Semisupervised production of speech corpora using existing recordings

Bartosz Ziółko, Bartłomiej Miga, Tomasz Jadczyk
Department of Electronics
AGH University of Science and Technology
Kraków, Poland
bziolko@agh.edu.pl
www.dsp.agh.edu.pl

***Abstract*—Software to generate professional speech corpora using audiobooks and corresponding to them texts is presented along with first small corpus created using it. The software allows much faster and cheaper production of speech corpora then traditional methods.**

***Index Terms*—speech corpora, Polish, speech recognition**

I. Introduction

Speech corpora are necessary to train models for several speech technologies applications like automatic speech recognition [1], [2], [3] or speech synthesis [4], [5], [6]. Their production is a very expensive and time consuming process which limits making such applications, especially for less popular languages like Polish [7]. Standard approach to making speech corpora needs recording equipment (often a studio), hiring speakers, technical assistants and phoneticians to provide transcriptions [8].

II. Concept

Our approach is too limit as many costs as possible, even accepting some flaws of the corpora. Audiobooks and existing recordings from seminars and conferences can be used instead of recording own audio. It reduces costs significantly, however, limit our choice of content of the corpus as well, both speakers and content of their speech. For many such data there are texts already available - papers or books. It reduces making a speech corpus to fitting these two data types. That is what our software allows.

III. Software

Our software 1 offers several other supports to a user. For example it sets start of a next phoneme if an end of a previous one was picked by a user. It is possible to modify the text file if the user decides it does not fit audio. Transcriptions can be exported as MLFs (Master Label Files) [9].

IV. Testing production of a speech corpus

The software was tested on around an hour of recordings which were practising of a presentation about automatic speech recognition and other speech technologies based on a prepared paper. The audio files were around sixty minutes long. Preparing one minute of a corpus took less then twenty minutes work of one person in average. The corpus is ready to use. Unfortunately, at the moment it has recordings of only one speaker. It is in Polish and using vocabulary about speech technologies. Below a fragment of produced MLF is presented:

```
#!MLF!#
"C:/Users/Bartek/Desktop/Nagrania/10a2.wav"
53420000 57750000 Podmiana
58030000 59940000 tego
60530000 65120000 typu
85830000 88490000 może
88490000 93720000 nastąpić
93720000 94210000 w
94210000 97740000 wyniku
97740000 102450000 błędnego
102450000 111010000 wypowiedzenia
121950000 122980000 i
```

All recordings are mono. They were saved in 32-bit standard with 16 [kHz] frequency 2.

V. Plan of further development

The next step in this software development is to add forced alignment to allow semisupervised segmentation into phonemes. The system will take a phonetic transcription of a particular word from a dictionary. Then it will use two automatic speech recognition systems to fit audio into the phonetic representation of the chosen word. The first system will be HTK [9] and the second will be our own system [7]. Both versions will be presented graphically on a screen with a shadow being a difference between them. This will help a user to focus on boundaries which were detected differently by both systems and trust the decisions in which they agree. As a result two version of MLFs will be created - transcriptions into words and into phonemes.

The next version will allow also automatic segmentation into words with an option of human corrections. This process will be more complex because the system has either work on large vocabulary or detect the end of a word on its own.

VI. Conclusions

The created software is a relatively simple but very useful tool for developing speech corpora, which are necessary and

