# TUKE-BNews-SK: Slovak Broadcast News Corpus Construction and Evaluation

Matúš Pleva and Jozef Juhár
Department of Electronics and Multimedia Communications
Technical University of Košice
Letná 9, 042 00 Košice, Slovakia
*E-mail: Matus.Pleva@tuke.sk & Jozef.Juhar@tuke.sk*

## Abstract

This poster presents an overview of the existing acoustical corpuses suitable for broadcast news automatic transcription task in the Slovak language. The TUKE-BNews-SK database created in our department was built to support the application development for automatic broadcast news processing and spontaneous speech recognition of the Slovak language. The audio corpus is composed of 479 Slovak TV broadcast news shows from public Slovak television called STV1 or "Jednotka" containing 265 hours of material and 186 hours of clean transcribed speech (4 hours subset extracted for testing purposes). The recordings were manually transcribed using *Transcriber* tool modified for Slovak annotators and automatic Slovak spell checking. The corpus design, acquisition, annotation scheme and pronunciation transcription is described together with corpus statistics and tools used. Finally the evaluation procedure using automatic speech recognition is presented on the broadcast news and parliamentary speeches test sets.

## Slovak acoustic databases

There are several *Czech, French, Thai, Norwegian, Slovenian, Iberian* or *English* broadcast news corpuses available on LDC, ELRA, etc. but most of them are not freely available. Some of them are available only for the project consortium described in the paper.

- A large Slovak speech database was created as a part of *SpeechDat-E* (II) project (*100 hours* of speech over public switched telephone network A-law compression 8kHz sampling frequency, mainly simple commands, *available as ELRA-S0095*)
- The *MobilDat* database (*100 hours*, similar corpus to SpeechDat recorded over mobile GSM network from different environments during IRKR project, *not publicly available*)
- The *Parliament* speech database (*136 hours* of annotated parliamentary speech from the Slovak parliament with $f_s$=48kHz, contains mainly monologues, *not publicly available,* UI SAV – Slovak Academy of Sciences)
- *APD* project database (*250 hours* of court proceedings, planned speech, contains only monologues, recorded in studio environment with $f_s$=48kHz, *not publicly available*)

Unfortunately no Slovak annotated database consisting of different dialogs, spontaneous speech or live coverage with different background conditions is available for automatic broadcast news processing and spontaneous speech recognition task.

## TUKE-BNews-SK description

- *265 hours* of recorded TV broadcast news
- *178'152 of extracted speech utterances*
- *186 hours of extracted annotated corpus*
- the transcribed utterances in the shows contain not only planned but also a *32.7 hours* of spontaneous speech (F1 – focus condition in the Table 1)
- 187'756 words in the dictionary extracted from 1'691'122 tokens in 166'938 utterances from 11'345 speakers

| | |
|---|---|
| F0 – prepared speech in studio | *94.38 h* |
| F1 – spontaneous speech in studio | *32.70 h* |
| F2 – prepared telephone speech (reduced-bandwidth) | *2.07 h* |
| F3 – speech with music in background (SNR<10dB) | *19.15 h* |
| F4 – speech under degraded acoustical conditions | *43.36 h* |
| F5 – speech performed by a non-native speaker | *1.24 h* |
| FX – combination of the focus conditions listed above (F1-F5) | *21.39 h* |

*Table 1: Focus conditions distribution*

| Speaker gender | Number of *utterances* | Percent from all |
|---|---|---|
| Female | 88 941 | *47%* |
| Male | 99 882 | *53%* |
| **Speaker gender** | **Number of *speakers*** | **Percent from all** |
| Female | 4 195 | *37%* |
| Male | 7 447 | *63%* |

*Table 2: Gender distribution*

## Corpus design and acquisition

- *annotation scheme* constructed from DARPA Hub4 project & LDC corpus building manual compiled during COST-278 project (BN pan-European database for segmentation and speaker clustering algorithms evaluation)
- SAMPA based SpeechDat set was used as the main *phonetic set* with 57 phonemes
- *pronunciation dictionary* was built using our Perl tool which uses reprogrammed & extended Ivanecky (2003) rules
- BN stream *captured* using Technisat AirStar PCI: MPEG1 Audio Layer 2 coded stereo in 128kbit/s and 48kHz sampling rate quality
- *manual* annotations (*no texts was provided*) were realized in the modified *Transcriber* tool (see Fig. 1) with native XML *trs* files (see Fig. 2) and exported *stm* (the NIST Scoring toolkit Sclite format – see Fig. 3)

