EURASIP Journal on
Audio, Speech, and Music Processing
a SpringerOpen Journal

RESEARCH Open Access

Speaker diarization of broadcast news in Albayzin 2010 evaluation campaign

Martin Zelenák*, Henrik Schulz and Javier Hernando

Abstract

In this article, we present the evaluation results for the task of speaker diarization of broadcast news, which was part of the Albayzin 2010 evaluation campaign of language and speech technologies. The evaluation data consists of a subset of the Catalan broadcast news database recorded from the 3/24 TV channel. The description of five submitted systems from five different research labs is given, marking the common as well as the distinctive system features. The diarization performance is analyzed in the context of the diarization error rate, the number of detected speakers and also the acoustic background conditions. An effort is also made to put the achieved results in relation to the particular system design features.

Keywords: Speaker diarization, Evaluation, Broadcast news

Introduction

Speaker diarization has attracted the interest of the scientific community already since several years. Given an audio recording, the goal is to answer the question: "*Who spoke when?*" 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. Over the years, most research effort was focused on speaker diarization in broadcast news domain, but recently there has been also a strong interest in lecture and conference meeting domain. This technology offers a strong application potential in many areas, in particular for transcription, indexing, searching, and retrieval of audiovisual information. Furthermore, diarization can contribute to increased robustness of other human language technologies like automatic speech recognition (ASR) by unsupervised adaptation of speech models to particular speakers. Speaker diarization task consists of two main steps. First is the segmentation of a conversation, involving multiple speakers, into speaker-homogeneous chunks. Second step aims to group together all the segments that correspond to the same speaker. The first part of the process is also referred to as speaker-change detection and the second is known as clustering.

A lot of diverse approaches to the speaker diarization task can be found in the literature, but in general, there are two predominant strategies. The *step-by-step* strategy deals with the main steps successively [1-3]. A limitation of this method is that it is not only difficult to correct the errors made in the segmentation later on, but these errors degrade the performance of the subsequent clustering step. An alternative approach, referred to as *integrated* strategy, is to optimize the segmentation and clustering jointly [4,5]. Both steps are performed simultaneously in an iterative procedure which uses, for instance, a set of Gaussian mixture models (GMMs) or an ergodic hidden Markov model (HMM). The drawback of this approach is the need to estimate these models using very short segments, even though the speaker models get refined along the process. Mixed strategies are also proposed, where classical step-by-step segmentation and clustering are first applied, and then the segment boundaries and clusters are refined jointly [6-8]. Fusion of both techniques can be found in [9]. The most popular strategies comprise Bayesian-information-criterion-based (BIC) segmentation [1] and agglomerative bottom-up clustering. With bottom-up clustering the optimal number of speaker clusters is determined by subsequent merging of a high number of clusters in an iterative process until a stopping criterion is met.

*Correspondence: martin.zelenak@upc.edu
TALP Research Center, Department of Signal Theory and Communications, Universitat Politècnica de Catalunya, C/Jordi Girona 1-3, 08034 Barcelona, Spain

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[a] and held under the FALA 2010 workshop [10]. Largely inspired by the NIST Rich Transcription evaluations [11], the Albayzin 2010 campaign focuses among others on the task of speaker diarization of broadcast news.

In this article, we present as the co-organizers of Albayzin 2010 responsible for speaker diarization section an overview of the evaluation and report the results achieved by five submitted speaker diarization systems. The evaluation was performed on Catalan broadcast news data. Although the presented systems have several features in common, such as the use of mel frequency cepstral coefficients (MFCCs) or agglomerative clustering, there are also many differences among them, e.g., online-optimized processing (for live audio input), speaker factor analysis [12], dot-scoring similarity [13], or acoustic fingerprinting [14]. Based on the observed results, we try to outline promising investigation directions.

Broadcast news is a challenging domain, because such shows contain an unpredictable number of different speakers speaking for a very variable amount of time and speakers sometimes talk simultaneously. However, overlapping speech issue was not very significant in this case (reference annotations show only 0.19 % of overlapping speech). Broadcast news data often contain a large amount of music and commercial breaks.

This article is organized as follows. Scoring metric and database used for the evaluation are explained in Section 'Speaker diarization evaluation'. The participants are listed in Section 'Evaluation participants' together with brief descriptions of their systems. The diarization results are reported in Section 'Evaluation results and system comparison' together with a discussion about the systems in the context of the achieved results. Conclusions follow in Section 'Conclusions'.