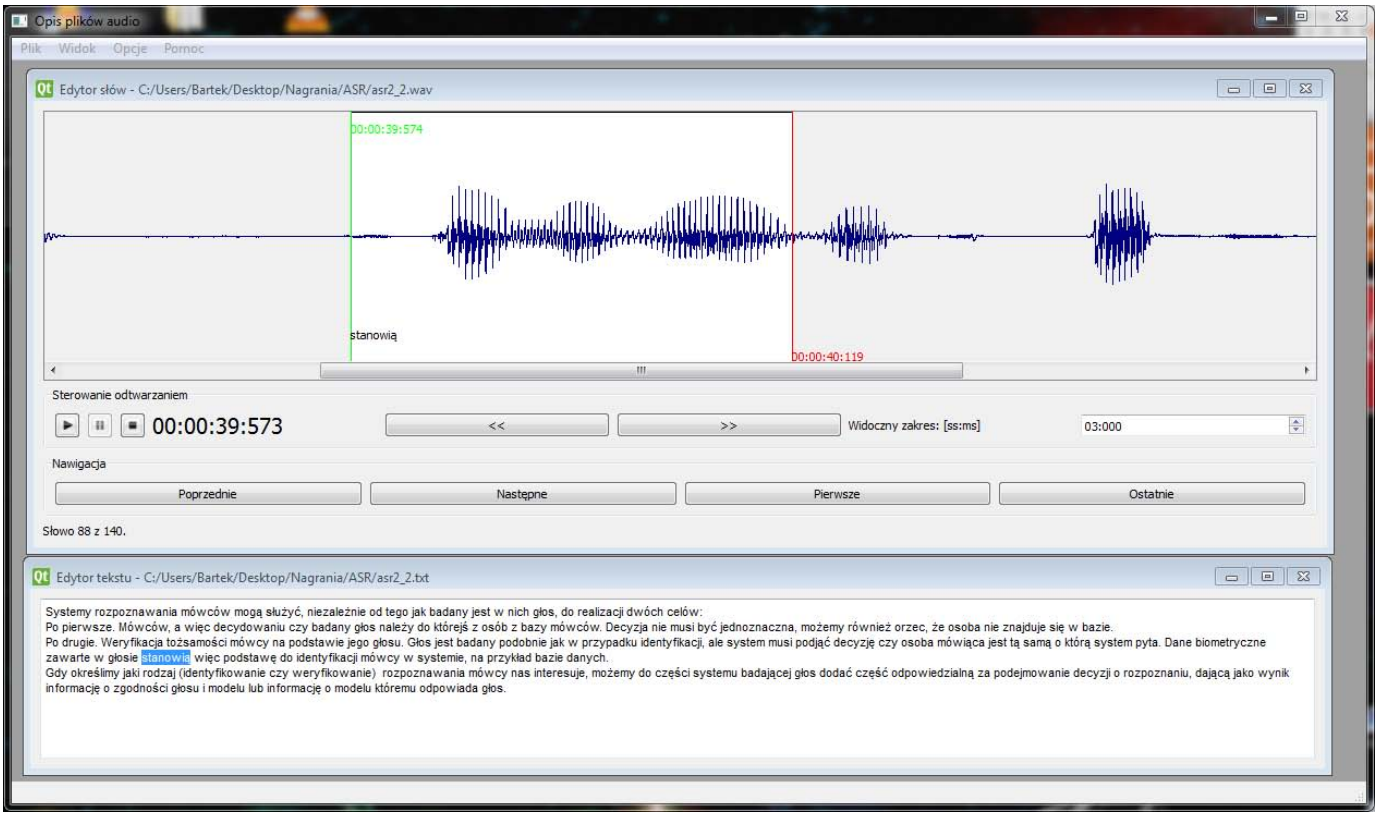

Fig. 1. Print screen of the developed software

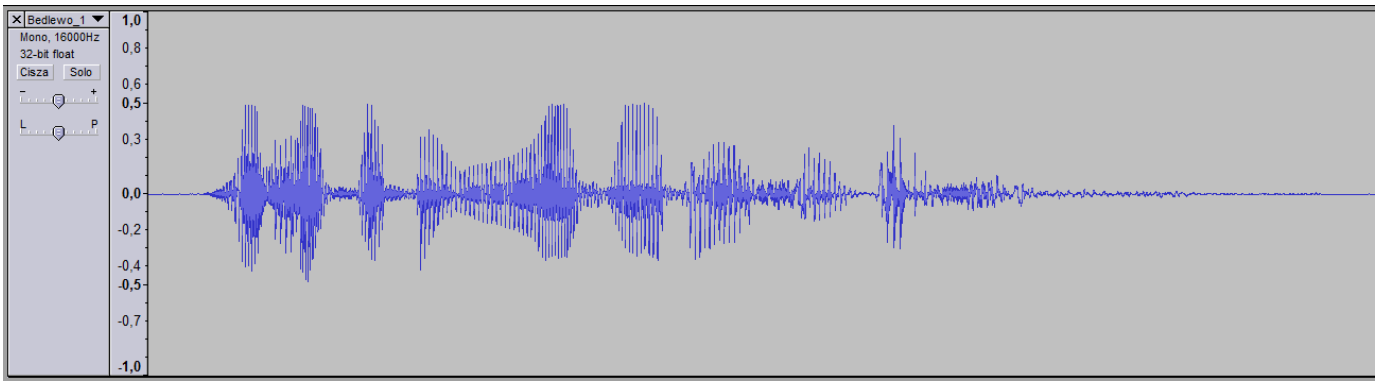

Fig. 2. Visualisation of the recordings shows that they were produced without visible noise

