Semi-supervised Annotator of Speech Corpora and AGH Speech Corpus of Polish

Bartosz Ziółko[1], **Bartłomiej Miga**[1], **Tomasz Jadczyk**[1] *

[1]Department of Electronics
AGH University of Science and Technology
al. Mickiewicza 30, 30-059, Kraków, Poland
www.dsp.agh.edu.pl

bziolko@agh.edu.pl, bartlomiej.miga@gmail.com, jadczyk@agh.edu.pl

Abstract. *Software to generate professional speech corpora using audiobooks and corresponding text books is presented. The software allows the creation of speech corpora much faster and cheaper than traditional methods. Existing speech resources of Polish are described with a brief introduction to Polish dialects. An example of a small corpus of Polish made with the described tool is presented as well.*

1. Introduction

Speech corpora are necessary to train models for speech technology applications like Automatic Speech Recognition (ASR) (Young, 1996; Rabiner and Juang, 1993; Huang and Lippman, 1988) or speech synthesis (King, 2003; Daelemans and van den Bosch, 1997; Hunt and Black, 1996). Creating a corpus is a very expensive and time consuming process which limits speech technology applications, especially for non-major languages like Polish (Ziółko et al., 2010). The standard approach to making speech corpora requires usage of recording equipment (often in a studio) and the employing of speakers, technical assistants and phoneticians to provide transcriptions (Demenko et al., 2008).

2. Concept

Our approach is to limit costs as much as possible, even accepting some flaws of the corpora. Audiobooks and existing recordings from seminars and conferences can be used instead of our own. It reduces costs significantly, however, on the other hand, it limits our choice of speakers and content of their speech in a corpus. Typically, there are also corresponding texts already available – papers or books. It reduces the work to fitting the speech with its text form to create a speech corpus. That is what our software allows.

3. Software

Our software (Fig. 1) offers several other supports to a user. For example, it sets a start point of the next phoneme if the end point of a previous one was picked out by a user.

*This work was supported by MNISW grant number OR00001905.

