Albayzin 2010 Evaluation Campaign: Speaker Diarization

Martin Zelenák, Henrik Schulz, and Javier Hernando

TALP Research Center
Department of Signal Theory and Communications,
Universitat Politècnica de Catalunya, Barcelona, Spain
{martin.zelenak,henrik.schulz,javier.hernando}@upc.edu

Abstract

In this paper we present the evaluation results for the task of speaker diarization in broadcast news domain as part of the Albayzin 2010 evaluation campaign of language and speech technologies. The evaluation data was a subset of the Catalan broadcast news database recorded from the 3/24 TV channel. Six competing systems from five different universities were submitted for the Albayzin 2010: Speaker diarization session and the lowest diarization error rate obtained was 30.4%.
Index Terms: speaker diarization, evaluation

1. Introduction

Objective evaluations became a valuable part of research and development in the field of spoken language processing. The comparison of performance of different approaches (systems) to a specific task helps setting new trends and stimulates the progress in a particular line of research. The Albayzin 2010 is the third in the series of evaluation campaigns (2006, 2008) organized by RTTH[1] and held under the FALA 2010 workshop. Largely inspired by the NIST Rich Transcription evaluations [1], the Albayzin 2010 campaign focuses among others on the task of speaker diarization of broadcast news.

Speaker diarization addresses the issue of segmenting a given audio stream according to different speakers and linking the speech regions which originate from the same person. In general, no kind of a priori speaker information is provided. In a broader sense, diarization also categorizes audio data according to music, background or channel conditions. Speaker diarization in broadcast news domain offers a strong application potential in many areas, in particular for transcription, indexing, searching and retrieval of audiovisual information.

In this paper we present an overview of the Albayzin 2010: Speaker diarization evaluation and report the results achieved by six submitted systems. The evaluation was performed on Catalan broadcast news data. Although the presented systems have several features in common (e.g. MFCCs, agglomerative clustering), there are also many differences among them (e.g. Poission-driven change rejection, online optimized processing, speaker factor analysis, dot-scoring similarity, or acoustic fingerprinting).

The rest of this paper is organized as follows. The conditions and database used for the evaluation are explained in Section 2. The participants are listed in Section 3 together with brief descriptions of their systems. The results are discussed in Section 4, followed by conclusions in Section 5.

This work has been funded by the Spanish project SAPIRE (TEC2007-65470). The first and second author are supported by a grant from the Catalan autonomous government.

[1] RTTH is the Spanish acronym for "Red Temática en Tecnologías del Habla" (the Spanish Speech Technologies Thematic Network)

2. Speaker diarization evaluation

2.1. Task and conditions

The organized evaluation campaign aims at evaluating the performance of automatic computer-based algorithms for speaker diarization, which can be also characterized as the "Who spoke when?" task. The participants could submit more than one system output, but only the primary hypothesis is considered here.

The minimum duration for a pause separating two utterances was set to 0.5 s, since pauses smaller than this value were not considered to be segmentation breaks in a speaker's speech (it is also complementary to the scoring collar discussed later).

The diarization error rate[2] (DER) defined by NIST [1] is the primary metric. DER is the ratio of incorrectly attributed speech time, (missed detections of speech, falsely detected speech, and speech assigned to the wrong speaker) to the total amount of speech time. Since there is no a priori relation between the system and reference speaker clusters, an optimum one-to-one mapping of reference speaker IDs to system output speaker IDs is computed separately for each audio file. A scoring "forgiveness collar" of 0.25 s around each reference segment boundary is used. This accounts for both the inconsistent annotation of segment times by humans and the philosophical argument of when does speech begin for word-initial stop consonants.

2.2. Database

The database contains broadcast news channel recordings, i.e., announcements, reports, interviews, discussions and short statements recorded from Catalan 3/24 TV channel throughout the program. Its original video recordings were supplied by a stationary digital video broadcasting (DVB-T) receiver. Their original audio tracks were extracted being available at 32 kHz sample rate, 16 bit resolution, but were downsampled to 16 kHz sample rate.

The annotated recordings encompass a total duration of 88 hours, but for the Albayzin 2010 speaker diarization evaluation a subset of 8 recording totaling approximately 30 hours was selected. Although TV3 is primarily a Catalan television channel, the recorded broadcasts contain a proportion of roughly $\frac{1}{6}$ of Spanish speech segments.

