The COST278 pan-European Broadcast News Database

An Vandecatseye[(1)], Jean-Pierre Martens[(1)], Joao Neto[(2)], Hugo Meinedo[(2)], Carmen Garcia-Mateo[(3)], Javier Dieguez[(3)], France Mihelic[(4)], Janez Zibert[(4)], Jan Nouza[(5)], Petr David[(5)], Matus Pleva[(6)], Anton Cizmar[(6)], Harris Papageorgiou[(7)], Christina Alexandris[(7)]

(1) Ghent University, Sint-Pietersnieuwstraat 41, B-9000 Ghent, Belgium, {avdecats, martens}@elis.ugent.be
(2) INESC ID, Tua Alves Redol 9, 1000- 029 Lisbon, Portugal
(3) University of Vigo, 36200 Pontevedra, Vigo, Spain
(4) University of Ljubljana, Trzaska 25, SI - 1000 Ljubjlana, Slovenia
(5) Technical University of Liberec, Halkova 5, 461 17 Liberec, Czech Republic
(6) Technical University of Kosice, Letna 9, 04120 Kosice, Slovakia
(7) ILSP, Artemidos 6 & Epidavrou, GR-151 25 Maroussi, Greece

Abstract

This paper describes a pan-European multilingual audio and video database of broadcast news shows. The database was constructed by seven institutions that are collaborating in the European COST278 action on Spoken Language Interaction in Telecommunications. At present, the database comprises broadcast news shows in seven languages, namely Dutch, Portuguese, Galician, Czech, Slovenian, Slovakian and Greek, but the policy is to attract new partners that bring in new data which are constructed and transcribed according to the rules and procedures outlined in this paper. The data comes with evaluation software that should facilitate a comparison of experiments.

1. Introduction

The automatic transcription of broadcast news material is a potentially attractive application of speech technology, and LDC has therefore created the Hub4 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. 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. Software for the evaluation of such modules is also provided with the data. By evaluating different algorithms using the same software, and by applying algorithms to a multilingual/ multi-style corpus, it may be possible to assess the strengths and weaknesses of existing approaches, and to conceive better approaches through collaborative research.

The rest of this paper is organized as follows. Section 2 gives a brief description of the data collection and distribution, whereas the transcription of the data is discussed in section 3. In section 4, we discuss the software that is presently available for the evaluation of segmentation and clustering algorithms. Some first evaluation results from different sites are reviewed in section 5. The paper ends with a discussion of future plans and a review of some first conclusions.

2. Data collection and distribution

Each institution collected a national data set consisting of a number of complete news broadcasts from public and/or private TV stations. The goal was to collect approximately 3 hours of material per data set. The complete database presently contains 23 hours of data originating from 10 TV stations, and covering 7 European languages, namely Dutch, Portuguese, Galician, Czech, Slovenian, Slovak and Greek.

Table 1: Database information

	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 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 (in minutes) and the sample frequency that was used for the digitalisation of the audio. Files that were not digitised at 16 kHz were digitally re-sampled by "sox <infile> -r 16000 <outfile> polyphase"[1] before archiving them in wave format (16 bit PCM).

A particular property of the COST278 database is that it also includes the video files. They can help to check the

[1] SoX: http://sox.sourceforge.net

correctness of speaker labels (see section 3), and support the development of multi-modal algorithms. To save memory, the video data were 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. The ftp server is organised according to the following directory structure: /<usage>/<language>/<type>/, where <usage> represents the train or test directory, <language> is the international country code and <type> corresponds to the audio, video and transcriptions respectively. There are also directories for documentation and software. 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. Each transcription directory contains the most recent transcriptions of all data as they were available at the mentioned release date.

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[2].

3. Transcription of the data

The annotation process follows the LDC transcription conventions for HUB4[3]. 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 finalised during a workshop that was organized at INESC-ID, and attended by 6 of the 7 people that later produced the transcriptions.

3.1. Transcription rules

During the workshop it was established that the following points needed to be harmonized:

- 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
- Transcription of interjections and non-lexemes

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. Usually, this is when the anchor person is talking or when a video story is commented by a journalist that is recorded in a studio. Medium fidelity refers to speech that is captured in the field, usually situations where the journalist is making a street interview. 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. Table 2 summarizes this coding 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

The **speech utterances** should not be too long and every speaker inspiration event should be regarded as a potential breakpoint.