Speaker diarization evaluation

Task and scoring

The organized evaluation campaign aims at evaluating the performance of automatic 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 was considered here.

The minimum silence duration separating two utterances was set to 0.5 s like in NIST RT '04 [11], 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[b] (DER) defined by NIST [11] is the primary metric. The audio file is divided into contiguous segments demarcated by all reference and system speaker change points so that the set of compared speakers in one segment does not change. Then, the metric is defined as follows:

$$\mathrm{D}ER = \frac{\sum_{\forall s} \mathrm{dur}(s) \cdot \big(\max(N_{\mathrm{r}ef}(s), N_{sys}(s)) - N_{corr}(s)\big)}{\sum_{\forall s} \mathrm{dur}(s) \cdot N_{\mathrm{r}ef}(s)}, \tag{1}$$

where d$ur(s)$ is the duration of a particular segment s, $N_{\mathrm{r}ef}(s)$ is the number of reference speakers speaking in s, $N_{sys}(s)$ is the number of system speakers in s and $N_{corr}(s)$ is the number of matching reference and system speakers who are speaking in s. DER represents the ratio of incorrectly attributed speech time The DER can be broken down into speaker errors (SPKE), which accounts for miss-attributed speaker segments, false alarms (FA), and missed speech errors (MS). The latter two account for non-speech labeled as speech and vice versa.

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 uncertainty when does speech begin for word-initial stop consonants.

Database

The specificity of the Albayzin 2010 campaign compared to other existing evaluation campaigns is that it focuses on Iberian languages (i.e., Castilian, Catalan, Basque, Galician, Portuguese). In this way, it is additional to NIST evaluations [11] concerned with English, Arabic and Chinese languages, or the ESTER evaluation [15] dealing with French. Broadcast news is one of the main domains of speaker diarization since it offers a strong application potential, especially in the context of improving the readability of automatic transcriptions. Performing evaluation on these data is also a good complement to evaluations on telephone conversations and meeting recordings. The used 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 comprise a total duration of 88 hours, but for the Albayzin 2010 speaker diarization evaluation a subset of 8 recordings totaling approximately 30 hours was selected. Although TV3 is primarily a Catalan television channel, the recorded broadcasts contain a proportion of roughly 8.5 % of Spanish speech segments.

Catalan language, mainly spoken in Catalonia, exhibits substantial dialectical differences and is divided into an eastern and western group. The eastern dialect includes northern Catalan (French Catalonia), central Catalan (the eastern part of Catalonia), and Balearic. The western dialect includes north-western Catalan and Valencian (south-west Catalonia) [16]. Presumably, the majority of recorded Catalan speakers features the central Catalan dialect.

A first manual 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 imbalance in favor of male speech data. The number of speakers per recording ranges from 30 to 250 with some speakers appearing in several recordings (newscaster, journalists). However, the majority of speakers are related only to a particular news and account to only a short duration.

The total durations of audio segments of specific conditions are given in Table 2. Besides, there are a few conditions featuring an overlap of all noted background sounds, but only with minor duration and are therefore omitted. 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. The channel "undefined" refers to segments that cannot be explicitly assigned to one of the other defined channels.

A second manual 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.

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

Table 2 Distribution of recording channel and background conditions

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

The number of segments are in parenthesis.

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 [17]. 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.

Evaluation participants

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

After submitting evaluation results one of the teams discovered that in half of the recording sessions their system was reading corrupt audio input. Therefore, their evaluation results cannot be considered representative and only five systems are presented in this article. The original description of the speaker diarization evaluation can be found in [18], where the systems are related to their corresponding research labs explicitly.

All teams except AhoLab also participated in another category of the Albayzin 2010 evaluation, in the audio segmentation section, where five acoustic classes were defined to segment the audio data [19]. The classes were as follows: music, clean speech, speech with music, speech

Table 3 Teams participating 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)

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.

System 1

The algorithmic concept of System 1 [18] facilitates an online execution, i.e., the complete process is performed in a single iteration. The parameterization of the signal involves static MFCCs with first and second-order derivatives, although for the detection of speaker turns the derivatives are not considered. An initial speech activity detection (SAD) employs a Viterbi segmentation [20] of the parameterized audio signal and distinguishes five acoustic classes (music, clean speech, speech with music, speech with noise, and other). Each class is modeled with a GMM.