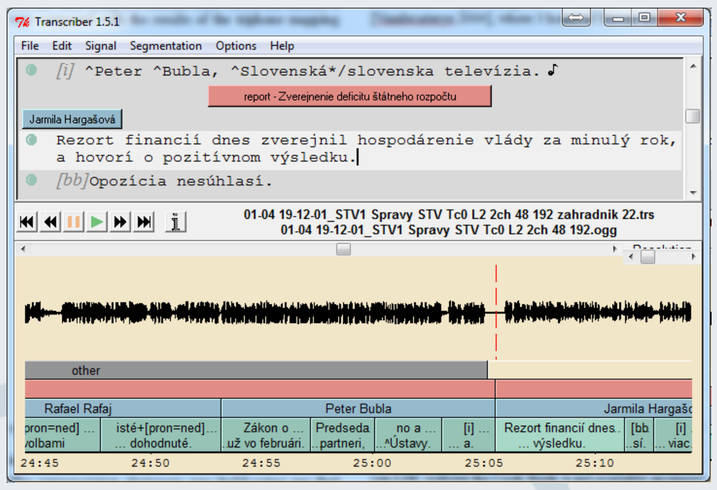


*Figure 1: Example of the Transcriber annotation*

```
<Event desc="i" type="noise" extent="instantaneous"/>
    Tí to však popierajú.
</Turn>
<Turn speaker="spk4" mode="planned" fidelity="high"
channel="studio"startTime="57.783" endTime="76.299">
<Sync time="57.783"/>
    V korupčnej kauze ide o nájomné byty v ^Košiciach
<Sync time="61.329"/>
     ktoré stavala firma ^Kame.
<Sync time="62.985"/>
```

*Figure 2: Example of the Transcriber trs - xml format*

```
stv1_hl_spravy_17 1 Jarmila_Hargašová 55.561 57.783
        <o,f0,female> [i] Tí to však popierajú.
stv1_hl_spravy_17 1 Katarína_Krajňáková 57.783 61.329
        <o,f0,female> V korupčnej kauze ide o nájomné
        byty v ^Košiciach
stv1_hl_spravy_17 1 Katarína_Krajňáková 61.329 62.985
        <o,f0,female> ktoré stavala firma ^Kame.
```

*Figure 3: Example of the stm NIST format*

## Evaluation

For *comparing the impact* of the acoustic similarity between testing and training set, the *acoustic model* (AM) based on Parliamentary speech (136h) and TUKE-BNews-SK (182h) database was used (Table 3):

| WER [%] | BN AM | Parliament AM |
|---|---|---|
| **BN test set** | *10.09* | 13.59 |
| **Parliament test set** | 17.28 | 12.62 |

*Table 3: Comparison test results*

*The BN Language model* built from Slovak text corpus ($10^9$ tokens, adapted to BN task) used with the open source *Julius recognition engine* for automatic speech recognition test.

- *BN test set*: 4h (4343 sentences) of TUKE-BNews-SK corpus extracted for testing
- *Parliamentary test set*: 1.3h (884 sentences) from database compiled on UI SAV

## Conclusion

Our goal was to develop a Broadcast News speech database for Slovak BN and spontaneous speech processing. Currently we are working on the web online automatic multimedia indexing database which will be available for the public (bn.kemt.fei.tuke.sk), where new media files could be uploaded and after automatic transcription process the subtitles for the corresponding media will be available. The resulting audio or video file could be played together with captions in optional karaoke format and edited afterwards.

Európska únia
Európsky fond regionálneho rozvoja
Operačný program VÝSKUM a VÝVOJ
Agentúra Ministerstva školstva, vedy, výskumu a športu SR pre štrukturálne fondy EÚ
We support research activities in Slovakia / This project is being co-financed by the European Union

## Acknowledgements

The poster is the result of the projects implementation under the codes ITMS-26220220141 (50%), ITMS-26220220182 (25%) & ITMS-26220220155 (25%) supported by Research and Development Operational Program funded by the ERDF.

*LREC 2014: The 9th edition of the Language Resources and Evaluation Conference, 26-31 May, Reykjavik, Iceland*
LREC 2014 Reykjavik