New bilingual speech databases for audio diarization

David Tavarez, Eva Navas, Daniel Erro, Ibon Saratxaga, Inma Hernaez

Aholab - Dept. of Communications Engineering. Faculty of Engineering.
University of the Basque Country. Alda. Urquijo s/n 48013 Bilbao
email: david, eva, derro, ibon, inma@aholab.ehu.es

Abstract

This paper describes the process of collecting and recording two new bilingual speech databases in Spanish and Basque. They are designed primarily for speaker diarization in two different application domains: broadcast news audio and recorded meetings. First, both databases have been manually segmented. Next, several diarization experiments have been carried out in order to evaluate them. Our baseline speaker diarization system has been applied to both databases with around 30% of DER for broadcast news audio and 40% of DER for recorded meetings. Also, the behavior of the system when different languages are used by the same speaker has been tested.

Keywords: Spoken Language Resource, Speaker Diarization, Bilingual Speech Database

1. Introduction

Speaker diarization is the process of identifying "who spoke when" in an audio recording, mainly without any previous information about the number of speakers present in the recording, the types of audio included or the amount of speech in the audio (Anguera Miro et al., 2012). This process usually requires the detection of speaker changes and the subsequent identification of each different speaker within the resulting speech segments. Speaker diarization has been primarily used in two different application domains:

- Broadcast news audio: Radio and television programs that usually contain commercial breaks, music, interviews, voice with different types and levels of background noise etc. They are generally recorded on a single channel. Speech has different levels of quality depending on the origin of the recording; it ranges from clean studio speech to speech recorded at the street with background noise and even telephone speech in some cases.
- Recorded meetings: Meetings and conferences where several people interact in the same room. Recordings are predominantly made with multiple microphones and contain different noises. The number of speakers is usually smaller in recorded meetings than in broadcast audio, but speaker turns are shorter due to the spontaneity of meeting speech, with many interruptions between participants and overlapping speech segments.

As it occurs in any other speech technology, there is a need for transcribed data in order to develop quality speaker diarization systems. In the last years, important efforts have been done to collect and transcribe several databases, such as the ISL audio corpus (Burger et al., 2002), the ICSI meeting corpus (Janin et al., 2003) and the AMI meeting corpus (Carletta et al., 2006) for the meeting environment and the COST278 broadcast database (Vandecatseye et al., 2004), the DiSCo German corpus (Baum et al., 2010) and the Catalan 3/24 database (Zelenák et al., 2012) for the broadcast audio domain. These diarization databases are mostly monolingual. The ones that include speech in more than one language, like COST278 database, have different speakers for different languages. The databases presented in this paper include bilingual speakers recorded both in Spanish and Basque. There are multilingual databases that include speakers talking in more than one language, however they have not been designed with audio diarization purposes but to study code switching between different language pairs (Han-Ping Shen et al., 2011), to perform speech and language recognition in non-native languages (Imseng et al., 2012) or to evaluate speaker recognition systems over the telephone in multilingual environments (Cieri et al., 2004).

The main purpose of the work described in this paper is the collection and recording of two new speech databases which will support the development and implementation of multilingual diarization systems for both main application domains.

The paper is organized as follows: it begins by describing the main characteristics of the two new databases, their recording purposes and the labeling available for each of them. Next, some diarization experiments performed in order to evaluate each database are presented. Finally, the paper concludes with a brief discussion on the performed work and possible future uses and availability of the databases.

2. Ahonews Database

The audio material for this broadcast news database was provided by ETB Television channel, the Basque public TV network. It consists of a collection of news clips in Spanish and Basque corresponding to the news broadcast in 2010. In addition to the voice of the journalists who tell the news, the audio files include interviews, music, noise, and simultaneous translation over the original audio. Together with each audio file, ETB provided a text file with information about the journalist who wrote the piece, the subject addressed and the script followed by the broadcaster. Unfortunately the completion level of these files was very variable: some of them were empty while others provided accurate information.