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.

The **sections blocks** are categorized as reports (news stories), fillers (headlines and short story descriptions) and nontrans (commercials and jingle segments).

All **jingle segments** are 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.

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

For each language a close set of **interjections and non-lexemes** was defined so as to pursue that the same tags are used for the same words/sounds in all the data of one language. No attempts were made to define a common set of tags for all languages.

3.2. Statistics across languages

Table 3 lists total number of words, number of different words, total number of sentences and percentage of sentences by focus condition (Pallet, 2002).

			% of sentences by focus conditions							
country	# words	different words	F0	F1	F2	F3	F4	F5	Fx	Total
BE	26456	5018	38.6	11.5	0.5	2.9	45.1	0.0	1.4	1857
CZ	27642	8834	52.9	18.2	1.2	5.4	21.7	0.1	0.4	1547
GA	33029	6463	29.5	5.1	0.2	6.5	43.3	4.0	11.4	1673
GR	23748	6065	48.7	16.6	0.0	3.1	31.3	0.1	0.2	1624
PT	33949	5719	10.5	7.5	0.0	2.5	76.4	0.5	2.6	1987
SI	22269	7237	61.9	15.3	3.7	8.6	8.1	0.6	1.7	1292
SK	25770	8887	35.7	15.7	4.1	7.6	34.1	0.0	2.8	2023

Table 3: Database statistics for each language

The significant differences across data sets are believed to reflect the multi-style character of the database.

3.3. Conversion to NIST STM format

All transcriptions were made using Transcriber[4] (Barras et al., 2001) and saved as XML formatted text files using different national character encodings. Since the STM files generated by Transcriber were often incorrect (a failure to support some national code pages),

[2] Interested parties have to contact the second author.

[3] http://www.ldc.upenn.edu/Projects/Corpus_Cookbook/transcription/broadcast_speech/english/index.html

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

we created a tool **trs2stm** for correctly converting the Transcriber XML files to NIST STM files (as used in e.g. the American Hub4). This tool is also made available together with the data.

4. Evaluation software

It appears that the NIST scoring software[5] cannot take time information into account the way we need this for the evaluation of speaker segmentation and clustering algorithms. Therefore, we did what many others did (e.g. Tritschler & Gopinath, 1999; Harris et al, 1999), namely provide our own evaluation software. Our objective was to provide software that can assess different types of segmentation approaches (e.g. with or without speech/non-speech segmentation), and with or without taking acoustic conditions into account. Our software expects an STM file providing the **reference segmentation** and an ASCII file describing the **computed segmentation**. The latter can contain 6 items per segment: start time (sec), end time (sec), speech/non-speech nature, cluster number, gender (male, female, child) and background (electrical noise, music, speech, other, clean).

The present version of the software implements the evaluation strategy of (Vandecatseye & Martens, 2003). There are three executables. The first one **eval_sns** is used to assess the speech/non-speech segmentation, the second one **eval_seg** to evaluate the change point detection (a change point is where a new speaker and/or acoustic condition starts), and the third one **eval_clus** to evaluate the clustering of the parts between change points. The aim is to separate as much as possible the different types of errors. E.g., when evaluating speaker cluster labels, one should be able to discern between errors that were caused by deficiencies in the segmentation and errors that were caused by deficiencies in the clustering of the segments.

4.1. Speech/non-speech segmentation

The speech/non-speech segmentation is evaluated at the frame level. Every frame has an reference and a computed speech/non-speech label. When non-speech segments shorter than a certain threshold value (1.5 seconds in our case), are encountered in either the reference or the computed segmentation, they are removed from that segmentation (replaced by speech). The program then lists the percentages of correctly labelled speech and non-speech frames both in total and per focus condition.

4.2. Change point detection

Many BN transcription systems comprise a change point detector. A **change point** (CP) is a point where a new speaker or acoustic condition starts, and a segment between two subsequent CPs is called a **turn**.

