IberSPEECH 2021
24-25 March 2021, Valladolid, Spain

The Biometric Vox System for the Albayzin-RTVE 2020 Speaker Diarization and Identity Assignment Challenge

Roberto Font[1], *Teresa Grau*[1]

[1]Biometric Vox S.L.

roberto.font@biometricvox.com, teresa.grau@biometricvox.com

Abstract

This paper describes the systems developed by Biometric Vox for the Albayzin Speaker Diarization Challenge organized as part of the Iberspeech 2020 conference. The two systems (primary and contrastive) we developed for the challenge are based on Deep Neural Network x–vector embeddings and a Probabilistic Linear Discriminant Analysis (PLDA) backend. The resulting x-vectors are grouped using Agglomerative Hierarchical Clustering (AHC) in order to obtain the diarization labels. Systems differ in the resegmentation stage. Our primary system achieves 14.96% DER on the test set of the RTVE2018 database and 21.35% on the 2020 evaluation set.

Index Terms: speaker diarization, speaker embeddings, x-vectors, speaker identification

1. Introduction

Speaker diarization is the task of segmenting a speech audio, marking speaker change points and categorizing those segments according to the speaker identity; in other words, speaker diarization answers the question of who speaks when.

The aim of this paper is to provide a complete description of the systems developed by Biometric Vox for the Albayzin-RTVE 2020 Speaker Diarization and Identity Assignment Challenge. The task for this challenge is to group together all speech segments belonging to the same speaker and, in case the speaker is known to the system, assign the corresponding identity. The material for the challenge consists in TV shows that belong to diverse genres and broadcast by the public Spanish Television (RTVE), covering a wide range of realistic and challenging conditions: spontaneous and read speech, different accents, background noise, songs, laughs, yells, quick short-turn conversations, etc.

Our two systems are based on the DNN x–vector paradigm [1] and share the same basic building blocks, which are described in more detail in Section 3:

- Acoustic feature extraction (23 MFCC features) and energy-based Voice Activity Detection (VAD).
- X–vector embedding extraction.
- Embedding post-processing: length-normalization, centering, whitening, Linear Discriminant Analysis (LDA).
- Probabilistic Linear Discriminant Analysis (PLDA) scoring.
- Agglomerative Hierarchical Clustering.

The primary system performs two diarization stages: the segments resulting from the first one are fed to a music/speech/noise classifier and music segments are removed. Once music segments are excluded, a second diarization stage is performed to obtain the final result. This system achieves a Diarization Error Rate (DER) of 21.35% on the 2020 evaluation set.

The contrastive system focuses on refining the speaker change boundaries, which we have found that can improve performance when there are long-turn conversations with no frequent speaker changes. This system obtains 31.57% DER on the 2020 evaluation set.

The rest of the paper is organized as follows: in Section 2 the RTVE database, the data used to train the embedding extractor and the data used as development and evaluation sets are described. Section 3 presents the basic components of our two systems and Section 4 describes the diarization process and characteristics for those systems. Section 5 covers our submitted system for the identity assignment task. Results are shown in Section 6, and, finally, we summarize our results and conclusions in Section 7.

2. Data Resources

2.1. RTVE database

The RTVE2018 database [1] has a total of 569 hours and 22 minutes of audio extracted from 17 different TV shows broadcast by RTVE (Radio Televisión Española) from 2015 to 2018. Most shows are related to news, debates, social gatherings and documentaries. The database is divided into 4 subsets: a training set, 2 development sets, *dev1* and *dev2*, and a test set, as summarized in Table 1. Around 37 hours, divided among the *dev2* set and a subset of the test set, have diarization and speech segmentation labels available. Additionally, the *dev2* set counts with several audio files for a limited set of speakers to allow for the creation of speaker models.

Table 1: *Composition of the different RTVE subsets.*

Subset	# hours	# different shows	# speaker models
RTVE2018			
train	500h	16	-
dev1	52:31:51	5	-
dev2	15:09:25	2	34
test	39:07:15	8	39
RTVE2020			
dev	03:55:31	2	18
test	67:23:29	15	161

The RTVE2020 database [2] is also a collection of TV shows from the public Spanish Television (RTVE) that were broad-

[1]http://catedrartve.unizar.es/reto2018/RTVE2018DB.pdf
[2]http://catedrartve.unizar.es/reto2020/RTVE2020DB.pdf

10.21437/IberSPEECH.2021-18

cast from 2018 to 2019. The database is composed of 70 hours and 18 minutes of audio belonging to 15 different shows (dev's shows are also in the test set). Only two shows are in common with the 2018 dataset: *Comando Actualidad* and *Millenium*. The database is divided into 2 subsets: *dev* and *test*. The latter one is the challenge evaluation set. Shows' genre for this set is related to entertainment: series, sketches, reports, entertainment shows, etc.

TV show's genre and format are quite different between the two sets; each set covers a different range of conditions. For the RTVE2018 database: spontaneous and read speech, long-turn conversations, different accents and background noise. And for RTVE2020: songs, laughs, yells, quick short-turn conversations and colloquial vocabulary.

2.2. Training Data

Our training material consists of the following datasets:

- NIST SRE 04-10
- MIXER6 as prepared by the Kaldi sre16 recipe.
- Switchboard phase1-3 and cellular1-2.
- Voxceleb 1 and 2. [2] [3]

We augment training data by generating four additional perturbed versions of each file using the Musan [4] corpus by adding:

- Reverberation
- Musan noise
- Musan music
- Musan speech

For the embedding extractor training, the training set consists of NIST SRE 04-10, MIXER6, Switchboard, VoxcelebCat and all augmented data from these datasets. VoxcelebCat is the result of concatenating all excerpts from the same video into one longer file and combining Voxceleb 1 train, Voxceleb 2 dev and Voxceleb 2 test. The samples from NIST SRE 04-10, MIXER6 and Switchboard were upsampled to 16kHz.