Subsequently, the speaker change detection employs a growing window approach [21] and the Bayesian information criterion (BIC) to measure the dissimilarity of two adjacent windows. The BIC metric estimates if windowed audio data is better modeled with two distributions or with only a single one. A change point is detected at positions where the BIC value exceeds a zero threshold. Even though the growing window scheme has higher computational cost, the authors of System 1 report its better performance compared to fixed-size sliding window approach and implemented a number of adjustments in order to decrease the computational load (skipping improbable places, window length limit). At this stage of the process, only static MFCC features with no derivatives are used. The speaker change detection of this system relies only on voiced audio data. During the system development it was observed that by discarding unvoiced frames it is possible to reduce the diarization error by up to 12 % [18].

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. The threshold is estimated in the same fashion as in [21]. The theoretically suboptimal online algorithm can in practice benefit from the fact that it is prone to combine adjacent segments rather than segments far apart and consecutive segments are likely to come from the same speaker.

System 2

System 2 [18] incorporates audio segmentation prior to the diarization to determine speaker turns and discard non-speech segments like silence and music. Signal parametrization uses a set of 16 MFCCs, 8 other features (energy, zero-crossing rate, spectral centroid, spectral roll-off, maximum normalized correlation coefficient and its frequency, harmonicity measure, spectral flux), and their derivatives. Audio segmentation is based on a hybrid multi-layer perceptron/hidden Markov model decoder and discriminates between five acoustic classes defined for the audio segmentation task.

In order to classify speakers the algorithm uses a simple Viterbi decoder. It begins with training a universal background model (UBM) with speech data of the entire audio file. Subsequently, the decoder determining the most likely mixture sequence detects (with high mixture transition penalization) the speaker turns. Homogeneous segments with speech of only one speaker are usually decoded during the most of the segment time with the same Gaussian mixture.

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 [22] and the other is based on BIC. If two segments are classified as similar, then the corresponding speaker labels are equated.

Audio fingerprinting [14] 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 [22]. This technique is convenient to discover repeated segments. Labels are determined according to a majority voting scheme in order to deal with classification inconsistencies in repeated segments.

System 3

System 3 [18] employs recent improvements in speaker segmentation of two-speaker telephone conversations using eigenvoice modeling [23] and the traditional agglomerative hierarchical clustering [1,24], similarly to the speaker segmentation technique proposed in [12].

The eigenvoice-based speaker segmentation in this system was originally designed for two-speaker telephone conversations [25], thus it works with a given number of speakers. Therefore, after separating the speech frames, every recording is split into 5 min slices and every slice is processed individually. The segmentation system is forced to find 10 speakers in every slice.

The parameterization of the signal consists of the extraction of 18 static MFCCs. Every speaker GMM is adapted from a background model using the eigenvoice approach [23]. Given a sequence of feature vectors, 20 speaker factors are estimated for every frame over a 100-frame window and then transformed with the within-class covariance normalization (WCCN) [26] in order to compensate for the intra-session variability. Afterwards, a 10-Gaussian GMM is estimated to model the stream of speaker factors (as in [12]), where each Gaussian will be assigned to a single speaker.

Once there are 10 clusters for every 5 min 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 MFCCs.

System 4

System 4 [18] consists of three decoupled elements: speech/non-speech segmentation, acoustic change detection, and clustering of speech segments. As far as signal parameterization is concerned, all of them rely on 13 static MFCC features. However, for clustering the static MFCCs are additionally augmented with their first and second-order derivatives.

Speech/non-speech segmentation makes use of an ergodic continuous HMM with 5 states (one per acoustic class mentioned earlier). In order to detect speaker change points, speech segments were further segmented by means of a conventional metric-based approach [3] evaluating the likelihood of the acoustic change in the center of a sliding window using normalized Cross-BIC (XBIC) metric [27]. The authors of the system state that with this approach, besides many additional acoustic changes, almost all the speaker changes were detected.