The recordings from Ahonews database were provided by the ETB TV channel in Windows PCM audio files (wav). Every sound file uses 16 bit samples and 48 kHz sampling rate. All the recordings are mono and have a length of about 2 minutes.

For the creation of the database, the news clips corresponding to the first fortnight have been used. Every audio file has been manually segmented, labeled and stored according to both the main speaker's identity and the used language, facilitating in this way the access to different audio files from the same speaker in two different languages. The files with no available information about the speaker are organized based only on the used language.

Wavesurfer label file format has been selected for the transcriptions. A text file with .lab extension is provided for each recording. This label file contains the time marks for different acoustic events such as speaker turns, silences, certain background conditions, etc. stored as follows:

START-TIME END-TIME ID

where start/end time values indicate the boundaries of the acoustic event in seconds, and ID is a character determining one the following categories of the acoustic event:

- a: speech from the main speaker of the clip (the journalist who tells the news). If known, the name of the journalist is used as label instead.
- o: speech from other speakers who take part in the audio file.
- e: peculiarities related to background conditions (for instance, different noises, music and overlapped speech).
- s: silence segments corresponding either to absence of signal or to breath pauses.

In total, 449 files have been manually labeled, 321 of which have a known speaker (177 in Spanish and 144 in Basque) and 128 an unidentified main speaker (86 in Spanish and 42 in Basque). Taking into account the files available in both languages, the number of files per speaker varies between 1 and 14.

Ahonews database contains separate clips organized by speaker and language. This fact allows the easy generation of artificial broadcast sessions where the number of speakers and the language used can be controlled. In fact, part of these files have already been concatenated in order to create three artificial sessions with different acoustic characteristics. None of these sessions include overlapped speech.

- *Ahonews_1*: The first session is 20 minutes long and includes speech from 9 different speakers with long appearances in low noise conditions.
- *Ahonews_2*: The second session has a total duration of 25 minutes with 40 different speakers who have short duration appearances, and includes also segments with background noise and music.
- *Ahonews_3*: To analyze the effect of using different languages during the diarization process, a third artificial session has also been created. This session is shorter than the previous ones (only 14 minutes long) and has only 4 different speakers with just two appearances per speaker, one in Spanish and one in Basque. Similarly to the Ahonews_1 session, Ahonews_3 includes long appearances of each speaker in low noise conditions.

Table 1 summarizes the main characteristics of Ahonews database. The first four lines contain general features of the database, including the different languages used by the speakers, the gender distribution of these speakers, the size of the database and its main purpose, which in this case refers to the application domain. The last two lines contain the minimum and maximum number of different speakers and turn changes that can be found in the database files.

Database	**Ahonews**
Language	Spanish & Basque
Purpose	News diarization
Gender distribution	46 female & 27 male
Total size	8 hours
Num. spk. per clip	1-5
Speaker turns per file	1-14

Table 1: Main characteristics of Ahonews database

3. Ahomeetings Database

This database was recorded in a meeting room inside our laboratory during 2013. It consists of separated sessions where different speakers present and discuss several topics, mainly in Spanish, although Basque is also used by some of them.

In a first approach, a professional audio mixer was used to feed and connect two high-quality boundary microphones to make the recordings. The signals from the two microphones were then stored in a stereo file using a portable PC. Due to some issues related to the sudden drop of the signal level after several minutes of recording, we decided to switch to the recording configuration shown in Figure 1, where a Kinect sensor is used to capture the audio. The Kinect sensor includes a four element linear supercardioid microphone array (shown in purple in Figure 1), which enables advanced signal processing including acoustic source localization, echo cancellation and noise suppression. This can be used to improve the performance of the diarization task and also simplifies the recording system by eliminating the need of an audio mixer.

For the Ahomeetings database, the signals from the four microphones were recorded at 16 kHz sampling rate with 16 bits of resolution and stored in quadraphonic files. Each session has a length of about 30 minutes.