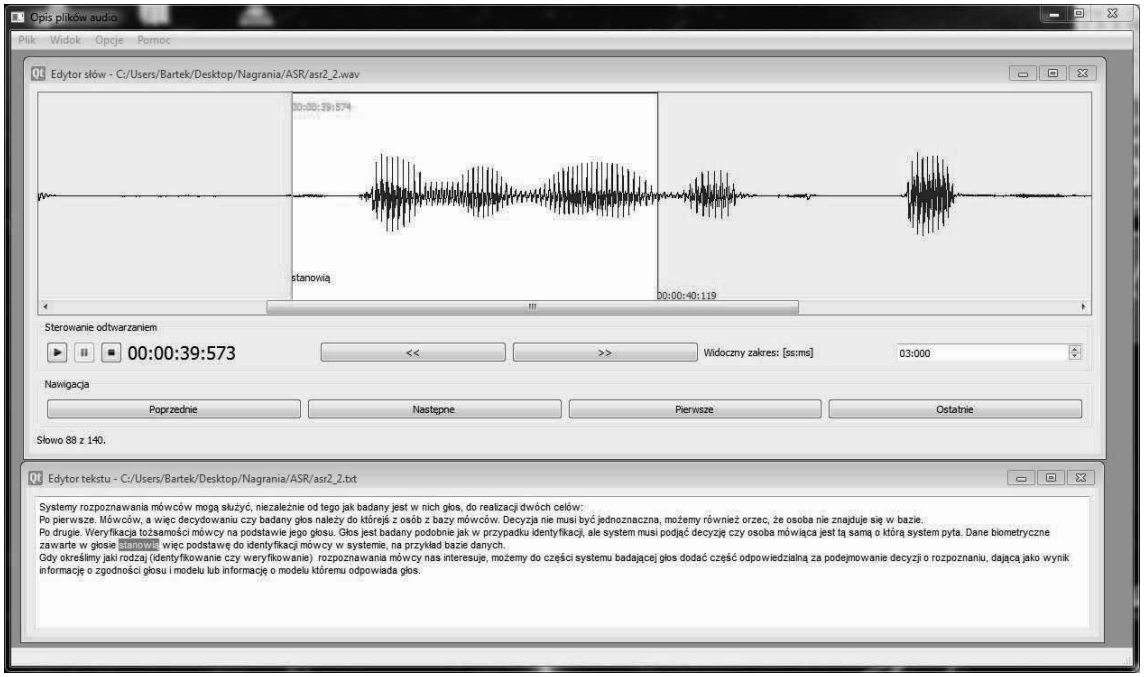

Figure 1. Print screen of the developed software

The next feature is playing (and pausing) a selected part of a *wave* file. Defining keyboard shortcuts by the user is also an advantage. It is possible to modify the text file if the user decides that it does not fit the recorded speech. Transcriptions can be exported as MLFs (Master Label Files) (Young et al., 2005) which is a commonly used standard.

4. Existing resources for Polish

Polish is an averagely common language with around 60 million speakers. For this number of people, speech corpora are quite scarce and inaccessible, which limits the possibilities of developing ASR systems.

Polish dialects are not very frequently used. Their diversity is rather small compared to other languages, mainly because of very little colonisation done by Poland, very strong migration after the World War II and political repressions of all minorities during communism time. This is why the dialect issues are not as important in speech processing applications for Polish as they are in the case of, for example, English, for which several dialects where developing separately for centuries with little influence on each other. However, there are several distinguishable main dialects (Fig. 2), (Urbańczyk, 1991):

- Silesian (Śląski) – sometimes referred to as a separate language,
- Wielkopolski,
- Małopolski (including several local dialects: Krakowski, Podhalański, Sądecki, Żywiecki and Łowicki),
- Mazowiecki (including Białostocki dialect),
- Chełmińsko-kociewsko-warmiński,
- Norther Borderland (Północnokresowy),

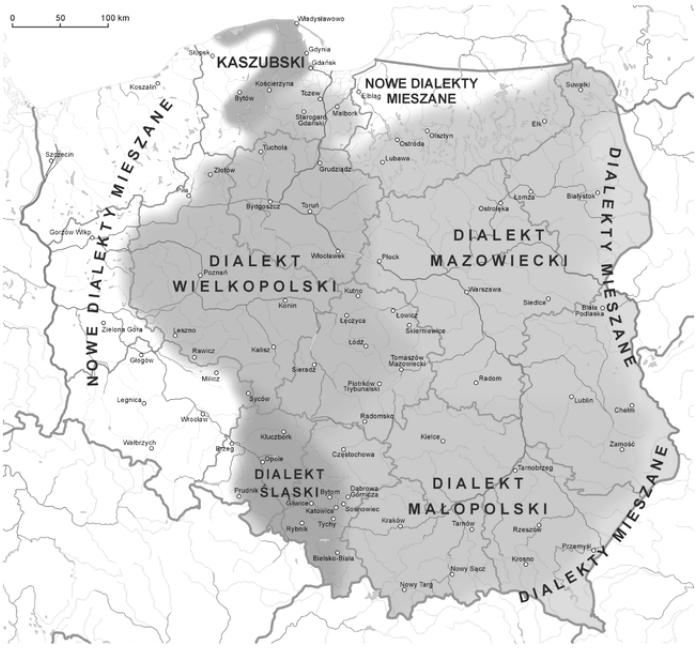

Figure 2. Map of Polish dialects from Wikipedia, according to K. Dejna

- Southern Borderland (Południowokresowy).

In some parts of the country (especially Northern and Western) there was a considerably large migration of people after World War II. As a result, there are no dialects there, and the languages of people who live in these parts of Poland are known as mixed dialects. Further disappearance of dialects is caused by mass media, especially television. Because of TV, the unification creates a tendency for the Mazowiecki dialect (being used in the capital – Warsaw) to spread all over the country. There are some other language changes observed thanks to a rapidly growing rate of migration of Poles (including in-country movements).

