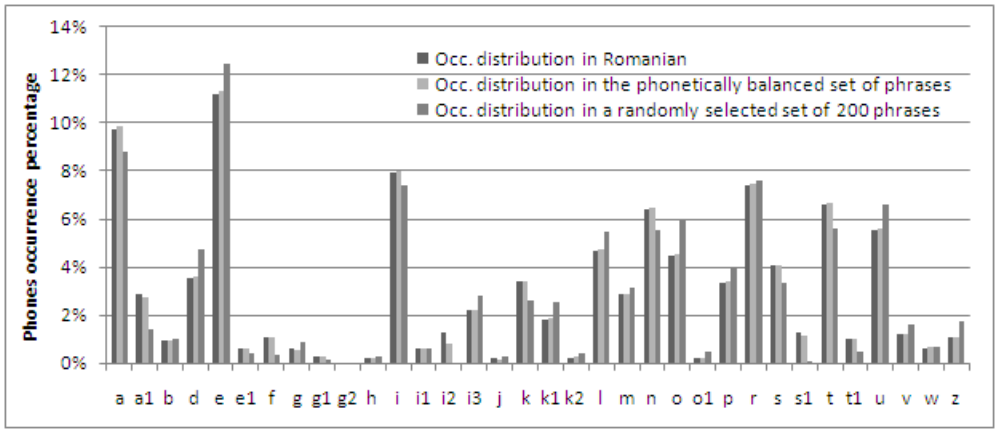

Figure 3. The phones occurrence distribution in the phonetically-balanced vs. a randomly-selected set of phrases

6. CONCLUSION

In this study we presented the methodology we have employed to select a set of phonetically balanced transcripts which will be further used to create a Romanian speech corpus. Several key issues starting with text corpora acquisition and processing, going to phonetic analysis and transcription and finally phrase selection procedures have been tackled.

Although for many internationally spoken languages the construction of text/speech corpora is no longer of large interest, we have shown that for low-resourced languages, such as Romanian, the need for high-quality speech resources is very real. The development of competitive speech applications is highly dependent on the construction of these resources.

This paper has also presented the acquisition and processing details for what is, to the best of our knowledge, the largest Romanian text corpus ever used for research. Based on this corpus and several other NLP tools which we have previously developed we managed to obtain the Romanian phones occurrence distribution. Given the large size of the data and the consistency experiments discussed in this paper we can assert that these statistics are representative for the Romanian language. To the best of our knowledge this is the first statistic of this kind published for our language.

Using all these resources we finally presented the various text selection steps and ended up with a phonetically balanced set of phrases that will be recorded in the near future by several hundred speakers. The resulted phonetically balanced speech corpus will be the first of this kind for Romanian.

7. ACKNOWLEDGEMENT

The work has been funded by the Sectoral Operational Programme Human Resources Development 2007-2013 of the Romanian Ministry of Labour, Family and Social Protection through the Financial Agreement POSDRU/88/1.5/S/60203.

8. REFERENCES

[1] V. Radová, P. Vopálka, "Methods of Sentences Selection for Read-Speech Corpus Design," *TSD 1999*, Pilsen, Czech Republic.

[2] M. Habib, F. Alam, R. Sultana, S.A. Chowdhury, M. Khan, "Phonetically balanced Bangla speech corpus," *HLTD 2011*, Alexandria, Egypt, 2011.

[3] S. Mandal, B. Das, P. Mitra, A. Bas, "Developing Bengali Speech Corpus for Phone Recognizer Using Optimum Text Selection Technique," *IALP 2011*, Penang, Malaysia, 2011.

[4] V. Pylypenko, V. Robeiko, M. Sazhok, N.Vasylieva, O. Radoutsky, "Ukrainian Broadcast Speech Corpus Development," SPECOM 2011, Kazan, Russia, 2011.

[5] A.A. Raza, S. Hussain, H. Sarfraz, I. Ullah, Z. Sarfraz, "Design and development of phonetically rich Urdu speech corpus," *Oriental COCOSDA 2009*, Urumqi, China, 2009.

[6] A. Stan, J. Yamagishi, S. King, M. Aylett, "The Romanian speech synthesis (RSS) corpus: Building a high quality HMM-based speech synthesis system using a high sampling rate", Speech Communication, vol. 53, pp. 442-450, 2011.

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

[8] A. Buzo, "Automatic Speech Recognition over Mobile Communication Networks", PhD Thesis, Bucharest, Romania, 2011.

[9] P. Koehn, "Europarl: A Parallel Corpus for Statistical Machine Translation," *MT Summit 2005*, Phuket, Thailand, 2005.

[10] A. Stan, M. Giurgiu, "Romanian language statistics and resources for text-to-speech systems," *ISETC 2010*, Timisoara, Romania, 2010.

[11] M. Stănescu, A. Buzo, H. Cucu, C. Burileanu, "Statistical Phonetic Analysis of the Romanian Language for Speech Recognition and Synthesis Tasks," submitted to *ELMAR 2012*.