The database has been recorded along several months. In order to keep the acoustic characteristics of the recordings unaltered, all the speakers had to hold the same place in each session. The position of the Kinect sensor was also

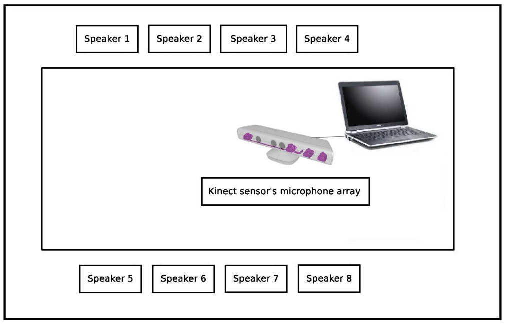

Figure 1: Audio set up for the Ahomeetings database acquisition

kept constant and the signal levels were tested at the beginning of each recording session to ensure that the recording conditions were unchanged. Not all the speakers are present in every session, thus the number of speakers varies between 5 and 10 per session.

To label the signals, an automatic segmentation process has been applied to locate the appearances of each speaker (Luengo et al., 2010). Then, a manual review has been completed and notes about specific background disturbances have been included in the label files. As the applied automatic segmentation system is able to assign only one speaker label at each time instant, overlapping speech segments have also been identified during the subsequent manual labeling process.

Wavesurfer label file format has been used for the Ahomeetings transcriptions, with the ID character determining one of the following categories of acoustic event:

- spkN: speech from a speaker identified by number N. Due to the automatic segmentation process, some speakers might have different IDs in each session they took part in. It is also possible that one speaker is identified by more than one ID within a session. Finally, in some cases, one ID might correspond to more than one speaker.

- o: overlapped speech segments produced by multiple speakers talking at the same time. Just the presence of overlap is indicated, without taking into account the individual speakers involved in it.

- n: specific disturbances related to background conditions and noises (knock on door or table, chair moving, laughs...).

- s: silence segments corresponding either to absence of signal or to breath pauses.

Table 2 summarizes the most important characteristics of Ahomeetings database. In this database, as usually corresponds to a meeting speech database, there are fewer different speakers, with more interaction among them which reflects in the higher number of speakers turns compared with Ahonews database.

Database	**Ahomeetings**
Language	Spanish & Basque
Purpose	Meetings diarization
Gender distribution	2 female & 8 male
Total size	4,5 hours
Num. spk. per file	5-10
Spk. turns per file	200-701

Table 2: Main characteristics of Ahomeetings database

4. Diarization Experiments

In order to evaluate the new databases gathered, a previously assessed diarization system has been applied to both of them and the results obtained have been analyzed with detail.

In these experiments, the diarization system built by our research group for the Albayzin 2010 speaker diarization evaluation campaign (Luengo et al., 2010) has been used. The applied algorithm is based on an efficient implementation of a BIC change detector and an off-line speaker clustering, and it was the top-performing system in Albayzin 2010 evaluation with 30% DER (Diarization Error Rate) (Zelenák et al., 2010). At the above mentioned campaign a Catalan broadcast news database from the 3/24 TV channel was used (Zelenák et al., 2012). Therefore, due to the similarities of both databases, the experiments developed with Ahonews database were expected to give similar DER as the ones obtained with the Catalan database. In the case of Ahomeetings database, the different characteristics of broadcast audio and audio from recorded meetings were likely to degrade the results of the diarization process.

The obtained results of the diarization system for the three Ahonews sessions are shown in Table 3. Due to the short length of Ahonews news clips (they are around 2 minutes long), Ahonews sessions have been used in this test in order to obtain more reliable DER values. These values have been calculated according to the criteria defined by NIST.

