The COST278 broadcast news segmentation and speaker clustering evaluation – overview, methodology, systems, results.

CONFERENCE PAPER · JANUARY 2005

Source: DBLP

The COST278 Broadcast News Segmentation and Speaker Clustering Evaluation - Overview, Methodology, Systems, Results

Janez Žibert[1], *France Mihelič*[1], *Jean-Pierre Martens*[2], *Hugo Meinedo*[3], *Joao Neto*[3], *Laura Docio*[4], *Carmen Garcia-Mateo*[4], *Petr David*[5], *Jan Nouza*[5], *Matus Pleva*[6], *Anton Cizmar*[6], *Andrej Žgank*[7], *Zdravko Kačič*[7], *Csaba Teleki*[8], *Klara Vicsi*[8]

[1]University of Ljubljana, Ljubljana, Slovenia,
[2]Ghent University, Ghent, Belgium,
[3]INESC ID, Lisbon, Portugal,
[4]University of Vigo, Vigo, Spain,
[5]Technical University of Liberec, Liberec, Czech Republic,
[6]Technical University of Kosice, Kosice, Slovakia,
[7]University of Maribor, Maribor, Slovenia,
[8]Budapest University of Technology and Economics, Budapest, Hungary
janez.zibert@fe.uni-lj.si

Abstract

This paper describes a large scale experiment in which eight research institutions have tested their audio partitioning and labeling algorithms on the same data, a multi-lingual database of news broadcasts, using the same evaluation tools and protocols. The experiments have provide more insight in the cross-lingual robustness of the methods and they have demonstrated that by further collaborating in the domains of speaker change detection and speaker clustering it should be possible to achieve further technological progress in the near future.

1. Introduction

The transcription of broadcast news (BN) poses a number of challenges, both in terms of computational complexity and transcription accuracy. Most present day transcription systems perform some kind of audio indexing (segmentation and labeling) as a first step in the processing chain [1]. Usually, the segmentation involves the partitioning of the audio in speech and non-speech intervals, and the further division of the speech intervals in speaker turns. The labeling of speech intervals is usually done in terms of gender and speaker identity (all turns of the same speaker are expected to get a unique label).

Audio indexing offers some practical advantages: no waste of time on the processing of non-speech intervals, no need to process very long speech chunks, facilitation of gender or speaker dependent acoustic model selection, etc. On the other hand, indexing errors may cause extra transcription errors, e.g. if a speaker change is hypothesized in the middle of an utterance, especially in the middle of a word.

In this paper algorithms developed at eight institutions are evaluated on the same multi-lingual data using the same evaluation tools and protocols. The major aim is to assess cross-language dependencies and to identify areas in which a further comparison of algorithmic details is bound to induce further technological progress.

The paper is organized as follows. Section 2 describes the experimental framework, whereas sections 3-6 review and discuss the experimental results. The paper ends with a short summary and some directions for future research.

2. Experimental framework

2.1. Evaluation database

The evaluation database is the pan-European COST278-BN database. At present it consists of 30 hours of news broadcast recordings, divided into ten equally large national data sets. Each national set was recorded and transcribed by one institution and contains some complete news shows broadcasted by TV stations in one country or region. The transcription was performed according to a protocol described in [2].

Since two institutions from Slovenia participated in the data collection, the database presently covers nine European languages: Belgian Dutch (BE), Portuguese (PT), Galician (GA), Czech (CZ), Slovenian (SI), Slovak (SK), Greek (GR), Croatian (HR) and Hungarian (HU).

Due to the limited size of the national data sets they cannot be used for transcription system training, but they are very suitable for the evaluation of acoustic model adaptation methods and audio indexing systems (which are presumed to behave language independently).

2.2. Tasks and tests

The following tasks are being considered: speech/non-speech classification (SNC), gender classification (male/female) (GC), speaker change detection (SCD) and speaker turn clustering (STC). Each tasks is evaluated under two experimental conditions:

C1: training and control parameter tuning is performed on external data and testing is performed on all national sets.

C2: training and control parameter tuning is performed on one national data set and testing is done on the remaining data sets, and this procedure is repeated four times using either BE, GA, PT or SK as the training set.

Figure 1: *Canonical structure of a system for audio data indexing*

The advantage of C2 is of course that everything is under control, whereas under C1, different institutions used different training databases. The advantage of C1 is that it permits a much better training of models, since under C2 the training data is limited to three hours.

2.3. Participants

Eight research institutions participated in this evaluation campaign: ELIS (Gent), INESC (Lisbon), TUB (Budapest), TUK (Kosice), TUL (Liberec), ULJ (Ljubljana), UMB (Maribor) and UVIGO (Vigo). Although all of them participated in task SNC, only three of them participated in all four tasks.