In a first stage computed CPs (from evaluated segmentation) are linked to reference CPs according to the following procedure: link reference CP n_{ri} to computed CP n_{cj} if (i) n_{cj} is the computed CP closest to n_{ri}, (ii) n_{ri} is the reference CP closest to n_{cj}, and (iii) the distance between n_{ri} and n_{cj} is smaller then a certain threshold (set to 1 second in our case). If there is a gap between two turns, the corresponding n_{ri} is allowed to float within that gap so as to minimize the number of errors.

After having linked computed and reference CPs, Recall and Precision are derived. Recall is the percentage of detected reference CPs (they are linked to a computed CP) and Precision is the percentage of correctly computed CPs (they are linked to a reference CP).

4.3. Clustering

A clustering algorithm is presumed to assign a cluster number to each computed turn. All frames of all turns with the same cluster number together constitute a cluster. By looking at the reference speaker/acoustic labels attached to the frames of a cluster, one can determine the cluster label as the best reference label for that cluster.

The quality of a cluster configuration is often given in terms of the total cluster purity P_t (the percentage of frames getting the correct label) and the average cluster purity P_a (the mean percentage of correct labels per cluster) (Harris et al, 1999). However, since cluster purity tends to increase with the number of clusters being hypothesised, this number has to be taken into account as well. Moreover, since the computed turns do not exactly coincide with the reference ones, even the ideal clustering which would assign to each computed turn the dominant reference label found in that turn, will not yield 100% pure clusters. Therefore, if N_c is the number of clusters, and if (P_{ti}, P_{ai}, N_{ci}) are the characteristics of the ideal clustering, the evaluation program also computes

$$D_t = |N_c - N_{ci}| / N_{ci} + (P_{ti} - P_t)$$

$$D_a = |N_c - N_{ci}| / N_{ci} + (P_{ai} - P_a)$$

as more appropriate indicators of the deficiency of the clustering algorithm.

5. First evaluation results

Two speech/non-speech segmentation systems based on GMMs were tested: one system (GU) was trained on American data, the other (TUL) on the COST278 training data. The found percentages of correctly classified speech and non-speech frames, measured on the COST278 test sets, are listed in Table 4. Note the significant differences between the percentages of correctly detected non-speech frames for the different sets.

		BE	CZ	GA	GR	PT	SI	SK
GU	Speech	97.0	98.3	97.9	95.3	98.2	96.7	98.4
	Non-sp	84.6	95.9	73.0	52.5	81.8	85.2	83.7
TUL	Speech	97.2	98.8	97.5	93.5	97.5	97.4	98.3
	Non-sp	82.4	93.0	73.0	62.2	75.7	93.7	87.3

Table 4: Percent frames correct (speech/non-speech)

There were three CP detection systems available for evaluation:

(1) The GU system (Vandecatseye & Martens, 2003) searches for CPs in the speech parts emerging from the speech/non-speech segmentation. It follows a two-step procedure and it uses the Bayesian Information Criterion (BIC) to decide. The parameters of the system were fixed on the American BN database.

[5] NIST. http://www.nist.gov/speech/tools.

(2) The VIGO system (Perez-Freire & Garcia-Mateo, 2004) also uses BIC, but no speech/non-speech partition. It also follows a totally different two-step segmentation strategy. The system was designed on Galician data.

(3) The LJU system comprises uses a variant of the DISTBIC algorithm (Delacourt & Wellekens, 2000). The system was tuned on Slovenian data (Zibert & Mihelic, 2004) that were no part of the COST278 database.

Since none of the systems used COST278 data during the design phase, it was possible to test them on the entire database. Figure 1 shows Recall and Precision for the pure speaker change detections (ignoring acoustics) of the three systems. Each system yields 7 symbols corresponding to the 7 national data sets.