Session	DER
Ahonews_1	35.65%
Ahonews_2	26.83%
Ahonews_3	1.73%

Table 3: Results of the baseline diarization system for the Ahonews sessions

As displayed in Table 3, the two first sessions get DER values comparable to the one obtained in the Albayzin 2010 evaluation campaign. However, it can also be seen that the DER value for Ahonews_1 session is significantly higher. This is due not to the Ahonews database itself, but to the differences between the acoustic characteristics of Ahonews sessions (mainly the number of speakers) and the default configuration of the diarization system.

On the other hand, the result achieved for Ahonews_3 session is close to ideal. Ahonews_3 is a rather less complex session than Ahonews_1 and Ahonews_2, with just 4 different speakers and only 2 turns per speaker, which results in an easier task to be solved. This session has been designed to analyze the behavior of the system when different languages are used by the same speaker, which does not seem to be a hard task in this case, given the low 1.73% value of DER obtained for this session. Nevertheless, additional experiments must be performed in order to consolidate this conclusion.

The baseline diarization system has also been used to evaluate the Ahomeetings database. This database has very different acoustic characteristics compared to Ahonews database, with background noise, reverberation, variable speech levels and lower signal to noise ratio, thus it can be presumed that the result for this database will be far from satisfactory. In Table 4, the results obtained with Ahomeetings first session (all four channels) are shown.

Session	DER
channel_1	77.03%
channel_2	69.51%
channel_3	72.08%
channel_4	71.47%

Table 4: Results of the baseline diarization system for the Ahomeetings first session

It can be confirmed that the DER value for the first Ahomeetings session is significantly higher than the one obtained for Ahonews database. To analyze the the origin of this high DER value, it has been studied with more detail. According to NIST criteria, DER is composed of three different errors: missed speaker time, percentage of time that a reference speech segment has been labeled as non-speech by the system; false alarm speaker time, percentage of time that the system has labeled as speech from a speaker a segment that corresponded to non-speech in the reference and speaker error time, percentage of time that an incorrect speaker ID has been assigned by the system. In Table 5, the three different sources of the total diarization error are detailed for channel_2 (all channels had a similar error distribution) and compared with the detailed error made by the baseline diarization system in Albayzin 2010 evaluation campaign. The source of error that has deteriorated the most is the missed speaker time. As the baseline diarization system has been designed for broadcast news audio, the VAD (Vocal Activity Detection) step is not suited to correctly detect the noises and silences of a meeting recording. Overlapping speech segments also contribute to the missed speaker time part of the diarization error. In the first Ahomeetings session 12% of data corresponds to overlapped speech. Due to the fact that the baseline system does not take into account that several speakers may be talking at the same time, all the remaining speakers that are not labeled in these segments are considered as missed speaker time when calculating the DER.

Diarization Error	channel_2	Alb. 2010
Missed Speaker Time	29.7%	2.80%
False Alarm Speaker Time	7.7%	2.20%
Speaker Error Time	32.1%	25.10%
DER	69.51%	30.11%

Table 5: Detailed results of the baseline diarization system for the Ahomeetings first session, channel 2 compared to the errors made with Albayzin 2010 database

5. Conclusions

Two new bilingual speech resources in Spanish and Basque for speaker diarization have been presented. Ahonews database includes around 8 hours of broadcast audio coming from news programs and Ahomeetings database is 4,5 hours long and contains speech from recorded meetings. These databases will support the development and implementation of diarization systems in two different application domains and the fulfillment of bilingual experiments. A preliminary diarization experiment made with our baseline diarization system has provided the first results for both databases: around 30% DER for Ahonews and 70% DER for Ahomeetings. These databases will be made available via METASHARE repository (`http://www.meta-share.eu/`).

6. Acknowledgments

The authors would like to thank Iker Luengo for the development of the baseline diarization system used in the experiments. This work has been partially supported by UPV/EHU (Ayudas para la Formación de Personal Investigador), the Basque Government (Ber2tek project, IE12-333) and the Spanish Ministry of Economy and Competitiveness (SpeechTech4All project, TEC2012-38939-C03-03).