The diversities of languages and impact of recording conditions make obtaining a corpus which is balanced in dialects and conditions, quite difficult. There are several existing speech corpora of Polish:

- Grocholewski CORPORA (Grocholewski, 1995) – the oldest and most often used corpus of Polish speech. Speech files were recorded with the sampling frequency 16 [kHz] equivalent to a sampling period 62.5 [μs] in an office with a working computer in the background. It makes the corpus not perfectly clean, however, the signal to noise ratio is not stated in the description of the corpus. CORPORA consists of 365 utterances (33 single letters, 10 digits, 200 names, 8 simple computer commands, and 114 short sentences), each spoken by 11 females, 28 males, and 6 children (45 people), giving 16,425 utterances in total. Pronunciation is careful and speech corresponds to transcriptions perfectly.
  One set spoken by a male and one by a female were hand-segmented into phonemes. The rest were segmented by the dynamic programming algorithm, which was trained on hand-segmented ones, and manually checked afterwards. The quality of all transcriptions can be assumed to be as good as hand-made transcription. The corpus is distributed between several Polish universities but the

official price is high.
- Jurisdic (Demenko et al., 2008) – a very large corpus created with speech of all dialects, segmented into words with a strong representation of lawyer language. At the time of writing (April 2011), Jurisdic is not available for public use.
- GlobalPhone (Vu et al., 2010) consists of about 20 hours of speech spoken by almost 100 native speakers (each person speaks different utterances). The sampling rate is 16 [kHz] and resolution 16 [bit]. Recordings were done using a close-talking microphone and some sound artefacts were observed. Transcriptions were not corrected in a few cases when a speaker pronounced a word wrongly. Each *wave* file is a sentence, automatically made word annotation can be provided by the authors. The corpus is offered at a reasonable price.
- Luna (Marciniak, 2010) – telephone conversation corpus with transcriptions without time annotations (but several other details). The corpus is distributed for free with the book (Marciniak, 2010).
- SpeechDat(E) (http://www.fee.vutbr.cz/SPEECHDAT-E/) – a very expensive corpus; the authors of this paper had no chance to use and test it.
- EPPS European Parliament corpus (RWTH Aachen University, (Lööf et al., 2009)) includes, among others, over 130 hours of Polish of which around 3 hours is transcribed. The quality of recordings is very high. The transcriptions are for whole phrases only. The corpus can be negotiated for exchange with RWTH.
- Szklanny corpus (PJWSTK) – 1443 sentences designed to cover as many diphones as possible, one speaker only. It contains transcriptions into phonemes. There is no official price of the corpus, but during our negotiations, the suggested one was quite high.
- AGH corpus – described in the next section.

5. AGH speech corpus

The software was tested on around an hour of recordings, which was a speech about ASR and other speech technologies from a prepared text. The audio files are seventy eight minutes long. The speaker was male. This part of the corpus consists of vocabulary related to speech technologies. It took less than twenty minutes of the work of one man to prepare one minute of a corpus. Below, a fragment of produced MLF is presented:

```
"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
```

The rest of the corpus consists of several other speakers' utterances, typically up to four minutes of recordings, each. They were taken from students' projects and were