Catalan (mainly spoken in Catalonia) exhibits substantial dialectical differences, dividing the language into an eastern and

[2] NIST scoring tool available at: http://www.itl.nist.gov/iad/mig/tests/rt/2006-spring/code/md-eval-v21.pl

western group. The majority of recorded Catalan speakers features the central Catalan dialect being part of the eastern dialect group [2].

A first annotation pass segmented the recordings with respect to background sounds, channel conditions, and speakers as well as speaking modes. Table 1 shows the speaker distribution. Since segments of overlapping speakers did not receive a gender tag, they form also a subset of the "unknown" gender account. The gender conditioned distribution indicates a clear misbalance in favor of male speech data. The number of speakers per recording ranges from 30 to 250. Some speakers appear in several recordings (newscaster, journalists), however, the majority of the speakers account to only a short duration, since they are connected to particular news.

Table 1: *Distribution of speakers*

Gender	# Speakers	Duration [h]	# Segments
male	1239	44:23:41	12869
female	507	25:43:54	7559
unknown	270	07:50:38	2579
overlapped	68	00:12:38	241

Besides the tabulated total durations of audio segments of specific conditions in Table 2, there are a few conditions featuring an overlap of all noted background sounds with minor duration. Few segments are indicated to originate from telephone speech. The recorded speech within these segments can be considered band-limited to frequencies from 300 Hz to 3.4 kHz.

A second annotation pass provided literal transcriptions and acoustic events of segments that feature planned and spontaneous speech, but no long term background noises. The non-speech acoustic events were furthermore tagged with time stamps indicating their beginning and end.

Because of the fact that silences were not manually annotated, the transcriptions were extended by passing the signal through the hierarchical audio segmentation described in [3]. This involved a simple low-energy silence detector to estimate regions with non-speech (silence). Furthermore, to avoid too short segments, a smoothing constraining the minimal non-speech duration to 0.5 s was applied.

Table 2: *Duration breakdown regarding recording environment and background conditions of speech segments (number of segments in parenthesis)*

Channel	Background [h]			
	None	Speech	Music	Noise
None	04:27:10 (2451)	00:18:54 (131)	04:36:06 (1945)	01:15:30 (1113)
Studio	15:04:24 (4752)	01:36:16 (594)	08:40:47 (1407)	00:57:12 (2067)
Telephone	00:00:40 (11)	00:00:10 (2)		00:06:47 (10)
Outside	14:49:44 (6558)	03:55:29 (1319)	01:52:52 (557)	18:55:19 (4342)

Table 3: *Participating teams in the Albayzin 2010: Speaker diarization section*

Team ID	Research institution
AhoLab	University of the Basque Country (EHU)
GSI	University of Coimbra (UC)
GTM	University of Vigo (UVigo)
GTC-VIVOLAB	University of Zaragoza (UZ)
GTTS	University of the Basque Country (EHU)
ATVS-UAM	Autonomous University of Madrid (UAM)

3. Evaluation participants

3.1. Teams

Six teams from five universities submitted their systems to the Albayzin 2010 speaker diarization evaluation. The list of participants is given in Table 3.

3.2. System descriptions

Several teams participated also in the Albayzin 2010: Audio segmentation section, where five acoustic classes were defined to segment the audio data [4]. The classes were as follows: music, clean speech, speech with music, speech with noise and other (e.g. noise, silence). Since audio segmentation normally constitutes a part of speaker diarization systems, we are referring in latter system descriptions to these five acoustic classes.

3.2.1. AhoLab system

The system from Aholab team was built to run online and thus the whole process is performed in a single iteration. A more detailed description of the selected algorithms and modifications is given in [5]. The speech activity detection (SAD) is based on Viterbi segmentation of the audio signal into five acoustic classes. Each class is modeled with a Gaussian mixture model (GMM) and signal parameterization involves MFCCs with first and second derivatives.

For speaker change detection, growing window architecture and the Bayesian information criteria (BIC) metric is applied. Though the growing window has higher computational cost, the authors report its better performance compared to fixed-size sliding window approach and implemented a number of adjustments in order to decrease the computation time. At this stage of the process, only MFCC features with no feature derivatives are used. Furthermore, only voiced frames are included in the speaker change detection.