2.4. System architecture and operating mode

Most of the tested algorithms fit into the canonical system architecture depicted on Fig. 1. However, in a few cases the SNC and SCD modules are interchanged. Most systems use a MFCC front-end (with or without delta's), but INESC uses PLPs instead.

Since no children appear in the data, gender classification is restricted to male/female. The speaker clustering is supposed to group all the turns of the same speaker.

Considering the four systems that include both SNC and SCD, three of them operate in batch, meaning that they always have access to the entire audio input in order to make their decisions. The ELIS system [3] works in a real-time, with a maximum look ahead of about 15 seconds.

3. Speech/non-speech classification (SNC)

The SNC is supposed to detect non-speech intervals of at least 1.5 seconds long.

3.1. Algorithmic differences

The ELIS, UMB, TUB and TUK algorithms work directly on the acoustic feature stream, whereas the ULJ, TUL, INESC and UVIGO algorithms also rely on the SCD output.

The ELIS, ULJ, UMB and UVIGO algorithms use GMMs as the acoustic models, whereas TUK, TUL and TUB use HMMs. In all cases, there were models for *speech*, *speech+music*, *speech+other*, *music* and *other*. The INESC system uses a totally different approach (cfr. [5]) involving a phone recognizer and an analysis of frames captured by the phone models. *[JPM: I think Hugo has to provide a compact description of what INESC does here]*

Six institutions performed C1 tests and thus used different training data: ELIS used the Hub-4 American English database, whereas ULJ, UMB, INESC, TUL and UVIGO used Broadcast News (BN) databases in their native language.

3.2. Performance measures

The performance measures are the percentages of frames, speech frames and non-speech frames that were classified correctly. The first one represents the accuracy of the SNC.

3.3. Experimental results

Figure 2 shows large discrepancies in the balance between per-

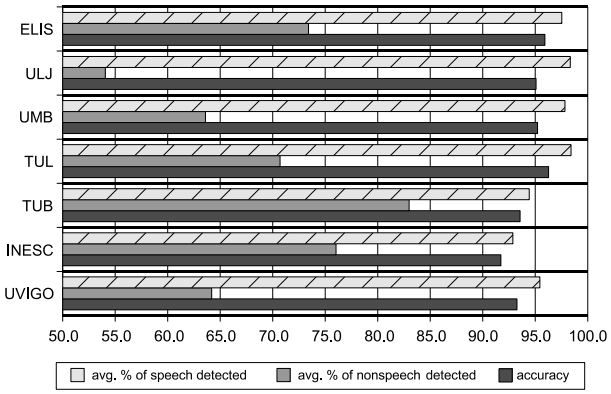

Figure 2: *C1 results for speech/non-speech classification*

cent speech and percent non-speech correct. One reason for this is that different institutions used different criteria for tuning their systems. We will compare algorithms on the basis of their accuracy, acknowledging that systems which were tuned on the basis of that criterion have a (small) advantage then.

Another cause of discrepancies is the composition of the training database. One of the main problems in SNC appears to be the detection of music intervals. A lack of training material of the different kinds of music appearing in the COST278-BN database will hurt the performance.

The C1 tests seem to suggest that interchanging the SNC and the SCD modules does not affect the attainable performance (compare ELIS, UMB to ULJ, TUL). Using GMMs or HMMs does not seem to matter either. A possible hypothesis for the lower accuracy of the INESC system could be that Portuguese phone models may not offer as much language independency as the more generic GMMs or HMMs used in the other systems. However, this hypothesis is not confirmed by the data since there were 4 languages for which the system performed better than for Portuguese.

The accuracies of the four C2 experiments are depicted on Figure 3. Note first that the average accuracy differences be-

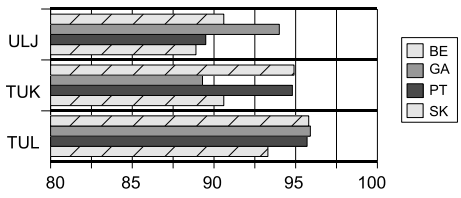

Figure 3: *C2 results for speech/non-speech classification*

tween the ULJ and TUL systems are almost identical to those under C1. Furthermore, the average error rate (100% - accuracy) appears to be about twice as large as under C1. Appar-

ently, three hours of training data is insufficient to achieve robust SNC.

4. Gender Classification (GC)

4.1. Algorithmic differences

Six institutions participated in this task. Four of them (ULJ, UMB, TUL and UVIGO) used GMMs, the other two (ELIS, INESC) used an MLP (Multi-Layer perceptron).

One institution (LJU) used different male and female models for telephone and broadband speech, for speech in the presence of music, etc.

4.2. Performance measures

The performance measures are the percentages of frames, male frames and female frames being classified correctly. The first one represents the accuracy of the GC.

