# Semisupervised production of speech corpora using existing records

Bartosz Ziółko, Bartłomiej Miga, Tomasz Jadczyk

www.dsp.agh.edu.pl

AGH

Department of Electronics

POLSKA PLATFORMA BEZPIECZEŃSTWA WEWNĘTRZNEGO

PPBW

**Abstract**

Software to generate professional speech corpora using audiobooks and corresponding to them texts is presented along with a created corpus of Polish. The software allows faster and cheaper production of speech corpora than traditional methods.

**Concept**

Our approach is to limit as many costs as possible, even accepting some flaws of the corpora. Audiobooks and existing records of speeches can be used instead of recording own audio material. For much of such data, there are texts already available - papers or books. It reduces making a speech corpus to fitting these two data types, written and spoken. It reduces costs significantly, however, it limits our choice of corpora content as well, both speakers and their speech.

Anotating **1** minute of corpus records takes **17** minutes of one man work

```
#!MLF!#
"C:/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
```

**Future development of the software**

- Phoneme forced alignment using HTK and our ASR system.
- Both hypotheses will be compared graphically with a shadow being a difference between them to help focusing on boundaries detected differently by both systems (large shadow).
- Two versions of MLFs will be created (words and phonemes).
- Automatic segmentation into words with an option of human corrections.

| Speaker | Time | no. of word tokens | no. of words | no. of sentences |
|---|---|---|---|---|
| BZ- male | 01:18:10 | 2258 | 7321 | 816 |
| TJ - male | 00:11:25 | 95 | 285 | - |
| Male students | 01:27:07 | 471 | 6161 | - |
| Female students | 00:54:52 | 151 | 3715 | - |
| **Total** | **03:51:34** | **~ 2 500** | **~17 500** | **816** |