During the online clustering algorithm, every time a speaker change is detected, the BIC value of the recent speech segment against all known clusters is computed. If the lowest BIC value falls below a certain threshold the segment is assigned to the given cluster. Otherwise, a new cluster is created.

3.2.2. GSI system

The diarization system proposed by team GSI [6] includes an audio segmentation system to determine speaker turns and discard non-speech segments like silence and music. It uses a set of 16 MFCCs, 8 other features (e.g. energy, zero-crossing rate, spectral measures) and their derivatives. Segmentation is based on a hybrid ANN/HMM Viterbi decoder and discriminates between five acoustic classes.

To classify speakers, the algorithm begins with training a

background GMM with data of the entire audio file. Then, a decoder that outputs the most probable mixture sequence is used (with high mixture transition penalization) to detect speaker turns. Homogeneous segments with speech of only one speaker tend to produce sequences with few mixtures turns.

Two passes of verification are then applied to the labeled speaker segments to test whether every pair of segments is homogeneous or not. The first pass involves an audio fingerprint system and the other is based on BIC. If two segments are classified as similar, then the corresponding speaker labels are equated.

Acoustic or audio fingerprinting refers to a condensed representation of an audio signal that can be used to identify an audio sample or quickly locate similar items in audio streams. A binary representation of spectral patterns computed by the convolution of spectrogram with a mask is used. This technique is convenient to discover repeated segments with high confidence. Labels are determined according to a majority voting scheme in order to deal with classification inconsistencies in repeated segments.

3.2.3. GTM system

The GTM system [7] starts by making a coarse segmentation with the distance changing trend segmentation (DCTS) algorithm. Then, a refinement or rejection of detected audio change-points by an adaptive threshold-based BIC algorithm follows in order to reduce the false alarm rate. The change-point rejection approach assumes that the occurrence times of change-points can be modeled by a Poisson process (cumulative density function). Initially, a change is accepted with a very high probability, but as the number of accepted changes increases and is close or over the expected number, they are more likely to be rejected.

After this segmentation stage, the system successively decides whether a particular segment is speech, whether the speech is male or female, and, based on the cross likelihood ratio (CLR) test, whether the two latest speech segments are spoken by the same speaker. In that case both speech segments are merged.

Finally, an agglomerative hierarchical clustering step is performed to classify the speech segments by speaker identity. Similarity between speech segments is evaluated with a cosine distance measure which uses information about the likelihood score. Specifically, each speech segment is characterized with a collection of scores against a set of GMMs adapted for every segment from an universal background model (UBM).

The audio signal is characterized by 12 MFCCs augmented with the log-energy. The speech/non-speech and gender classification modules also consider the first and second derivatives.

3.2.4. GTC-VIVOLAB system

The speaker diarization systems submitted by the GTC-VIVOLAB team for the Albayzin 2010 speaker diarization evaluation [8] combines recent improvements in the field of speaker segmentation of two-speaker telephone conversations, using eigenvoice modeling, with the traditional BIC-based agglomerative hierarchical clustering approach.

The JFA-based (JFA stands for joint factor analysis) speaker segmentation system works with a given number of speakers (since it was designed for two-speaker dialogues). Because of that, after running speech activity detection, every recording is split into 5 minute slices and every slice is processed separately. The segmentation system is forced to find 10 speakers in every slice.

Once there are 10 clusters for every 5-minute slice, clustering over the whole recording is performed to merge those clusters belonging to the same speakers. For this purpose, BIC is considered as both a clustering metric and a stopping criterion. Clusters are modeled with a single full-covariance Gaussian function using 18 MFCCs.

3.2.5. GTTS system

The GTTS system detailed in [9] consists of three decoupled elements: speech/non-speech segmentation, acoustic change detection and clustering of speech segments. All of them rely on 13 MFCC features, which are augmented for clustering with first and second-order deltas.

Speech/non-speech segmentation is based on an ergodic continuous HMM with 5 states (one per acoustic class). With the aim to detect speaker changes, speech segments are further segmented by means of a naive XBIC-metric-based approach, which locates the most likely spectral change points. The authors state that almost all the speaker changes and many other additional changes were detected.

