Statistical Phonetic Analysis of the Romanian Language for Speech Recognition and Synthesis Tasks

Miruna Stănescu (Paşca), Andi Buzo, Horia Cucu, Corneliu Burileanu
University "Politehnica" of Bucharest, Splaiul Independentei nr. 313,
Bucharest, Romania
mirunapasca@yahoo.com

Abstract - This article provides a statistical phonetic analysis based on the largest Romanian text corpus collected so far for research purposes. Several types of phonetic events are analyzed: phones, diphones, triphones, and phone clusters based on the general classification of phones in the Romanian language. Some interesting conclusions are drawn, such as the fact that less than half the diphones cover 99% of the whole text. The article also discusses some usages of these phonetic statistics for spoken language technology tasks.

Keywords – Spoken language technology; Text-to-speech; Automatic speech recognition; Phonetic event

I. INTRODUCTION

Research in the field of spoken language technology includes two main subfields: building text-to-speech (TTS) systems, and building automatic speech recognition (ASR) systems.

For both tasks, the main resource is the corpus, whether it is a speech corpus, or text corpus. From both points of view, Romanian is still an under-resourced language. TTS and ASR systems are being built or optimized in several research teams, like [1], [2], [3], [4], but text corpora and speech corpora are still relatively small and not always freely available (compared to languages like English, French or German, etc.). As such, corpora acquisition is an ongoing challenge for most spoken language technology tasks, and statistics that would help this process are almost non-existent.

For TTS or ASR systems, one of the major inputs is the speech database; when creating such a database for an under-resourced language like Romanian, the first step is selecting the phrases to be recorded. This paper tries to assess the need for taking into account phonetic statistics during this selection phase. This would result in resources and efforts being focused on the most frequent phonetic events, for example creating better trained models for the triphones found most often during speech recognition. We obtained the needed statistics based on the largest text corpus collected for the Romanian language, and the conclusions drawn from these results could influence the future development of Romanian speech processing tasks.

The rest of this paper is organized in four sections. Section 2 examines the phonetic particularities of the Romanian language, as well as other research relevant to the task at hand. Section 3 details the corpus acquisition and processing steps to bring this corpus to a phonetically transcribed form, needed for the phonetic statistics. Section 4 discusses the experimental results and in the end, Section 5 draws some conclusions.

II. RELATED WORK

To our knowledge, there is only one relevant paper [5] showing different phonetic events (phones, diphones and quinphones) statistics for Romanian. However, these statistics are computed on a very small corpus of approximately 2500 newspaper sentences, called the *RSS-text* (Romanian Speech Synthesis-text) corpus. These sentences have been selected from a larger (about 1.7 million words) corpus collected from online newspaper articles. The selection was made for the best coverage of the Romanian diphones with at least 10 occurrences in the words of the DEX (Romanian Dictionary) online database. A similar text selection technique is used by another TTS system, in [4], and ASR systems [1] currently take phone coverage into account when selecting phrases for the training database.

However, since state-of-the-art ASR systems are based on triphones (context-based training for phones), it is clear that such statistics should also be taken into account when developing speech or text corpora.

Phonetic statistics have been recently published for other under-resourced languages, like Polish [6] and Turkish [7]. For Polish [6], the authors take into account the gap between words (as a phone pronounced at the beginning of a word is not the same as one pronounced in the middle or at the end of the word), while for Turkish they do not.

One of the particularities of the Turkish language [7] is that there are some interesting conclusions to be drawn from organizing the consonants into groups like soft, hard, sustainable, and unsustainable. For the Romanian language, this classification is different, as detailed below.

The Romanian language has some phonetic particularities [4, 8] that have to be taken into account when creating text or speech corpora. Compared to English [9], some phones are missing (e.g. the *th* in *thin*); there are two vowels typical for Romanian (ə, ɨ), and a whole new group called *semi-vowels* (e̯, j, o̯, w). They are similar to the English glides (vowels whose position in the syllable require them to be a little shorter and cannot be stressed), but for Romanian they are classified separately from both vowels and consonants. The sonant – non-

sonant classification is the same as for English: sonant phones (l, r, m, n) are generated in a way similar to vowel generation (the so-called musical tones), but within the syllable they behave like consonants. Based on the manner of articulation, the non-sonant phones are classified into plosives (e.g. b), fricatives (e.g. v) and affricates (e.g. ʦ), and this classification is obviously the same for Romanian and English.

III. Corpus Acquisition and Processing

The text corpus under discussion has been collected using the Web-as-Corpus (WaC) approach: it consists of news from various Romanian online newspapers, as well as the transcripts of the discussions in the European Parliament [10]. We were able to collect an extremely large corpus (about 170 million words) due to the completely automated manner in which the text collection and processing are being done. The natural language processing (NLP) steps taken to convert the html articles to a form suited to our needs are described below [1]:

- Text normalization (consisting of html-to-text conversion, expansion of the abbreviations in the original text, converting numbers to plain text, and removal of special characters such as punctuation marks);
- Diacritics restoration (a critical step, due to the fact that for short and elliptical texts, such as the output of an ASR system, the lack of diacritics could make them ambiguous or even incomprehensible; the training of the ASR should also take into account the diacritical characters);
- Phonetic transcription (using an automated graphemes-to-phonemes tool).

