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], *Jindrich Zdansky*[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 during recognition, 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, or even worse, 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 task 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

All the tested algorithms fit into the canonical system architecture depicted on Fig. 1, with the exception that the SNC and the acoustic change detection can be interchanged. Most systems use an 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

ELIS, UMB, TUB and TUK work directly on the acoustic feature stream, whereas ULJ, TUL, INESC and UVIGO classify segments emerging from an acoustic change detector.

All systems use several Gaussian Mixture Models (GMMs) to model speech and non-speech frames respectively (e.g. GMMs for *speech, speech+music, background*, etc.). If SNC is performed before SCD, the GMMs are embedded in a looped automaton and a Viterbi algorithm is used to attain the optimal segmentation.

INESC adopts a totally different approach (cfr. [5]). A Multi-Layer perceptron (MLP) first computes a phone posterior probability vector for each frame. Then it derives for each segment a mean entropy of this vector and a mean difference between consecutive vectors. During speech one expects the entropy to be low (usually 1 dominant probability) and the probability differences high (sudden changes at phoneme boundaries), whereas during non-speech one expects the opposite.

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 are 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-

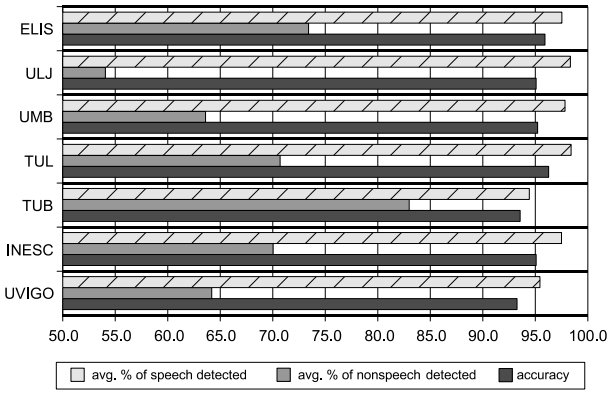

Figure 2: *C1 results for speech/non-speech classification*

cent speech and percent non-speech correct. One reason for this is that different institutions used different criteria for tuning their systems. Another reason could be the composition of the training database. One of the main problems in SNC appears to be the detection of music intervals. A lack of examples of the different kinds of music which appear in the COST278-BN database could hurt the performance.

We will compare algorithms on the basis of their accuracy, or equivalently their error rate (100% - accuracy), acknowledging that systems which were tuned on the basis of accuracy have an advantage then.

The C1 tests seem to suggest that all systems yield very comparable results. Interchanging the SNC and the SCD modules does not make a difference (compare ELIS, UMB to ULJ, TUL) and the alternative approach of INESC does not seem to be superior either.

The accuracies of the four C2 experiments are depicted on Figure 3. The accuracies of 91 and 95% obtained with ULJ

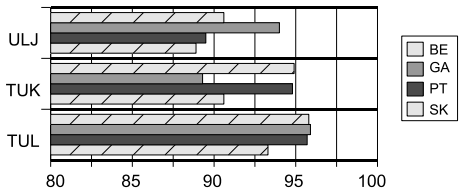

Figure 3: *C2 results for speech/non-speech classification*

and TUL are not much smaller than the corresponding 95 and

97% found under C1. Nevertheless, the corresponding percentages of misclassified frames are substantially higher now: an increase of 80 and 67% relative.

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 a MLP (Multi-Layer perceptron).

One institution (ULJ) 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

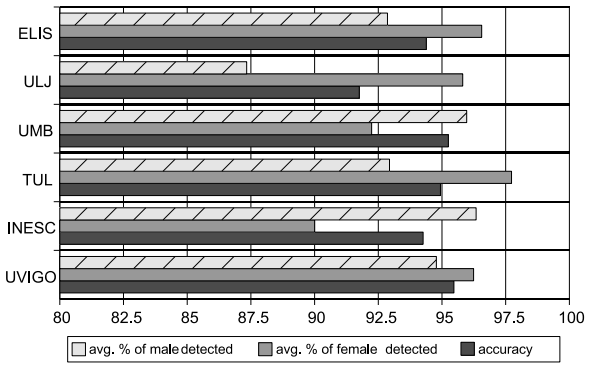

Figure 4: *C1 results for gender classification*

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

The C2 experiments (Figure 5) show that reducing the amount of training data to 3 hours does not hurt the performance

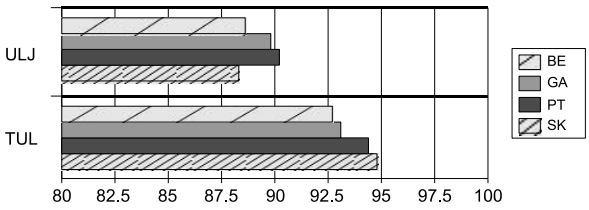

Figure 5: *C2 results for gender classification*

as much as in the case of SNC. 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 different male and female models (due to less training data), but their *unequal* quality (due to unbalanced training data).

5. Speaker Change Detection (SCD)

5.1. Algorithmic differences

Five institutions (ELIS, INESC, TUL, ULJ, UVIGO) participated in this task.