The third element is based on a dot-scoring speaker verification system, where speech segments are represented by MAP-adapted GMM zero- and first-order statistics. The dot scoring is then applied to compute a similarity measure between segments (or clusters) and finally an agglomerative clustering algorithm is used until no pair of clusters exceeds a similarity threshold.

3.2.6. ATVS-UAM system

The front-end parameterization of the ATVS-UAM speaker diarization involves the extraction of 19 MFCCs concatenated to their deltas, followed by cepstral mean normalization (CMN), RASTA filtering and feature warping. All speech data detected by a preceding audio segmentation step is used to train an UBM. Given this UBM, sufficient statistics are extracted for every segment. The next steps involve a factor analysis to model the total variability subspace resulting in so-called iVectors.

The MFCC feature stream is divided into 90-second audio slices. Compensated iVectors in each slice are clusterized based on their cosine distance. Cluster centroids are representing candidate speakers. Candidate speaker models are accumulated over all the slices in the test session together with the frequency of appearance of their clusters.

Speakers are expected to appear in several slices and thus a secondary clustering is used to merge the initial centroids, obtaining an enhanced set of candidate speakers. A prior probability is assigned to each of the candidate speakers according to its presence in the entire session. Likelihoods for each candidate speakers are estimated in a second pass over the iVector stream using the cosine distance and the prior probability of each candidate speaker. The final diarization labels are obtained with a Viterbi decoding of these scores. A more detailed description of the system can be found in [10].

4. Results

The DER results for six submitted systems in Albayzin 2010 are given in Table 4. In addition, the DER composition is also depicted in Figure 1. The best result of 30.4% DER was obtained by the AhoLab system, followed by similar performances of GTTS, GTC-VIVOLAB and ATVS-UAM systems. The performance rankings are closed with the DERs of GSI and GTM teams.

Note, that the most significant portion of DER is caused

Table 4: *Speaker diarization results for all participants in terms of Missed speech rate (*MS*), False alarm speech rate (*FA*), Speaker error rate (*SPKE*) and Diarization error rate (*DER*). All values are in given in (%).*

Team	MS	FA	SPKE	DER
AhoLab (EHU)	4.9	1.5	23.9	**30.4**
GSI (UC)	1.1	2.3	52.4	**55.8**
GTM (UVigo)	8.8	4.1	45.1	**58.0**
GTC-VIVOLAB (UZ)	3.7	1.5	28.6	**33.8**
GTTS (EHU)	2.2	2.2	28.8	**33.2**
ATVS-UAM	1.1	10.8	22.9	**34.7**

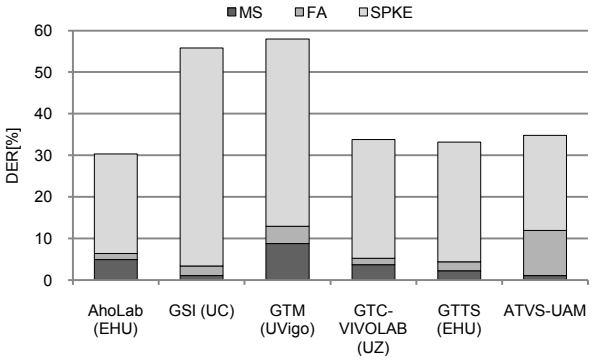

Figure 1: DER *distribution of missed-speech detections (*MS*), false-alarm detections (*FA*) and speaker error (*SPKE*).*

by incorrectly assigned speaker labels. This is very likely due to the high number of speakers in the evaluation corpus and variable background conditions. Lowest speaker error was achieved by the ATVS-UAM system with Viterbi decoding of iVector-stream scores over candidate clusters. Interesting question would be the impact of the score normalization according to cluster appearance probability on the error rates. Noteworthy is also the speaker error achieved by AhoLab, where the clustering happens in only a single iteration. The lowest error accounting to speech/non-speech detection produced the GSI system with a hybrid ANN/HMM approach.

The operation of the systems in terms of detected speaker count is shown in Figure 2. Here, the ATVS-UAM and GTTS systems exhibit the highest number of true detected speakers, but at the same time suffer from even higher counts of false speakers. The AhoLab system for instance, though detecting less correct speakers, maintains a significantly lower number of false speakers. Similarly the GTC-VIVOLAB system.