normally expensive resources. The created transcriptions will be on sale or freely available on a website depending on a type of further funding source of this project.

REFERENCES

[1] S. Young, "Large vocabulary continuous speech recognition: a review," *IEEE Signal Processing Magazine*, vol. 13(5), pp. 45–57, 1996.

[2] L. Rabiner and B. H. Juang, *Fundamentals of speech recognition*. New Jersey: PTR Prentice-Hall, Inc., 1993.

[3] W. Huang and R. Lippman, "Neural net and traditional classifiers," *Neural Information Processing Systems, D. Anderson, ed.*, pp. 387–396, 1988.

[4] S. King, "Dependence and independence in automatic speech recognition and synthesis," *Journal of Phonetics*, vol. 31, no. 3-4, pp. 407–411, 2003.

[5] W. Daelemans and A. van den Bosch, "Language-independent data-oriented grapheme-to-phoneme conversion," *Progress in Speech Synthesis, New York: Springer-Verlag*, 1997.

[6] A. Hunt and A. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, 1996. ICASSP-96*, vol. 1, pp. 373 – 376, 1996.

[7] M. Ziółko, J. Gałka, B. Ziółko, T. Jadczyk, D. Skurzok, and J. Wicijowski, "Automatic speech recognition system based on wavelet analysis," *Proceedings of 2010 IEEE International Conference on Semantic Computing*, 2010.

[8] G. Demenko, S. Grocholewski, K. Klessa, J. Ogórkiewicz, A. Wagner, M. Lange, D. Śledziski, and N. Cylwik, "JURISDIC Polish speech database for taking dictation of legal texts," *Proceedings of the International Conference on Language Resources and Evaluation*, pp. 1280–1287, 2008.

recording	time (min:sec)	number of word tokens	number of words	number of sentences
asr2_1.wav	01:57	170	228	25
asr2_2.wav	01:07	100	140	17
asr2_3.wav	01:12	106	152	11
asr2_4.wav	02:21	172	282	29
asr2_5.wav	01:40	137	213	21
asr2_6.wav	01:05	105	130	13
asr2_7.wav	01:40	121	176	16
asr2_8.wav	01:43	143	196	27
asr2_9.wav	01:19	110	136	12
asr2 TOTAL	14:04	733	1653	171
Galka1_2a.wav	01:06	109	129	17
Galka1_3a.wav	02:07	175	235	37
Galka1_4b.wav	01:32	141	173	18
Galka1_5a.wav	03:30	211	306	29
Galka1_6a.wav	03:11	227	321	20
Galka1_7b.wav	03:17	260	365	37
Galka1_8b.wav	01:49	155	210	23
Galka1 TOTAL	16:32	923	1739	181
Matryce1.wav	01:05	106	135	11
Matryce2.wav	00:40	62	70	12
Matryce3.wav	00:43	67	77	10
Matryce4.wav	00:46	74	87	14
Matryce5.wav	01:07	102	120	12
Matryce6.wav	01:20	118	148	15
Matryce7.wav	00:51	86	109	17
Matryce8.wav	00:41	69	86	13
Matryce TOTAL	07:13	484	832	104
Bedlewo2010.wav	22:06	922	1632	197
Bedlewo2010.wav	18:15	789	1465	163
Bedlewo TOTAL	40:21	1483	3097	360
TOTAL ALL	**01:18:10**	**2258**	**7321**	**816**

TABLE I
LIST OF RECORDINGS CONSTITUTING THE FIRST CORPUS CREATED WITH THE DESCRIBED TOOL

[9] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *HTK Book*. UK: Cambridge University Engineering Department, 2005.