The clustering employs linear dot-scoring, a fast and simple technique for scoring test segments against target models which employs the first-order Taylor-series approximation to the GMM log-likelihood. For each speech segment a GMM was MAP-adapted from a universal background model, and zero- and first-order sufficient statistics are computed. The similarity between different segments is then estimated with TZ-normalized [28] dot scores. Finally, an agglomerative clustering algorithm is used until no pair of clusters exceeds a similarity threshold.

System 5

The front-end parameterization of the speaker diarization System 5 [18] involves the extraction of 19 static MFCCs with their derivatives, followed by cepstral mean normalization (CMN), RASTA filtering [29], and feature warping [30]. All speech segments from training data are used to train a UBM. Given this UBM, sufficient statistics are extracted for every segment labeled by a preceding audio segmentation step. The next steps involve a factor analysis to model the total variability subspace resulting in so-called iVectors [31] and a linear discriminant analysis (LDA) transformation of the iVectors [32] computed over 1 s windows. The LDA technique discriminates between speakers and was also estimated from the training data.

The MFCC feature stream is divided into 90 s audio slices with certain overlap. Computed LDA-projected iVectors in each slice are clustered based on their cosine distance with a constrained maximum allowed distance between an iVector and a cluster centroid. Cluster centroids, the mean of iVectors in each cluster, represent 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 presumably appear in several slices, thus a secondary clustering merges 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. Finally, the output diarization labels are obtained by a Viterbi decoding of so-calculated speaker scores.

Evaluation results and system comparison

In this section, the performances of the five systems are presented and analyzed from different perspectives. Firstly, we review the main evaluation metric (DER) and discuss the distributions of its three components (misses, false alarms, speaker errors.) In the following, the performance for individual test sessions and also for different background conditions is presented. Next, the number of detected speakers by the systems is analyzed. Finally, we highlight the main outcomes of this performance evaluation.

Diarization performance

The DER results for five submitted systems in Albayzin 2010 are given in Table 4, where it is clear that the most successful system was the System 1 with 30.4 % DER. Furthermore, a decomposition considering missed-speech detection, false alarms, and false speaker labeling is also depicted in Figure 1.

Figure 1 indicates incorrect assigned speaker labels as the most significant proportion of the DER. The challenge seems to be the fact that many speakers speak only short

Table 4 Speaker diarization results for all participants

Team	MS	FA	SPKE	DER
System 1	4.9	1.5	23.9	**30.4**
System 2	1.1	2.3	52.4	**55.8**
System 3	3.7	1.5	28.6	**33.8**
System 4	2.2	2.2	28.8	**33.2**
System 5	1.1	10.8	22.9	**34.7**

Speaker diarization results are in terms of missed speech rate (MS), false alarm rate (FA), speaker error rate (SPKE), and diarization error rate (DER) in (%).

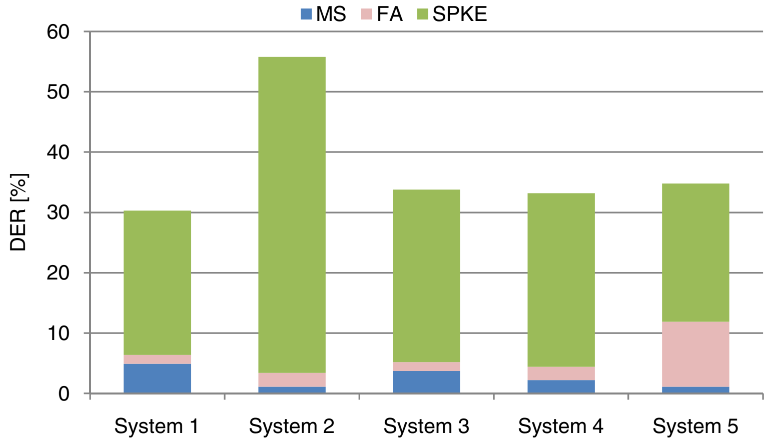

Figure 1 Overall speaker diarization results. DER distribution of missed speech rate (MS), false alarm rate (FA), and speaker error rate (SPKE) .

segments of time, while a speaker may feature different background conditions.

The speaker error achieved by the first system is very remarkable, since all the clustering happens in only a single iteration. Furthermore, System 1 relies on the most popular approaches of the state-of-the-art diarization systems. Even though it is not possible to directly derive a conclusion from this result, the strategy to discard unvoiced frames in speaker change detection may have been the crucial factor of the best performance. The SAD of System 1 was tuned for hypothesizing more misses than insertions (false alarm). The SAD performance is reflected by the amount of MS and FA errors.