Subsequently, to count the phones, diphones and triphones in our phonetically transcribed text, we used the *ngram-count* tool of the SRI Language Modeling (SRI-LM) Toolkit [11].

IV. Phonetic Analysis – Results

In order to verify that the statistics would be consistent and representative for the Romanian language we started with the following experiment:

- We randomized the order of the phrases within the corpus;
- We split the corpus into three equal-sized sub-corpora;
- We computed the statistics on these three sub-corpora.

The correlation coefficients (Pearson's) 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). This meant that the phones/diphones/triphones occurrence distribution was approximately the same for the corpus and its sub-corpora. The experiment and its result certify that the corpus on which the statistics were computed is large enough for these statistics to be considered representative for the Romanian language.

A. Phones Occurrence Distribution

The first step of our phonetic statistic task was computing the phones occurrence distribution in our corpus, presented in Table 1.

This distribution differs from the one reported in [5] for the RSS-text corpus, due to the difference in corpus size. However, the most frequent phones are the same, even if their frequencies are not.

Table 1 depicts the highly unbalanced occurrence distribution for the Romanian phones: the most frequent phone occurs as often as the least frequent 18 phones altogether. Also, the most frequent 6 phones cover 50% of all phone occurrences.

TABLE I. Romanian phones occurrence distribution

Phone (IPA)	Word Example	Freq [%]
e	mar**e** (sea/large)	11.20%
a	s**a**t (village)	9.77%
i	l**i**ft (elevator)	7.97%
r	**r**isc (risk)	7.41%
t	**t**o**t** (all)	6.61%
n	**n**as (nose)	6.40%
u	ș**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**a**ră (wild animal)	2.21%
ʧ	**c**enușă (ash)	1.83%
ɨ	**î**ntre (between)	1.31%
ʃ	co**ș** (basket)	1.30%
v	**v**apor (ship)	1.23%
f	**f**ața (the face)	1.10%
z	**z**ar (dice)	1.09%
ʦ	**ț**ăran (peasant)	1.04%
b	**b**ar (bar)	0.94%
ʲ	tar**i** (strong)	0.65%
e̯	d**e**al (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%
ɣ	un**gh**i (angle)	0.03%

This fact emphasizes the importance of selecting a small corpus that is phonetically balanced (the phones appear with their real frequency) to be used for training ASR systems, just as it is shown in [12]. For such systems 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. A similar argument could be made for TTS systems [5], but in this case the diphone, triphone and even quinphones have to be taken into account.

B. The Most Frequent Diphones and Triphones

The diphone distribution computed from our corpus shows that less than half of the diphones present in the corpus are enough to cover 99% of the text (529 diphones of 1092 found in the text). A similar conclusion can be drawn from the triphone statistics: less than a third of the triphones (7263 of 25335 found in the corpus) sum up to 99% of the triphones in the full corpus.

Table 2 provides the 10 most frequent diphones and ten most frequent triphones, as resulted from the statistical analysis of our corpus. Most of the entries in this table (e.g. *de*, *ar*, *are*) can be either syllables or words – they are counted the same in our current statistical approach. This led to the idea that a more in-depth analysis could be performed by taking into account the spaces between words. For TTS and ASR systems, differentiation should be made between a phone pronounced at the beginning, in the middle or at the end of the word. Thus, we computed the same statistics in a different manner, resulting diphones and triphones that may contain the space between words in a phrase (marked by #).

A sample of such statistics is shown in Table 3. We see that the most frequent triphone not containing # is *are* (it can be either a part of a word, or a verb meaning *has*) – the same as in Table 2 (where the space between words is not taken into account). However, there are fourteen more frequent triphones than *are*: short words, or word beginnings or endings. For example, a special case is the diphone *de*, which can be either: a preposition, a beginning, or an ending of a word. [5] also reports *de* as being the most common word in their newspaper text corpus (which contains about 1.7 million words).