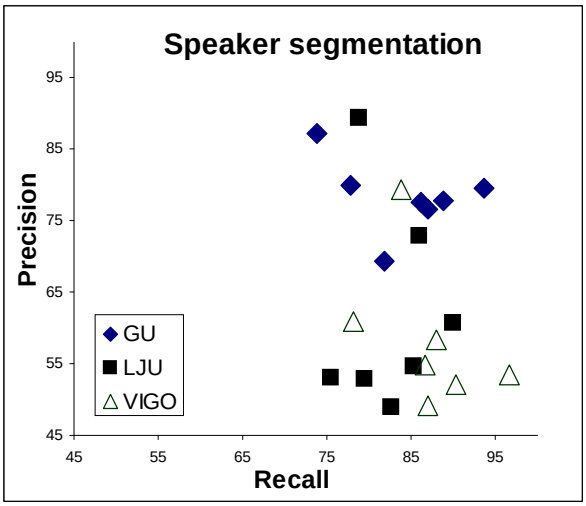

Figure 1: Evaluation of speaker segmentation

Seemingly very similar algorithms yield significantly different Precisions (e.g. from 50 to 85%) for the same data set. All algorithms yield similar dispersions in the Recall values (from about 75 to 95%) across data sets. Both results definitely need further investigation.

As there is at present only one system (GU) that also incorporates speaker clustering, we can only present (Table 5) cluster purities and deficiencies for that system.

	BE	CZ	GA	GR	PT	SI	SK
P_t (%)	93.2	82.0	96.4	93.4	95.2	92.6	91.9
D_t	0.44	0.21	0.28	0.63	0.33	0.48	0.43
P_a (%)	91.8	82.9	95.1	91.1	93.3	90.4	90.8
D_a	0.45	0.19	0.29	0.64	0.34	0.49	0.43

Table 5: Evaluation of clustering part of GU system.

For most languages, N_c is close to 1.4 times N_{ci}, but for CZ, the two numbers are about equal, leading to a low deficiency, even with a cluster purity being far from ideal.

6. Future work

Although the present evaluation software already permits the performance of comparative tests, the general feeling is that it should be further extended and improved . Shortly, there will be a workshop on the topics of evaluation and further planning of experiments. This workshop will be attended by people from all interested institutions participating to COST278.

In the coming year, more segmentation and clustering algorithms will be evaluated on the COST278 database. Then, an in depth analysis of the weaknesses and strengths of the existing algorithms will become possible. The hope is that the collaborative effort which started with the design of a database will soon result in the proposal of better segmentation and clustering algorithms.

7. Conclusion

By joining forces it was possible to construct a modest multilingual/multi-style BN database of audio and video files. The database is first of all found useful for the design and evaluation of speaker segmentation and clustering algorithms. The first evaluation results reveal significant dispersions in change point detection performances that need further investigation. The seven 3 hour national data sets can also be valuable as independent test sets for full transcription systems that were previously trained on other data. By making the data available to institution that bring in new data set (even in a language that is already covered), the present makers hope to have initiated the creation of a very valuable multilingual/multi-style tool for research in BN transcription at large.

8. Acknowledgments

This work is performed in COST278, the European COST action on Speech and Language Interaction in Telecommunications. L. Perez-Freire (Vigo) is thanked for making available his segmentation program.

9. References

Barras, C., Geoffrois, E., Wu, Z. & Liberman, M. (2001). Transcriber: Development and use of a tool for assisting speech corpora production. *Speech Communication*, Volume 33, Issues 1-2, 5-22.

Delacourt, P. & Wellekens, C.J. (2000). DISTBIC: A speaker-based segmentation for audio data indexing. *Speech Communication* 32, Issues 1-2, 111-126.

Perez-Freire, L. & Garcia-Mateo, C. (2004). A multimedia approach for audio segmentation in TV Broadcast News. *Proceedings ICASSP*.

Tritschler A., Gopinath R. (1999). Improved speaker segmentation and segments clustering using the Bayesian Information Criterion. *Proceedings Euro-speech*, 679-682.

Harris M., Aubert X., Haeb-Umbach R., Beyerlein P. (1999). A study of Broadcast News audio stream segmentation and segment clustering. Proceedings Eurospeech, 1027-1030.

Pallett, D. S. (2002). The role of the National Institute of Standards and Technology in DARPA's Broadcast News continuous speech recognition research program, Speech Communication 37, 3-14.

Vandecatseye, A. & Martens, J.P. (2003). A fast, accurate and stream-based speaker segmentation and clustering algorithm. *Proceedings Eurospeech*, 941-944.

Zibert, J. & Mihelic, F. (2004). Development of Slovenian Broadcast News Speech Database. *Proceedings LREC*.