The balanced SAD of System 4 and robust techniques applied for speaker segmentation resulted in the second best DER according to Table 4. The factor analysis technique used in speaker segmentation of System 3 proved to be well-suited for this task. The overall DER and speaker error rate in particular were very similar for Systems 3 and 4.

The factor analysis approach was also employed in System 5, which achieved the lowest speaker error with Viterbi decoding of iVector-stream scores over candidate clusters. It remains an open question, how the score normalization according to cluster appearance probability impacts the error rates.

System 2, with its hybrid ANN/HMM approach, displays the lowest error accounting to speech/non-speech detection, but it cannot benefit from this advantage in the overall performance. It is unclear what was the major reason for the higher overall DER score. It may have been the very simple initial speaker change detection, or the fingerprinting technique, which was observed to study well for audio segmentation [19], is not so appropriate for clustering speaker segments. Eventually, using the same set of acoustic features (and their derivatives) in all three stages of the process may not have been the optimal choice.

Session variability

A more detailed analysis of the DER for each testing session shows (see Figure 2) that the performance of the systems was rather stable. In practice, a relatively high variability across different recordings, despite having comparable characteristics, is nothing unusual with diarization [33]. The DER standard deviation over the eight test recordings for each system lies between 4.6 and 8.0 % DER. All systems were operating well (with respect to their average performance) in the test session 23. The absolutely lowest error of 21.6 % DER was achieved by the System 4 on session 19. It seems that the biggest difficulties for the majority of evaluated systems posed the session 22, probably also due to a rather high (but not the highest) number of involved speakers. The session DER and the number of speakers are only very vaguely correlated, though.

Acoustic background conditions

The speech signal can be divided according to acoustic background conditions into three categories: clean speech, speech over noise, and speech over music. A particular difficulty of the diarization task is due to the nature of broadcast news data, which may exhibit different background conditions for one and the same speaker. It makes it very challenging for the clustering algorithm to put such speaker segments with different background conditions into the same cluster. The proportions of the durations of the background conditions (music, noise, silence), including both speech and non-speech segments, reflect how these three classes roughly contribute to the overall diarization error. Clean-speech, speech-over-noise, and

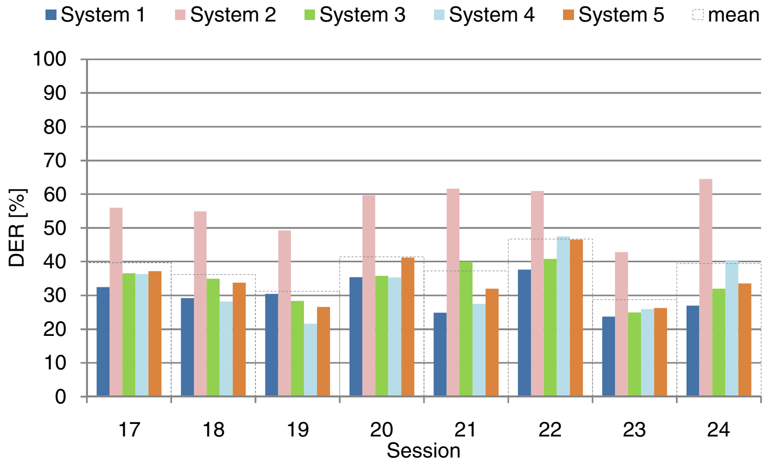

Figure 2 Speaker diarization per session. DER performance for each of the eight test recordings from 3/24 TV broadcast news corpus.

speech-over-music segments are influencing the DER by 36, 46, and 18 %, respectively. Looking at the individual DER performances (evaluating each speech class independently), given in Figure 3, it is not surprising that the DERs of clean speech are usually the lowest.

Number of speakers

The operation of the systems in terms of detected speaker count is shown in Figure 4. Here, the Systems 5 and 4 exhibit the highest number of true detected speakers, but at the same time suffer from even higher counts of false speakers. The System 1, for instance, though detecting less correct speakers, maintains a significantly lower number of false speakers. Similar observation applies also for the operation of System 3.