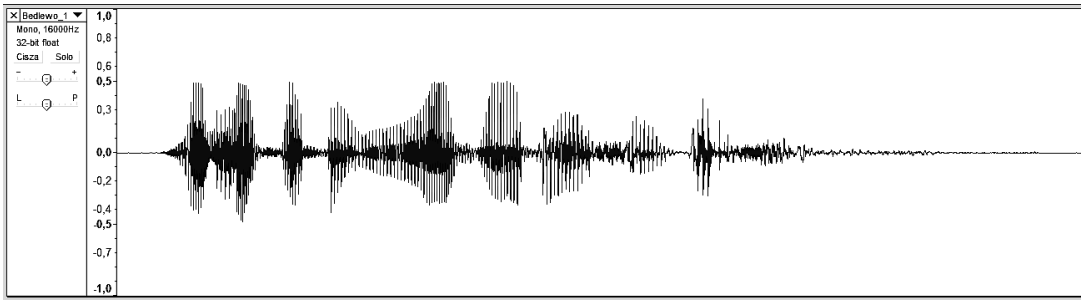

Figure 3. Visualisation of the recordings shows that they were produced without distinguishable noise

recorded under different conditions, using different hardware, including professional ones like a students' radio studio. These recordings are not related to audiobooks and the original idea of creating the corpus, but were added as an offshoot of student laboratories to enlarge the AGH corpus. The student recordings are about various topics. All recordings are mono versions. They were recorded in 16 [bit] standard with 16 [kHz] sampling frequency (Fig. 3). The speakers are people from Krakow, or living in Krakow, of both genders.

The corpus is available (along with the software) with a licence from AGH or in exchange for other resources. Apart from the training data described in Tables 1 and 2 it contains over 30 minutes of test data without time annotations.

6. Plan of future system enhancements

The next step in the software development is to integrate forced alignment to allow semi-supervised segmentation into phonemes. The system will take a phonetic transcription of a particular word from a dictionary. Then it will use two ASR systems to fit audio into the phonetic representation of the chosen word. The first system will be HTK (Young et al., 2005) and the second one will be our own system (Ziółko et al., 2010). Both hypotheses will be presented graphically on a screen with a shadow being the difference between them. This will help a user to focus on boundaries which were detected differently by both systems and trust the decisions on which they agree. As a result, two versions of MLFs will be created – transcriptions into words and into phonemes.

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

7. Conclusions

The created software is a relatively simple but very useful tool for reducing time for the developement of speech corpora, which is necessary to conduct scientific research or to provide commercial solutions. Speech corpora are normally expensive resources. The created transcriptions (or the whole corpus) will be on sale or freely available on a website depending on the type of further funding source for this project.

Table 1. List of recordings about speech technologies constituting the corpus created with the use of the described tool

recording/speaker	gender	time (min:sec)	no. of word tokens	no. of words	no. of sentences
asr2_1.wav	M	01:57	170	228	25
asr2_2.wav	M	01:07	100	140	17
asr2_3.wav	M	01:12	106	152	11
asr2_4.wav	M	02:21	172	282	29
asr2_5.wav	M	01:40	137	213	21
asr2_6.wav	M	01:05	105	130	13
asr2_7.wav	M	01:40	121	176	16
asr2_8.wav	M	01:43	143	196	27
asr2_9.wav	M	01:19	110	136	12
asr2 TOTAL	M	14:04	733	1653	171
Galka1_2a.wav	M	01:06	109	129	17
Galka1_3a.wav	M	02:07	175	235	37
Galka1_4b.wav	M	01:32	141	173	18
Galka1_5a.wav	M	03:30	211	306	29
Galka1_6a.wav	M	03:11	227	321	20
Galka1_7b.wav	M	03:17	260	365	37
Galka1_8b.wav	M	01:49	155	210	23
Galka1 TOTAL	M	16:32	923	1739	181
Matryce1.wav	M	01:05	106	135	11
Matryce2.wav	M	00:40	62	70	12
Matryce3.wav	M	00:43	67	77	10
Matryce4.wav	M	00:46	74	87	14
Matryce5.wav	M	01:07	102	120	12
Matryce6.wav	M	01:20	118	148	15
Matryce7.wav	M	00:51	86	109	17
Matryce8.wav	M	00:41	69	86	13
Matryce TOTAL	M	07:13	484	832	104
Bedlewo2010.wav	M	22:06	922	1632	197
Bedlewo2010.wav	M	18:15	789	1465	163
Bedlewo TOTAL	M	40:21	1483	3097	360
TOTAL BZ1	M	**01:18:10**	**2258**	**7321**	**816**

