BUILDING EUROPEAN BROADCAST NEWS DATABASE

Ing. Matúš Pleva
supervisor: prof. Ing. Anton Čižmár, CSc.

Department of Electronics and Multimedia Communications,
Technical University of Košice, Park Komenského 13, 041 20 Košice,
Slovak Republic.
Tel: +421-55-6022334/2298
E-mail: Matus.Pleva@tuke.sk

ABSTRACT

This paper describes building of an European multilingual multimodal audio and video database of broadcast news shows and its transcriptions.

At this time, the database consists of broadcast news shows in seven languages, namely Dutch, Portuguese, Galician, Czech, Slovenian, Slovakian and Greek. The database was constructed by seven institutions that are collaborating in the European COST278 action on Spoken Language Interaction in Telecommunications [1].

The data distribution policy is to make the data freely available for scientific use to institutions that bring in a new data set which is constructed according to the conventions that were used for the present data sets.

1 INTRODUCTION

"The automatic transcription of broadcast news material is an attractive application of speech recognition technology, and LDC[1] has therefore created the Hub4[2] American Broadcast News corpus to support research in this domain. However, since there are large differences between the American and European national broadcasts, seven institutions collaborating in the European COST278 action on Spoken Language Interaction in Telecommunication joined together to compile a pan-European Broadcast News Database" [1].

The objective was not to create a large database for the training of complete transcription systems, but rather a modest database for accommodating system adaptation and development of weakly language dependent parts such as speaker segmentation and clustering modules.

Department of Electronics and Multimedia Telecommunication participate on this initiative. The second initiative of COST278 we participate is MASPER – Multilingual Speech recognition with large telephone databases [2].

2 DATA COLLECTING

Each institution collected a national data set consisting of approximately 3 hours of material per data set. The material was manually transcribed to text form using the Transcriber[3] tool. Also the focus conditions, speaker turns, silence inside a speaker turn, jingle segments and foreign language utterances was described in the transcription files.

The complete database contains 23 hours of data originating from 10 TV stations (public or commercial), and covering 7 European languages, namely Dutch, Portuguese, Galician, Czech, Slovenian, Slovak and Greek.

	BE	**SI**	**SK**	**PT**	**CZ**	**GA**	**GR**
public TV	VRT	RTVSLO1		RTP1 RTP2	CT1	TVG	NET
commercial TV			TA3		Prima Nova		
nr. of shows	6	3	9	6	10	3	3
nr of anchors	8	6	7	6	12	3	5
data size (min)	162	182	190	211	181	225	174
sample-freq (kHz) at digitalisation	16	16	44,1	44.1	44.1	16	22.05

Table 1: Database information

Table 1 shows the names of the TV stations and their public/commercial status, the number of collected shows, the number of different anchor persons appearing in these shows, the total data size and the sampling frequency used for the digitalisation of the audio files [1].

The database also includes the video files archived in Real Media Video format with a resolution of 352x288.

Each national data set was divided into a training set (about two hours) and a test set (about one hour). All data are stored on an ftp server which can be accessed by all the participating institutions. Data files have unique names identifying the TV channel and the date of broadcast. Since there can be different versions of the transcriptions (because errors were discovered), they are kept in different directories whose names reveal the release date.

[1] http://www.ldc.upenn.edu/

[2] http://www.nist.gov/speech/tests/bnr/hub4_96/h4spec.htm

[3] http://www.etca.fr/CTA/gip/Projets/Transcriber/

3 TRANSCRIPTION RULES

The annotation process follows the LDC transcription conventions for HUB4[4]. However, some ambiguities had to be resolved, especially since the 7 annotators had to work independently at different places. The segmentation and transcription rules were finalized during a workshop that was organized at INESC-ID in July 2003.

3.1 Transcription parameters

During the workshop was decided that the following points need to be standardized:

- Channel and fidelity attributes of speaker turns
- Speech utterance segmentation
- Silences inside speaker turns
- Labeling of section blocks
- Identification of jingle segments
- Marking of foreign language utterances

The major speaker turn attributes were channel (studio/ telephone) and fidelity. Fidelity low/medium/high has different meanings for different channel conditions. For the studio speech, high fidelity is used for conversations that take place inside a studio. Medium fidelity refers to speech that is captured in the field. Low fidelity refers to situations where there is noise in the transmission channel. In the case of telephone speech, *high* refers to clear (clean) speech, *medium* to noisy speech that is still easy to understand though, and *low* to speech that is difficult to understand [3]. Table 2 summarizes this scheme.

		Channel	
		Studio [Bandwidth > 4kHz]	**Telephone** [4kHz Bandwidth]
Fidelity	**Low**	Channel noise	Not intelligible
	Medium	Field	Noisy
	High	Studio	Sounds clear

Table 2: Coding of channel and fidelity attributes

As regard annotating of silence blocks marking, when a silence inside a speaker turn is less than 0.5 seconds it is not marked at all. When it is between 0.5 and 1.5 seconds, a breakpoint in the middle of the silence is inserted. When the silence is longer than 1.5 seconds, two breakpoints delimiting the silence are inserted [4].

All jingle segments have to be identified as such and marked by a noise event tag. When the TV station uses different jingles at the beginning and the end of a show, each jingle gets an additional suffix indicating its begin/end category [5].

Foreign language utterances are marked with language event tags and are *not transcribed*.

4 CONCLUSIONS

We found it very useful that we are participating in this database building. Slovak transcriptions were made on our department and therefore we have access to whole multilingual European database. We hope that this database will be a good basis for training multilingual speech recognition and multilingual speech partitioning software.

ACKNOWLEDGEMENT

This research is performed with support of the grant project N^{o} 1/1057/04 of the Grant Agency for Science (VEGA), Slovak Republic and COST 278 (http://cost278.org/).

REFERENCES

[1] Vandecatseye, A. - Martens, J. P. - Neto, J. - Meinedo, H. - Mateo, C. G. - Dieguez, J. - Mihelic, F. - Zibert, J. - Nouza, J. - David, P, - Pleva, M. - Cizmar, A. - Papageorgiou, H. – Alexandris, Ch.: *The COST278 pan-European Broadcast News Database.* Proc. LREC‘ 2004 – 4th International Conference On Language Resources And Evaluation, 2004 (to be published)

[2] Zgank, A. – Kacic, Z. – Moreno, A. – Caballero, M. – Diehl, F. – Vicsi, K. – Szaszak. G. – Juhar, J. – Lihan, S.: The COST 278 initiative – crosslingual speech recognition with large telephone database. Proc. LREC‘ 2004 – 4th International Conference On Language Resources And Evaluation, 2004 (to be published)

[3] Gauvain, J. L. - Lamel, L. - Adda, G.: *Partitioning And Transcription Of Broadcast News Data.* Proc. ICSLP'1998 - 4th International Conference on Spoken Language Processing, 1998

[4] Hain, T. - Woodland, P.C.: *Segmentation And Classification Of Broadcast News Audio*. Proc. ICSLP'1998 - 4th International Conference on Spoken Language Processing, 1998.

[5] Meinedo, H. - Neto, J.: *Automatic Speech Annotation and Transcription in a Broadcast News task.*, 2003, http://citeseer.nj.nec.com/571279.html

4 http://www.ldc.upenn.edu/Projects/Corpus_Cookbook/transcription/broadcast_speech/english/index.html