4.3. Evaluation results

Figure 4 shows that under C1 all systems except ULJ offer very

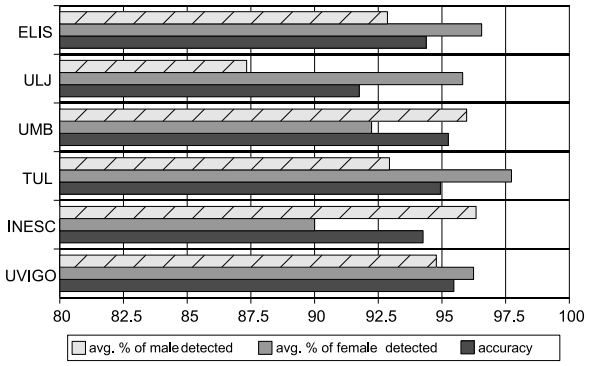

Figure 4: *C1 results for gender classification*

similar accuracies of around 95%. The type of classifier (GMM or MLP) seems rather irrelevant.

The C2 experiments show that 3 hours of training data is enough to attain close to maximal performances. The accura-

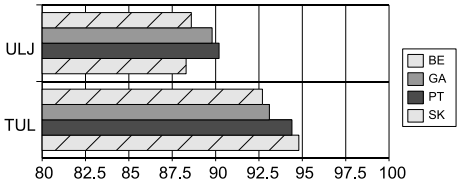

Figure 5: *C2 results for gender classification*

cies of both ULJ and TUL are only 2% lower than under C1. The fact that ULJ does not degrade more than TUL suggests that the problem with ULJ is maybe not the absolute quality level of the models (due to less training data), but the unequal quality of the different models (due to unbalanced training data).

5. Speaker Change Detection (SCD)

5.1. Algorithmic differences

Five institutions (ELIS, ULJ, TUL, UVIGO, INESC) participated in this task. They used the approaches described in [3, 6, 7, 8, 4] respectively.

In all cases the segmentation is performed in two stages: the first stage identifies potential candidate change points and the second stage tries to remove some of them on the basis of a more reliable analysis. Five groups used fixed length sliding windows in the first stage whereas one group (ULJ) continuously changes the window size until a new change point is found.

All partners use the Bayesian Information Criterion (BIC) in stage 2, and two of them also in stage 1. However, INESC and ELIS work with a Kullback-Leibler distance and a normalized log-likelihood ratio (LLR) instead. ELIS also applies its LLR normalization in conjunction with BIC in stage 2.

5.2. Performance measures

The performance measures are Recall (% of detected speaker change points), Precision (% of detected points which are genuine change points) and F-rate (defined as $2RP/(R+P)$). In order to compute these figures, a one-to-one computed-to-reference points mapping (cfr. [3]) with a maximum tolerance of 1 second on the time difference between mapped points is performed.

5.3. Evaluation results

According to Figures 6 and 7 there are substantial differences

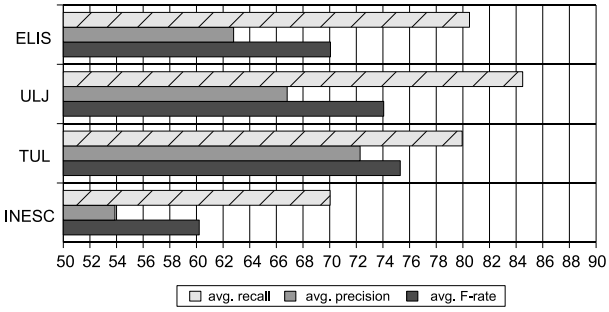

Figure 6: *C1 results for speaker change detection.*

between the results of the three best systems (ELIS, ULJ, TUL) and the other two, but it is difficult to explain them in terms of the cited algorithmic differences. It is hard to believe for instance that the different distance measure that was used in stage 1 would be the cause of the performance difference between ELIS and INESC. The results of the three leading systems are very comparable, suggesting that using fixed length (ELIS, TUL) or variable length (ULJ) windows in the first stage makes little difference.

Since none of the SCD approaches involves a training of models, one would expect them to perform equally well un-

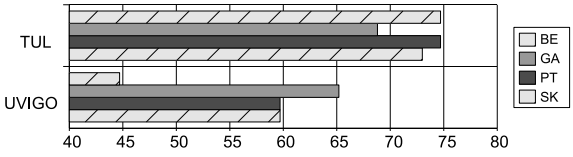

Figure 7: *C2 results (avg. F-rate) for speaker change detection.*

der C1 and C2. This is confirmed by the results of TUL. This finding also allows us to predict that UVIGO would have an accuracy of about 55% under C1.

The main conclusion here is that a more detailed comparison of algorithms is necessary to explain the observed differences. A nice opportunity for further collaboration.

6. Speaker Clustering (SC)

The speaker clustering algorithms were run with a reset of the cluster configuration at the beginning of a new file.