The possible reason for the high number of false speakers of System 4 could be the substantial initial over-segmentation (reported in Section 'System 4') in a combination with a too strictly defined merging threshold of the dot-scoring similarity. Nevertheless, since the overall DER is not much different from System 1 or 3, the affected speaker segments were probably very short.

In the case of System 5, the probable cause of the high number of falsely detected speakers lies in the substantial false alarm rate (FA error, see Table 4) of the speech/non-speech detection rather than clustering algorithm, because speaker error rate (SPKE) is very good compared with other systems.

Summarizing points

The analysis of speaker diarization results and the characteristics of the submitted systems revealed several observations which can be summarized as follows:

- The use of only voiced frames for performing speaker segmentation, which was implemented in one of the

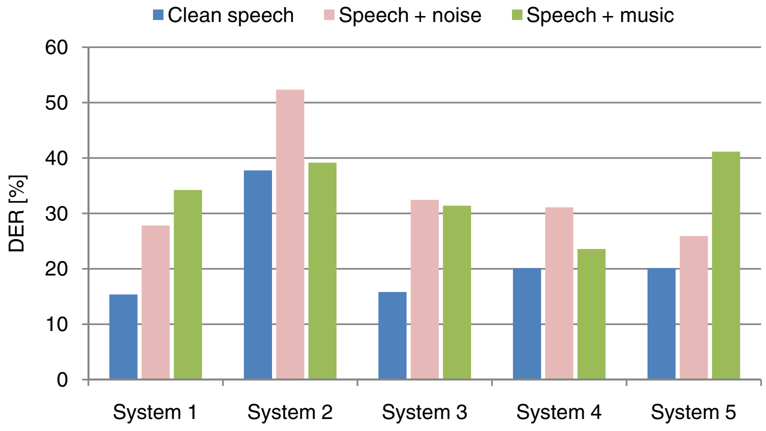

Figure 3 Speaker diarization per background condition. DER performance according to three acoustic background conditions: clean speech, speech with noise, and speech over music.

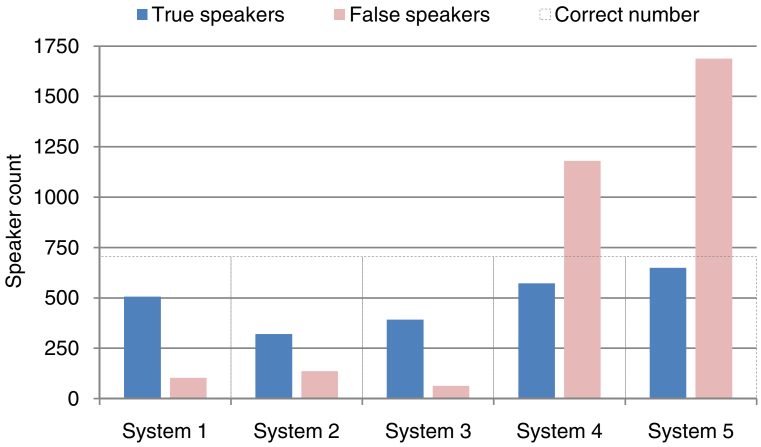

Figure 4 Number of detected speakers. Correctly detected (True) and falsely introduced (False) number of speakers by evaluated systems.

systems, seems a very interesting step in context of the very good speaker error result of that particular system.

- The speaker factor analysis technique which received attention in the field of speaker verification was successfully adopted in two presented diarization systems. Both of them delivered competitive results compared to the best system.
- Almost all systems rely exclusively on MFCC features (13–19 coefficients), for clustering also their derivatives can be used.
- BIC maintains as the most popular and effective cluster merging metric and/or clustering stopping criterion. It can be accompanied with other segmentation passes applying other metrics, but in all the cases the BIC is present at some point. The growing window strategy provides better results than a fixed-size sliding window (the winner system also relied on growing window), but the computational cost is normally also larger.
- All the systems used the conventional bottom-up agglomerative clustering approach. Even though it can sometimes suffer from merging instability or stopping criteria difficulties, it is usually robust and is also the most popular in other state-of-the-art systems. Some systems do not manage to estimate very well the approximate number of speakers, probably due to a too high initial over-segmentation. However, the durations of the segments for the majority of falsely introduced speakers amount to only very short times, and hence the speaker error rate is not increased substantially.