7. References

Xavier Anguera Miro, Simon Bozonnet, Nicholas Evans, Corinne Fredouille, Gerald Friedland, and Oriol Vinyals. 2012. Speaker Diarization : A Review of Recent Research. *IEEE Transactions on Audio, Speech and Language Processing*, 20(2):356–370.

Doris Baum, Daniel Schneider, Rolf Bardeli, Jochen Schwenninger, Barbara Samlowski, Thomas Winkler, and Joachim Köhler. 2010. DiSCo-a German evaluation corpus for challenging problems in the broadcast domain. In *7th International conference on Language Resources and Evaluation (LREC)*, pages 1695–1699, Valletta, Malta.

Susanne Burger, Victoria MacLaren, and Hua Yu. 2002. The ISL meeting corpus: the impact of meeting type on speech style. In *7th International Conference on Spoken Language Processing (ICSLP)*, pages 301–304, Denver, Colorado, USA.

Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Mael Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, Guillaume Lathoud, Mike Lincoln, Agnes Lisowska, Iain McCowan, Wilfried Post, Dennis Reidsma, and Pierre Wellner. 2006. The AMI Meeting Corpus: A Pre-announcement. In *Machine Learning for Multimodal Interaction Lecture Notes in Computer Science Volume 3869*, pages 28–39, Edinburgh, UK. Springer-Verlag.

Christopher Cieri, Joseph P. Campbell, Hirotaka Nakasone, David Miller, and Kevin Walker. 2004. The mixer corpus of multilingual, multichannel speaker recognition data. In *Proc. 4th International Conference on Language Resources and Evaluation (LREC)*, pages 26–28, Lisbon, Portugal.

Han-Ping Shen, Chung-Hsien Wu, Yan-Ting Hsu, and Yang Chun-Shan. 2011. CECOS: A Chinese-English code-switching speech database. In *International Conference on Speech Database and Assessments (Oriental COCOSDA)*, pages 120–123, Hsinchu, Taiwan.

David Imseng, Hervé Bourlard, Holger Caesar, Philip N. Garner, Gwénolé Lecorvé, and Alexandre Nanchen. 2012. MediaParl: Bilingual mixed language accented speech database. In *IEEE Spoken Language Technology Workshop (SLT)*, pages 263–268, Miami, USA.

Adam Janin, Adam Baron, Jane Edwards, Dan Ellis, David Gelbart, Nelson Morgan, Barbara Peskin, Thilo Pfau, Elizabeth Shriberg, Andreas Stolcke, and Chuck Wooters. 2003. The ICSI meeting corpus. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 364–367, Hong Kong.

Iker Luengo, Eva Navas, Ibon Saratxaga, Inma Hernáez, and Daniel Erro. 2010. AhoLab Speaker Diarisation System for Albayzin 2010. In *FALA 2010*, pages 393–396, Vigo, Spain.

An Vandecatseye, Jean-Pierre Martens, Joao Neto, Hugo Meined, Carmen Garcia-Mateo, Javier Dieguez, France Mihelic, Janez Zibert, Jan Nouza, Petr David, Matus Pleva, Anton Cizmar, Harris Papageorgiou, and Christina Alexandris. 2004. The COST278 Pan-European Broadcast News Database. In *4th international conference on Language Resources and Evaluation (LREC)*, pages 873–876, Lisbon, Portugal.

Martin Zelenák, Henrik Schulz, and Javier Hernando. 2010. Albayzin 2010 Evaluation Campaign: Speaker Diarization. In *FALA 2010*, pages 301–304, Vigo, Spain.

Martin Zelenák, Henrik Schulz, and Javier Hernando. 2012. Speaker diarization of broadcast news in Albayzin 2010 evaluation campaign. *EURASIP Journal on Audio, Speech, and Music Processing*, pages 1–9.