ELIS follows a two-stage approach [3]: stage 1 uses fixed length windows and a normalized log-likelihood ratio (LLR) between gaussian models to generate candidate change points, stage 2 iteratively applies an equally normalized ΔBIC to consecutive variable length candidate segment pairs so as to eliminate some of these points.

INESC uses a simple single stage process to find the positions where the Kullback-Leibler distance between consecutive fixed length windows [4] reaches a local maximum.

Conceptually, the TUL method searches for the best set of n change points, defined as the set maximizing the BIC for a model consisting of $n+1$ full covariance gaussian models (one per formed segment), and it identifies the best segmentation by repeating this process for different values of n. In practice, the whole method is implemented in the form of a single-pass Dynamic Programming process.

ULJ and UVIGO [7] use a two-stage algorithm: in stage 1 a standard BIC algorithm [6] is applied to produce candidate change points, and in stage 2 candidates can be rejected on the basis of a BIC-analyses performed on fixed length windows centered around these candidates.

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

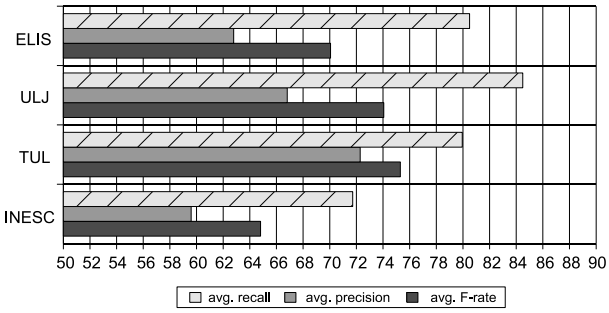

Figure 6: *C1 results for speaker change detection.*

between the different systems. Since none of the SCD approaches requires the training of any models, one would expect the C1 and C2 experiments to yield equivalent results. This is confirmed by the results of TUL.

Taking all the results (C1+C2) into account, it is clear that the different algorithms yield different outputs. There lies a nice opportunity here to study these differences in more detail and to use the so gathered information in search of a new algorithm that can outperform any of the algorithms tested in this study.

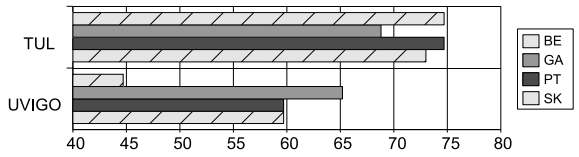

Figure 7: *C2 results (avg. F-rate) for speaker change detection.*

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 works sequentially [3], but nevertheless, short consecutive segments in one speech interval are jointly clustered. In that case, segments are merged with existing clusters on the condition that there is no evidence for selecting other segments as new clusters. Another feature of the algorithm is that clusters are not permitted 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 the ELIS and ULJ systems cluster merging or creation is based on BIC for full covariance gaussian distribution models. In the INESC system the diagonal covariance models are used, and the merging of adjacent segments is favored over that of distant segments.

6.2. Performance measures

In order to evaluate the clustering, a bi-directional one-to-one mapping of reference speakers to clusters is computed (as in NIST rich text transcription evaluation script). The mapped speaker-cluster pairs define the correct cluster for the speaker and the correct speaker for the cluster. Unmapped speakers (clusters) have no correct speaker (cluster).

Using the correct speakers/clusters, the Q-measure is defined as the geometrical mean of (1) the percentage of cluster frames referring to the correct speaker and (2) the percentage of speaker frames labeled with the correct cluster. Since unmapped clusters (speakers) have no correct speaker (cluster), we have also computed a Q_{map} on the basis of percentages derived for frames with mapped clusters and speakers respectively.

Another performance measure (related to Q) is the Diarization Error Rate (DER). Its is defined by NIST 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 the DERs (Fig. 8 (b)) for different systems. The INESC system has the largest DER, mainly because it generates more clusters. Since all systems generate more clusters than speakers, more clusters means more unmapped (incorrectly classified) clusters. On the other hand, more clusters tend to raise the number of speakers that can be mapped, and explains why the Qs of the INESC

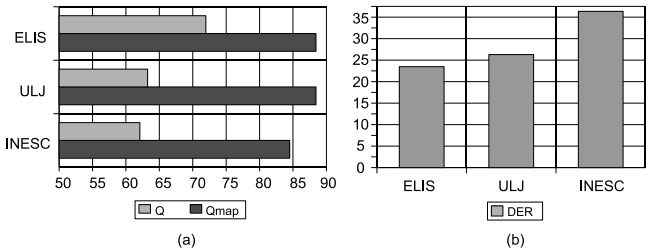

Figure 8: *C1 results for speaker clustering.*

and ULJ systems are much less different than their DERs. A more detailed analysis of differences is needed to find out why the ELIS and ULJ systems have such different Qs, in spite of their comparable DERs and numbers of clusters. Such an analysis can eventually result in the conception of a new clustering algorithm that can outperform the present ones.

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 more thoroughly analyzing these differences in relation to algorithmic details it should be possible for the participating institutions to make further progress in the near future.

In the present study the audio indexing systems were evaluated as independent systems. However, in the future the emphasis will be more 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 presented work was performed in the Broadcast News Special Interest Group within the European 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] Perez-Freire, L. and Garcia-Mateo, C., "A multimedia approach for audio segmentation in TV broadcast news", Procs. ICASSP 2004, 369–371, 2004.