Conclusions

The Albayzin 2010 speaker diarization evaluation results were presented for five of the 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 design of the presented systems confirmed the popularity of cepstral features, BIC metric (or its modification) and agglomerative bottom-up clustering approach for the diarization task. The evaluation results also showed the effectiveness the factor analysis approach adopted from speaker recognition.

The evaluation data turned out to be relatively challenging, since the DER results in other comparable evaluations, e.g., the NIST RT'04 evaluation [34] or the ESTER evaluation on French broadcast news [15], were considerably lower than in this case. The high number of speakers in Catalan TV 3/24 broadcast news corpus was perhaps also the reason why no system managed to determine the correct speaker count in neither recording.

Endnotes

[a] RTTH is the Spanish acronym for "Red Temática en Tecnologías del Habla" (the Spanish Speech Technologies Thematic Network), http://www.rthabla.es/.

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

Competing interests

The authors declare that they have no competing interests.

Acknowledgements
This study had been funded by the Spanish project SARAI (TEC2010-21040-C02-01). The authors would like to thank the organizers of the FALA 2010 workshop for their effort and also the participants for the help with the system descriptions.

Received: 1 June 2011 Accepted: 17 July 2012
Published: 31 July 2012

References

1. S Chen, P Gopalakrishnan, Speaker, environment and channel change detection and clustering via the Bayesian information criterion. in *Proc. DARPA Broadcast News Transcription and Understanding Workshop* (1998), pp. 127–132
2. J Gauvain, LF Lamel, G Adda, Partitioning and transcription of broadcast news data. in *Proc. 5th International Conference on Spoken Language Processing (ICSLP'98)* (Sydney, Australia, 1998), paper 0084
3. MA Siegler, U Jain, B Raj, RM Stern, Automatic segmentation, classification and clustering of broadcast news audio. in *Proc. DARPA Speech Recognition, Workshop* (Chantilly, VA, USA, 1997), pp. 97–99
4. J Ajmera, C Wooters, A robust speaker clustering algorithm. in *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU'03)* (St. Thomas, VI, USA, 2003), pp. 411–416
5. S Meignier, J Bonastre, S Igounet, E-HMM approach for learning and adapting sound models for speaker indexing. in *Proc. 2001: A Speaker Odyssey - The Speaker Recognition Workshop (Odyssey-2001)* (Crete, Greece, 2001), pp. 175–180
6. L Wilcox, F Chen, D Kimber, V Balasubramanian, Segmentation of speech using speaker identification. in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'94),* vol. 1 (Adelaide, Australia, 1994), pp. I/161–I/164
7. DA Reynolds, RB Dunn, JL McLaughlin, The Lincoln Speaker Recognition System: NIST Eval2000. in *Proc. ICSLP'00,* vol. 2 (Beijing, China, 2000), pp. 470–473
8. D Moraru, L Besacier, E Castelli, Using a priori information for speaker diarization. in *Proc. Odyssey-2004 The Speaker and Language Recognition Workshop* (Toledo, Spain 2004), pp. 355–362
9. S Meignier, D Moraru, C Fredouille, J Bonastre, L Besacier, Step-by-step and integrated approaches in broadcast news speaker diarization. Comput. Speech Lang. **20**(2-3), 303–330 (2006)
10. FALA: FALA 2010 Proceedings (2010), http://fala2010.uvigo.es/images/proceedings/. Accessed 30 May 2011
11. NIST: The NIST rich transcription evaluation project website (2009), http://www.itl.nist.gov/iad/mig/tests/rt/. Accessed 30 May 2011
12. F Castaldo, D Colibro, E Dalmasso, P Laface, C Vair, Stream-based speaker segmentation using speaker factors and eigenvoices. in *Proc. ICASSP'08* (Las Vegas, NV, USA, 2008), pp. 4133–4136
