TECHNICS AND INFORMATICS IN EDUCATION
6th International Conference, Faculty of Technical Sciences, Čačak, Serbia, 28–29th May 2016
TEHNIKA I INFORMATIKA U OBRAZOVANJU
6. međunarodna konferencija, Fakultet tehničkih nauka, Čačak, Srbija, 28–29. maj 2016.

UDK: 004.6:811.163.41`342.4 Professional paper

Serbian language speech database "Phonemes_1.0": Design and application

Branko Marković[1], Vladimir Milićević[1], Dragana Petrović[1], Dejan Nešković[1] and Gordana Marković[2]
[1]Čačak Technical College, Čačak, Serbia
[2]Technical School, Čačak, Serbia
e-mail brankomarko@yahoo.com

Abstract: *In this paper we explained how to create Serbian speech database called "Phonemes_1.0" and how to use it for pattern match tests. This database contains a list of all 30 phonemes that cover the Serbian alphabet called "Azbuka". The database is divided in two parts: vowels and consonants. For vowels we applied an initial DTW comparison.*

Keywords: *Serbian speech database; vowels; consonants; DTW algorithm.*

1. INTRODUCTION

Automatic speech recognition (ASR) systems are very popular nowadays. They are based on different approaches. Some of them are related to isolated phonemes, syllables or words and some of them are related to continuous speech. Also they are divided to speaker independent and speaker dependent systems.

In order to do comparison between speech patterns some referential data must be provided. So, this paper is related to problem how to create a database [1] [2] with speech patterns. In this case the question is: "How to collect phonemes for Serbian language speech and how to organize them in the database?"

The phonemes of "Azbuka" is recorded in a special acoustic room where noise is suppressed. For this project 20 volunteers (students from Cacak Technical College) are participated. All recordings are labeled on specific way ease for later use. On some elements of database (vowels) initial DTW (Dynamic Time Warping) test is conducted and results are presented here.

This paper is organized on the following way: Section 2 explains how data are recorded, and what kind of equipment is used. Section 3 explains how data are labeled and how they are stored in "Phonemes_1.0" database. In section 4 we presented some initial test of vowels recognition based on pattern-matching technology. The last section is Conclusion and there are remarks related to this project.

2. RECORDING DATA

The "Phonemes_1.0" database was recorded in a quiet laboratory room by using an Optimus omni-directional microphone with good frequency response up to 16 kHz. (Figure

1.) and lap-top computer Fujitsu-Siemens Esprimo Mobile with Adobe Audition 1.5 software for speech recording.

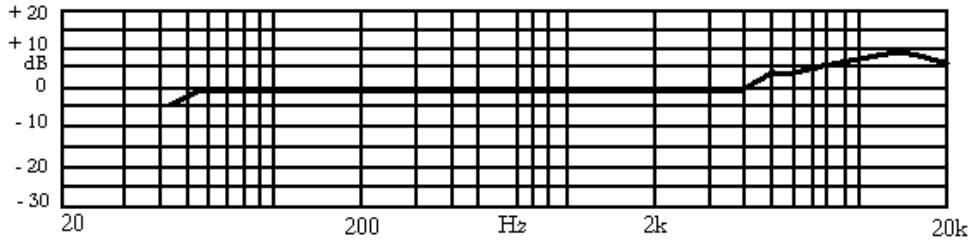

Figure 1. *Frequency characteristic for Optimus microphone*

The microphone was at a distance of about 25 cm from the mouth of a speaker. The speech was digitized by using the sampling frequency of 22.050 Hz, with 16 bits per sample, and stored in the form of Windows PCM wave files.

The sessions of recording were organized four times so as to collect a sufficient number of good quality representatives (two of four were eliminated). During a single session speakers had read 30 phonemes of "Azbuka" two times. Then the whole set of recordings was segmented manually and the quality control applied to it. If the examined patterns was satisfactory, it was labeled and stored in the "Phonemes_1.0" database; otherwise, it was eliminated. It is on this basis that a collection of more than 1.200 phonemes was generated, but only 1.200 of them were stored in the "Phonemes_1.0" database.