6.1. Algorithmic differences

Only three institutions (ELIS, INESC, ULJ) participated in this task and they all worked under condition C1.

Since the ELIS system works in real-time, it basically performs its clustering in a sequential manner [3]. However, when a number of consecutive segments are jointly clustered, the algorithm only merges a segment with an existing cluster if there is no evidence for introducing another segment as a new cluster. Another feature of the algorithm is that it does not permit the created clusters to accumulate more than a predefined number of frames.

The INESC and ULJ systems use a bottom-up agglomerative clustering procedure which iteratively merges the two most similar clusters into one new cluster.

In all systems the merging/creation of clusters is based on BIC.

6.2. Performance measures

In order to evaluate the clustering, a bi-directional one-to-one mapping of reference speakers to clusters is computed (NIST rich text transcription evaluation script). It defines the correct speaker/cluster for a cluster/speaker. Obviously, unmapped clusters/speakers have no correct speaker/cluster.

On the basis of this information, the Q-measure is defined as the geometrical mean of the percentage of cluster frames belonging to the correct speaker and the percentage of speaker frames labeled with the correct cluster. Since these percentages are zero for unmapped clusters/speakers, we have also provided a Q_{map} which is computed solely over the mapped cluster-speaker pairs.

Another performance measure is the Diarization Error Rate (DER) which is defined as the percentage of frames with an incorrect cluster-speaker correspondence.

Since no cluster information was passed between different files, the evaluation is also done on a file per file basis, and the shown performances are averages over different files.

6.3. Evaluation results

Figure 8 shows the two Q-measures (Fig. 8 (a)) and DER results (Fig. 8 (b)) for different systems.

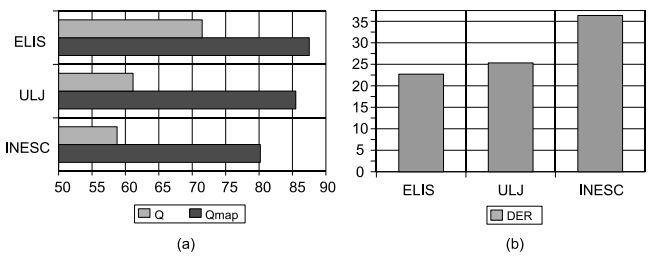

Figure 8: *C1 results for speaker clustering.*

Although INESC and ULJ follow more or less the same strategy, their DERs are substantially different. Could this mean that PLP features convey less speaker information than the MFCC features, or is INESC handicapped by its lower SCD accuracy, or is there still another explanation?

If ELIS outperforms ULJ, it is mainly because it produces less clusters for the same speaker, thus leaving less clusters unmapped. Since there are two differences in the clustering procedures, it would be interesting to find out which one is the main responsible for the performance differences.

7. Summary

By testing different audio indexing systems on the same data, using the same evaluation tools and protocols, it has been possible to identify some interesting performance differences in the areas of speaker change detection and speaker clustering. By deeper analyzing these differences in relation to algorithmic details it should be possible to make further progress.

So far, audio indexing systems were evaluated as such, but in the future the emphasis will be on the relation between the audio indexing accuracy and the speech and speaker recognition accuracy of a system making use of that indexing.

8. Acknowledgements

The work presented in this paper was performed in the Broadcast News Special Interest Group within the COST278 action on Spoken Language Interaction in Telecommunications.

9. References

[1] Articles in Issues 1-2, Speech Communication 37, Issues 1-2, 1–159, 2002.

[2] Vandecatseye, A., et al., ”The COST278 pan-European Broadcast News Database”, Procs. LREC 2004, Lisbon, 873–876, 2004.

[3] Vandecatseye, A., Martens, J. P., ”A fast, accurate and stream-based speaker segmentation and clustering algorithm”, Procs. Eurospeech 2003, Geneva, 941–944, 2003.

[4] Siegler, M. A., Jain, U., Raj, B., and Stern, R. M. ”Automatic segmentation, classification and clustering of broadcast news”, In Procs. DARPA Speech Recognition Workshop, Chantilly VA, 97–99, 1999.

[5] Williams, G. and Ellis, D., ”Speech/music discrimination based on posterior probability features”, Procs. Eurospeech 1999, Budapest, 687–690, 1999.

[6] Chen, S. S., et al. ”Automatic transcription of Broadcast News”, Speech Communication 37, 69–87, 2002.

[7] Chickering, D., Heckerman, D., ”Efficient approximations for the marginal likelihood of incomplete data given a Bayesian network” Procs. 12th Conf. on Uncertainty in Artificial Intelligence, Portland, 158–168, 1996.

[8] Perez-Freire, L. and Garcia-Mateo, C., ”A multimedia approach for audio segmentation in TV broadcast news”, Procs. ICASSP 2004, 369–371, 2004.