13. M Diez, M Penagarikano, A Varona, L Rodriguez-Fuentes, G Bordel, On the use of dot scoring for speaker diarization. in *Pattern Recognition and Image Analysis, vol. 6669 of* Lecture Notes in Computer Science, ed. by J Vitrià, Ja Sanches, and M Hernández. (Springer, Berlin, 2011), pp. 612–619
14. J Haitsma, T Kalker, A highly robust audio fingerprinting system. in *Proc. ISMIR'02* (Paris, France, 2002)
15. S Galliano, E Geoffrois, D Mostefa, K Choukri, JF Bonastre, G Gravier, The ESTER phase II evaluation campaign for the rich transcription of French broadcast news. in *Proc. Interspeech'05* (Lisbon, Portugal, 2005), pp. 1149–1152
16. MW Wheeler, *The Phonology of Catalan* (Oxford University Press, Oxford, 2005)
17. M Aguiló, T Butko, A Temko, C Nadeu, A hierarchical architecture for audio segmentation in a broadcast news task. in *Proc. Workshop on Speech and Language Technologies for Iberian Languages* (Porto Salvo, Portugal, 2009), pp. 17–20
18. M Zelenák, H Schulz, J Hernando, Albayzin 2010 evaluation campaign: speaker diarization. in *Proc. FALA 2010* (Vigo, Spain, 2010), pp. 301–304
19. T Butko, C Nadeu, H Schulz, Albayzin-2010 audio segmentation evaluation: evaluation setup and results. in *Proc. FALA 2010* (Vigo, Spain, 2010)
20. L Rabiner, A tutorial on hidden Markov models and selected applications in speech recognition. Proc. IEEE. **77**(2), 257–286 (1989)
21. A Tritschler, RA Gopinath, Improved speaker segmentation and segments clustering using the Bayesian information criterion. in *Proc. Eurospeech'99* (Budapest, Hungary 1999), pp. 679–682
22. C Neves, A Veiga, L Sá, F Perdigão, Fingerprinting system for broadcast streams. in *Proc. Conference on Telecommunications (ConfTele2009)* (2009), pp. 481–484
23. R Kuhn, JC Junqua, P Nguyen, N Niedzielski, Rapid speaker adaptation in eigenvoice space. IEEE Trans. Speech Audio Process. **8**(6), 695–707 (2000)
24. RO Duda, PE Hart, DG Stork, *Pattern Classification,* 2nd ed. (Wiley-Interscience, New York, 2000)
25. C Vaquero, A Ortega, J Villalba, A Miguel, E Lleida, Confidence measures for speaker segmentation and their relation to speaker verification. in *Proc. Interspeech'10* (Makuhari, Japan, 2010), pp. 2310–2313
26. A Hatch, A Stolcke, Generalized linear kernels for one-versus-all classification: application to speaker recognition. in *Proc. ICASSP'06* vol. 5 (Toulouse, France, 2006), p. V
27. X Anguera, J Hernando, J Anguita, XBIC: nueva medida para segmentación de locutor hacia el indexado automático de la señal de voz. in *III Jornadas en Tecnología del Habla* (Valencia, Spain, 2004), pp. 237–242
28. R Auckenthaler, M Carey, H Lloyd-Thomas, Score normalization for text-independent speaker verification systems. Digital Signal Process. **10**(1-3), 42–54 (2000)
29. H Hermansky, N Morgan, RASTA processing of speech. IEEE Trans. Speech Audio Process. **2**(4), 578–589 (1994)
30. J Pelecanos, S Sridharan, Feature warping for robust speaker verification. in *Proc. 2001: A Speaker Odyssey—The Speaker Recognition Workshop (Odyssey-2001)* (Crete, Greece, 2001), pp. 213–218
31. N Dehak, P Kenny, R Dehak, P Dumouchel, P Ouellet, Front-end factor analysis for speaker verification, 788–798, (2011)
32. N Dehak, R Dehak, J Glass, D Reynolds, P Kenny, Cosine similarity scoring without score normalization techniques. in *Proc. Odyssey 2010: The Speaker and Language Recognition Workshop* (Brno, Czech Republic, 2010), pp. 71–75
33. N Mirghafori, C Wooters, Nuts and flakes: a study of data characteristics in speaker diarization. in *Proc. ICASSP'06* vol. 1 (Toulouse, France, 2006), p. I
34. J Fiscus, A Le, G Sanders, MDE Tasks and Results (2004), http://www.itl.nist.gov/iad/mig/tests/rt/2004-fall/rt04f-mde-nist.pdf. Accessed 30 May 2011

doi:10.1186/1687-4722-2012-19
Cite this article as: Zelenák *et al.*: Speaker diarization of broadcast news in Albayzin 2010 evaluation campaign. *EURASIP Journal on Audio, Speech, and Music Processing* 2012 **2012**:19.