The quality control of recordings found various type of error. Some of them were related to an incorrect articulation, a wrong pronunciation, blown microphone etc. Multiple new recordings were sometimes required to solve these problems.

The patterns stored in the database were divided in two sub-corpora: vowels (5 patterns) and consonants (25 patterns). They are presented in Table 1 [4] with the IPA notation for each of them.

Table 1. *Phonemes captured in "Phonemes_1.0" database (with IPA notation)*

Type	Phoneme	IPA	Type	Phoneme	IPA
vowel	/a/	/a/	cons.	/љ/	/ ʎ /
vowel	/e/	/e/	cons.	/м/	/m/
vowel	/и/	/i/	cons.	/н/	/n/
vowel	/o/	/o/	cons.	/њ/	/ ɲ /
vowel	/y/	/u/	cons.	/п/	/p/
cons.	/б/	/b/	cons.	/р/	/r/
cons.	/в/	/v/	cons.	/с/	/s/
cons.	/г/	/g/	cons.	/т/	/t/
cons.	/д/	/d/	cons.	/ћ/	/ tɕ /
cons.	/ђ/	/ dʑ /	cons.	/ф/	/f/
cons.	/ж/	/ ʒ /	cons.	/х/	/h/
cons.	/з/	/z/	cons.	/ц/	/ ʦ /
cons.	/j/	/j/	cons.	/ч/	/ tʃ /

cons.	/к/	/k/	cons.	/ц/	/ *dʒ* /
cons.	/л/	/l/	cons.	/ш/	/ *ʃ* /

From the aspect of speech recognition the vowels are more interesting then consonants because they are more frequent in speech and they can stay alone.

For one particular speaker we provided waveforms for each of five vowels in Figure 2.

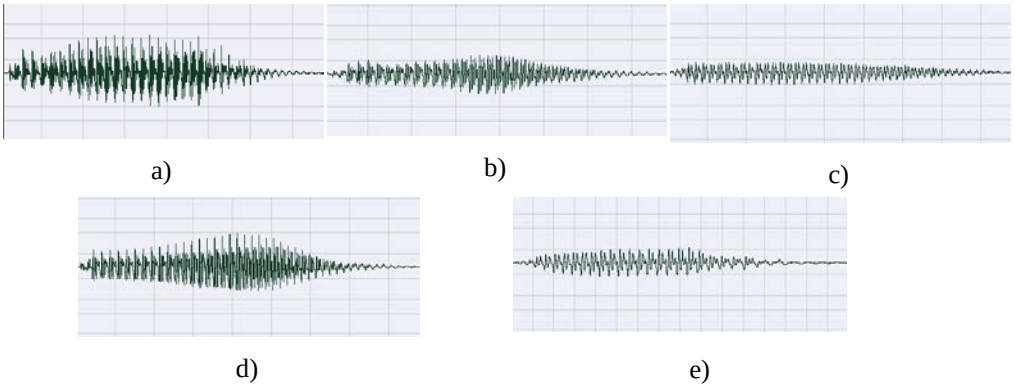

Figure 2. *Waveforms for vowels a) for /a/, b) for /e/, c) for /i/, d) for /o/ and e) for /u/*

From this figure we can see the most shapes of waveforms are similar. But when some methods for spectral analysis are applied the specters will appear different for all of them.

3. LABELING DATABASE

In order to make easy and automatic way for test with this database appropriate labeling must be performed. The labels are chosen to be self-descriptive. So, for vowels all wave files are labeled in the following way: ***vn_m_p.wav***. Letter "***v***" indicates vowel and "***n"***, "***m"*** and "***p"*** are numbers with the following meaning:

- ***"n"*** is a number which indicates an order of vowels (1 - means /a/, 2 - means /e/ etc.)
- ***"m"*** is a number which indicates speaker (1 - means the first speaker, 2 - means the second speaker etc.)
- ***"p"*** is a number that indicates the number of utterance of the same speaker (1 - means the first utterance, 2 - means second utterance etc.)