5. Conclusions

The Albayzin 2010 speaker diarization evaluation results were presented for six teams from four Spanish (EHU, UVigo, UZ, UAM) and one Portuguese (UC) university. The system which obtained the best result was also designed to run online and relies on modified growing-window BIC-based speaker-change detection and on a BIC-based clustering algorithm.

The evaluation data turned out to be relatively challenging, since the DER results in other comparable evaluations, e.g., the NIST RT'04 evaluation [11] or the ESTER evaluation on French broadcast news [12], were considerably lower than in this case. The high number of speakers in Catalan TV 3/24

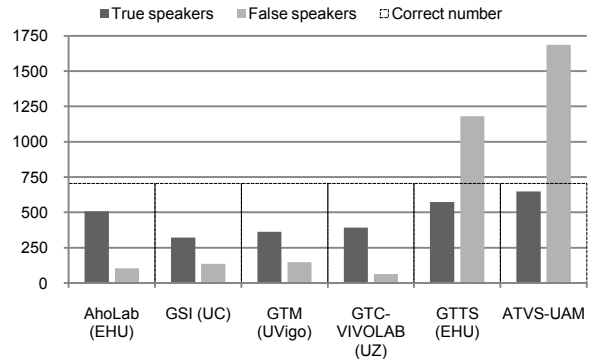

Figure 2: *Correctly detected (True) and falsely introduced (False) number of speakers by evaluated systems.*

broadcast news corpus was perhaps also the reason why no system managed to determine the correct speaker count in neither recording.

6. Acknowledgements

The authors would like to thank the evaluation organizers for their effort and also the participants for the help with the system descriptions.

7. References

[1] NIST. (2009) The NIST Rich Transcription evaluation project website. [Online]. Available: http://www.itl.nist.gov/iad/mig/tests/rt/

[2] M. W. Wheeler, *The Phonology of Catalan*. Oxford, UK: Oxford University Press, 2005.

[3] M. Aguilo, T. Butko, A. Temko, and C. Nadeu, "A hierarchical architecture for audio segmentation in a broadcast news task," pp. 17–20, September 2009.

[4] T. Butko and C. Nadeu, "A Hierarchical Architecture with Feature Selection for Audio Segmentation in a Broadcast News Domain," in *Proc. FALA 2010*, Vigo, Spain, 2010.

[5] I. Luengo, E. Navas, I. Saratxaga, I. Hernáez, and D. Erro, "AhoLab Speaker Diarisation System for Albayzin 2010," in *Proc. FALA 2010*, Vigo, Spain, 2010.

[6] A. Veiga, C. Lopes, and F. Perdigão, "Speaker Diarization Using Gaussian Mixture Turns and Segment Matching," in *Proc. FALA 2010*, Vigo, Spain, 2010.

[7] L. Docio-Fernandez, P. Lopez-Otero, and C. Garcia-Mateo, "The UVigo-GTM Speaker Diarization System for the Albayzin'10 Evaluation," in *Proc. FALA 2010*, Vigo, Spain, 2010.

[8] C. Vaquero, A. Ortega, and E. Lleida, "VIVOLAB-UZ Speaker Diarization System for the Albayzin 2010 Evaluation Champaign," in *Proc. FALA 2010*, Vigo, Spain, 2010.

[9] M. Diez, M. Penagarikano, A. Varona, L. J. Rodriguez-Fuentes, and G. Bordel, "GTTS System for the Albayzin 2010 Speaker Diarization Evaluation," in *Proc. FALA 2010*, Vigo, Spain, 2010.

[10] J. Franco-Pedroso, I. Lopez-Moreno, D. T. Toledano, and J. Gonzalez-Rodriguez, "ATVS-UAM System Description for the Audio Segmentation and Speaker Diarization Albayzin 2010 Evaluation," in *Proc. FALA 2010*, Vigo, Spain, 2010.

[11] J. Fiscus, A. Le, and G. Sanders. (2004) MDE Tasks and Results. [Online]. Available: http://www.itl.nist.gov/iad/mig/tests/rt/2004-fall/rt04f-mde-nist.pdf

[12] S. Galliano *et al.*, "The ESTER Phase II Evaluation Campaign for the Rich Transcription of French Broadcast News," in *Proc. Interspeech '05*, Lisbon, Portugal, 2005, pp. 1149–1152.