Table 2. Students recordings (lines with "?" have hand labelled phoneme transcriptions but lack of word statistics)

speaker	gender	file time (min:sec)	speech time	no. of word tokens	no. of words
AW1	W	04:19	02:47	296	34
EM1	W	03:16	02:25	237	48
EM2	W	14:25	07:04	660	47
IM1	W	04:43	02:14	334	48
JL1	W	03:40	02:31	267	21
KK1	W	04:19	03:05	389	32
KbiKK1	W	20:10	16:56	1532	15
All women	W	54:52	37:02	3715	151
AM1	M	00:14	00:14	?	?
AS1	M	01:28	01:25	138	24
BD1	M	01:10	00:29	59	29
BM1	M	04:19	03:03	390	29
BN1	M	03:33	02:27	239	42
BP1	M	02:06	01:33	150	33
BR1	M	03:46	02:11	208	22
DK1	M	00:17	00:17	?	?
EP1	M	03:58	02:18	346	29
GD1	M	01:22	01:08	84	26
GM1	M	04:50	03:26	312	28
JB1	M	02:05	01:07	120	26
JM1	M	05:26	04:39	648	33
JP1	M	02:45	01:21	114	29
KC1	M	03:37	02:36	332	21
KM1	M	00:22	00:22	?	?
KS1	M	04:19	03:03	301	32
MB1	M	05:05	02:54	427	31
MD1	M	02:01	00:48	78	26
MDB1	M	03:25	01:37	264	32
MF1	M	08:30	01:57	238	31
MK1	M	04:11	02:49	257	32
MK2	M	02:55	00:39	100	22
ML1	M	03:57	02:27	222	30
MM1	M	00:33	00:31	70	14
MO1	M	00:14	00:14	?	?
MR1	M	02:21	01:24	151	29
MW1	M	00:55	00:54	140	21
PD1	M	01:12	01:12	?	?
PF1	M	00:14	00:13	27	11
PS1	M	00:36	00:34	87	27
TW1	M	04:37	03:02	379	31
WW1	M	03:03	02:31	280	36
All men	M	01:29:12	55:11	6161	471
Students	M	**02:21:59**	01:30:08	9876	574
Sawa-TJ	M	11:25	04:22	285	95
TOTAL ALL	M/W	**03:53:39**	**∼03:00:00**	∼ 17 500	∼ 2 500

References

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

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

Grocholewski, S. Założenia akustycznej bazy danych dla języka polskiego na nośniku cd rom (Eng. Assumptions of acoustic database for Polish language). *Mat. I KK: Głosowa komunikacja człowiek-komputer, Wrocław*, pages 177–180, 1995.

Huang, W. and Lippman, R. Neural net and traditional classifiers. *Neural Information Processing Systems, D. Anderson, ed.*, pages 387–396, 1988.

Hunt, A. and Black, A. 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*, 1:373 – 376, 1996.

King, S. Dependence and independence in automatic speech recognition and synthesis. *Journal of Phonetics*, 31(3-4):407–411, 2003.

Lööf, J., Gollan, C., and Ney, H. Cross-language bootstrapping for unsupervised acoustic model training: Rapid development of a polish speech recognition system. *Proceedings of Interspeech, Brighton*, pages 88–91, 2009.

Marciniak, M. *Anotowany korpus dialogów telefonicznych [Eng. Anotated corpus of telephone dialogues]*. Exit, Warszawa, 2010.

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

Urbańczyk, S. *Encyklopedia języka polskiego*. Ossolineum, 1991.

Vu, N. T., Kraus, F., and Schultz, T. Multilingual a-stabil: a new confidence score for multilingual unsupervised training. *Proceedings of IEEE Workshop on Spoken Language Technology, SLT 2010, Berkley*, 2010.

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

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

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