Using this principle we also generated labels for consonants. So, consonants have the following names: ***cn_m_p.wav***. Meaning of "***n"***, "***m"*** and "***p"***is identical as it explained for vowels.

4. AN INITIAL DTW TEST

In order to evaluate data in this database some initial tests are conducted. The goal of these tests is to see how this database can be used for automatic speech recognition (from the aspect of phonemes) and what will be the recognition rate for vowels.

As a front-end for this ASR the LPC (Linear prediction coding) features are used [5] with the order of autocorrelation ***p=12***. As a back-end for this comparison the DTW algorithm is

used [6].

The DTW algorithm is based on dynamic programming and the goal is to find an optimal path between the starting and ending points of two pattern representations. The speech patterns are represented by a set of vectors. The first set of patterns (5 vowels) is used as a reference, and the other patterns (nine sets, each consisting of 5 vowels) are test data. For local constraints the type I proposed by Sakoe and Chiba [7] is used where a diagonal step is preferred. Global constraints are not used. The system was not trained.

The results in form of the word recognition rates are given in Table 2. The diagonal of this matrix shows a number of successful recognition.

Table 2. *Word recognition rate for vowels with confusion matrix*

Ref/Test	/a/	/e/	/i/	/o/	/u/
/a/	7				1
/e/	2	5	1		
/i/		4	7	1	
/o/			1	6	
/u/				2	8
Average	77.78	55.56	77.78	66.67	88.89
Summary	**73.33**				

Based on result from Table 2 vowels /e/ and /o/ had lower scores. The best result is for vowel /u/ . The average recognition rate was 73.33%.

5. CONCLUSION

This paper gives one of ways how to create speech database in this case it is for Serbian language and for phonemes from “Azbuka”. Using appropriate techniques and labeling the database should be well organized, easy for access and convenient for use.

For test purposes different algorithms can be applied. Here, with LPC as a front-end and DTW as a back-end we showed how to do a particular test and how to obtain the word recognition rates for vowels. Similar scenario can be used for consonants and can be applied to whole words.

Further work will be focus on these areas.

REFERENCES

[1] B. Marković, S.T. Jovičić, J. Galić, Đ. Grozdić: “Whispered Speech Database: Design, Processing and Application”, 16th International Conference, I. Habernal and V. Matousek (Eds.): TSD 2013, LNAI 8082, Springer-Verlag Berlin Heidelberg, pp. 591-598. (2013).

[2] S. Itahashi, “A Japanese Language Speech Database”, ICASSP 86, Tokzo, pp. 321-324.

[3] L. Rabiner, B-H. Juang, “Fundamentals of speech recognition”, (Prentice Hall, New Jersey) (1993).

[4] S. T. Jovičić, "Govorna komunikacija – fiziologija, psihoakustika i percepcija", Nauka, Beograd, 1999.
[5] B. R. Marković and Đ. T. Grozdić, „The LPCC-DTW Analysis for Whispered Speech Recognition", Proceedings of 1st International Conference of Electrical, Electronic and Computer Engineering, IcETRAN 2014, pp. AKI1.1.1-4, Vrnjačka Banja, Serbia, June 2-5, 2014.
[6] G. Marković, B. Marković, "Vizuelni DTW kao nastavno sredstvo za poređenje govornih uzoraka", Tehnika i informatika u obrazovanju, TIO '08, str. 409-415, Tehnički fakultet, Čačak, 9-11. maja.
[7] H. Sakoe and S. Chiba, „Dynamic programming optimization for spoken word recognition", IEEE Trans. Acoustics, Speech, Signal Proc., pp 43-49, 1978.