TABLE II. THE MOST FREQUENT DIPHONES AND TRIPHONES

Diphones		**Triphones**	
Diphone (IPA)	**Freq [%]**	**Triphone (IPA)**	**Freq [%]**
re	1.84%	are	0.51%
de	1.57%	ent	0.41%
ar	1.47%	est	0.39%
in	1.43%	zeʧ	0.38%
te	1.42%	pre	0.36%
at	1.21%	ntr	0.63%
er	1.1.8%	tru	0.35%
ul	1.12%	ede	0.33%
ɨn	1.09%	owə	0.33%
ri	1.04%	din	0.31%

TABLE III. THE MOST FREQUENT TRIPHONES INCLUDING SPACE (#)

Triphone (IPA)	**Freq [%]**
#de	1.29%
de#	1.05%
te#	0.83%
#in	0.79%
ul#	0.64%
#ʃi	0.63%
re#	0.63%
ʃi#	0.62%
#pe	0.55%
le#	0.53%
ɨn#	0.53%
#pr	0.53%
ij#	0.52%
#a#	0.46%
are	0.44%

C. Phone clusters statistics

Next, we grouped the phones into several clusters based on voicing and manner of articulation:

- *Vowels* (V): a, ə, e, i, ʲ, ɨ, o, u;
- *Semi-vowels* (SV): e̯, j, o1, w;
- *Plosives* (P): b, p, d, t, g, k, c, ɟ;
- *Fricatives* (F): v, f, z, s, ʃ, ʒ, h;
- *Affricates* (A): ʦ, ʧ, ʤ;
- *Sonant* (S): l, r, m, n.

We designed these clusters based on the general phoneme classification for the Romanian language [4], discussed in the introductive chapter of this article.

The occurrence distribution of the clusters in the corpus is presented in Table 4. It clearly shows two interesting particularities of the Romanian language: a) the vowels are the most frequent group, and b) the Sonant group (which contains only four phones) is more frequent than the Plosives group (which contains twice as many phones – eight).

An even more interesting conclusion extracted from the clustering of the corpus came after we computed statistics based on diphone categories, as descried below.

All the diphones in the Romanian language can be grouped into 38 categories – for example, a category named *V – SO* contains all possible combinations between a vowel and a sonant. The frequency of each diphone category can be computed the same as for the non-clustered corpus, and Table 5 shows the 7 most frequent categories.

These seven categories make up 73% of the whole corpus; furthermore, the five most common categories make up 61% of the whole corpus.

TABLE IV. CLUSTER DISTRIBUTION

Cluster	Freq [%]
V	43.86%
SO	21.37%
P	18.81%
F	9.13%
SV	3.72%
A	3.12%

TABLE V. MOST FREQUENT DIPHONE CATEGORIES

Diphone category	Freq [%]
V – SO	16.94%
SO – V	14.06%
P – V	13.60%
V – P	10.68%
V – F	6.43%
F – V	6.42%
V – V	4.98%

Summing up our results, less than half the diphones are enough to cover 99% of the text and five diphone categories cover 61% of the text. These observations could help to improve the efficiency for speech processing tasks if efforts could be directed towards the most frequent phones, diphones, or clusters. Of course, the less frequent phonetic events should not be neglected, but excessive resources should not be invested in them, either.

V. CONCLUSIONS

This article has shown phones, diphones, triphones, and clusters statistics obtained from the largest available Romanian text corpus.

We have drawn some conclusions from these statistics, such as: the phones occurrence distribution is highly unbalanced, less than 50% of all diphones cover 99% of the corpus, and the five most common diphone categories make up 61% of the whole corpus.

To our knowledge, so far the corpus selection for spoken language technology tasks has been performed based on a targeted uniform phones or diphones coverage; afterwards, statistics have been computed to assess the phonetic events distribution in these selected corpora. We have tried to assess the possibility of a new approach: using techniques such as the one in [12], the statistics computed on a much larger corpus could be used as an input for sub-corpus selection. This would assure a better quality of the key resources in a TTS or ASR system (text or speech corpora), as well as a better channelization of the computing effort towards the most frequent phonetic events.

ACKNOWLEDGMENT

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.

REFERENCES

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

[2] A. Stan, "Romanian HMM-based text-to-speech synthesis with interactive intonation optimization," PhD Thesis, Cluj-Napoca, Romania, 2011.

[3] C. Ungurean, "Interfețe de comunicare prin voce cu dispozitive portabile", PhD Thesis, Bucharest, Romania, 2011.

[4] D. Burileanu, "Basic research and implementation decisions for a text-to-speech synthesis system in Romanian", in International Journal of Speech Techonology, Kluwer Academic Publishers, 2002.

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

[6] B. Ziołko, J. Gałka, S. Manandhar, R. C. Wilson, and M. Ziołko, "Triphone statistics for Polish language", in Human Language Technology. Challenges of the Information Society, Springer-Verlag Berlin, Heidelberg, 2009.

[7] I. B. Uslu, A. E. Yilmaz and H. G. Ilk, "Statistical digram and trigram analysis of Turkish in terms of coverage and entropy for possible language and sppech based applications", EUSIPCO 2010, Aalborg, Denmark.

[8] D. Burileanu, "Contribuții la sinteza vorbirii din text pentru limba română", PhD Thesis, Bucharest, Romania, 1999.

[9] X. Huang, A. Acero, and H. Wuen-Hon, Spoken Language Processing – A Guide to Theory, Algorithm, and System Development, Prentice Hall, 2001, pp.36-48.

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

[11] A. Stolcke, "SRILM - an extensible language modeling toolkit", ICSLP 2002, pp. 901-904, Colorado, USA.

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