The total number of utterances is $1,109,458$ from a total of $13,682$ speakers.

2.3. Development and Evaluation Data

In order to evaluate the speaker diarization performance of our systems, we used the datasets provided by the organizers for this challenge:

- RTVE2018DB dev2, with 12 episodes from 2 shows, which are manually transcribed and with speaker time references for all episodes and identity references for one episode. Includes an enrollment set with audio files provided for 34 speakers.
- RTVE2018DB test, with 61 episodes from 9 shows. A subset of 40 of these episodes have speaker time references for speaker diarization available. The enrollment set has 39 speakers.
- RTVE2020DB dev, with 9 episodes from 2 shows with speaker time references and identity references for some of the speakers in each episode. Also includes an enrollment set with audio files provided for 18 speakers.

The 2020 evaluation set is composed of 87 episodes from 15 shows. Of those, 54 episodes from 10 shows are used for the speaker diarization challenge. Additionally, an enrollment set for 161 speakers is provided.

3. System Components

The diarization system, based on x-vector neural embeddings [1] and a PLDA backend, was implemented using the Kaldi [5] toolkit. The rest of this section provides a more detailed description of the main building blocks.

3.1. Feature Extraction and Voice Activity Detection

The input of the embedding extractors are 23-dimensional Mel Frequency Cepstral Coefficients (MFCCs) which are extracted from 25 ms windows with 15 ms overlap. Features are normalized using cepstral mean and variance normalization over a sliding-window of 300 frames.

Initial segmentation and silence removal is made using Kaldi standard energy-based VAD.

3.2. Embedding Extractor

To train the embedding extractor we used the baseline TDNN x-vector architecture as in the Kaldi SRE 16 recipe (Table 2).

Table 2: *Baseline x-vector architecture.*

Layer type	**Layer context**	**Size**
TDNN-ReLU-batchnorm	t-2:t+2	512
TDNN-ReLU-batchnorm	t-2, t, t+2	512
TDNN-ReLU-batchnorm	t-3, t, t+3	512
ReLU-batchnorm	t	512
ReLU-batchnorm	t	1500
Stats Pooling (mean+stddev)	T	2x1500
ReLU-batchnorm		512
ReLU-batchnorm		512
Softmax		# speakers

The model was trained for 3 epochs, with batch size of 64, on an Nvidia GeForce GTX 2080.

3.3. Back-end

All systems use a back-end that follows a classical LDA-PLDA scoring scheme:

- Embeddings are projected to unit length, centered and whitened.
- Linear discriminant analysis (LDA) is used to project the embeddings to a lower dimension. (From 512 to 150 in our case.)
- The segments are scored using PLDA.

Both LDA and PLDA are trained on VoxcelebCat. No score normalization or calibration was applied.

4. Diarization Systems

In this section, we provide a description of our submitted diarization systems. As discussed above, both systems share the same basic building blocks. However, the primary system performs two diarization stages to try and remove all music segments, which is a challenging feature of the RTVE dataset, while the contrastive system performs a resegmentation around the speaker change points to refine the segment boundaries.

4.1. Primary system

The speaker diarization process starts with feature extraction: 23-dimensional MFCC features are extracted, and Cepstral Mean Variance Normalization (CMVN) sliding window is applied before performing a VAD segmentation to remove silence portions.

DNN x-vectors embeddings are extracted for the resulting segments. A sub-segmentation with a sliding window of 3 seconds and a second and a half hop is used. The embeddings are clustered using Agglomerative Hierarchical Clustering (AHC) with single linkage.

The segments resulting from this first diarization stage are passed through a music/speech/noise classifier and music segments are removed. Finally, a second diarization stage is performed to obtain the final diarization result. Figure 1 shows the flowchart of our primary diarization system.

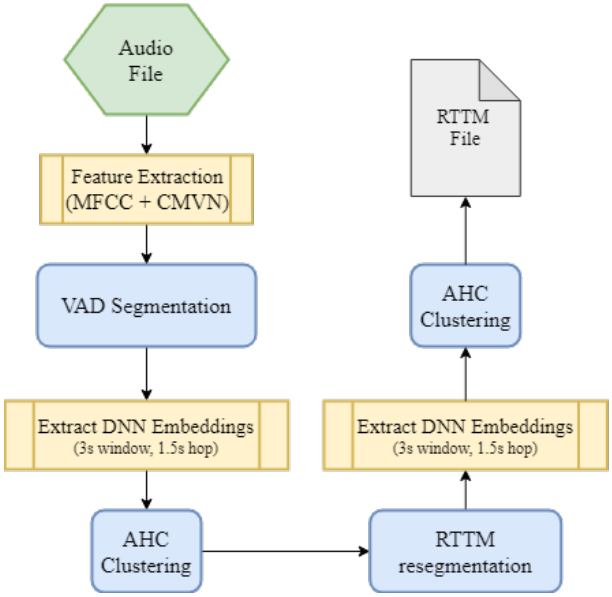

Figure 1: *Primary system's flowchart*

4.2. Contrastive system

The contrastive system uses the same feature extraction and VAD segmentation than the primary system, but no subsegmentation is used for embedding extraction. Embeddings are extracted for the whole segments resulting from the VAD segmentation. This embeddings are clustered using Agglomerative Hierarchical Clustering with single linkage. Then, a resegmentation algorithm refines the speaker transition boundaries: DNN x-vectors embeddings with a sliding window of 3 seconds and a second and a half hop are extracted only for the segments involved in a speaker transition. The embeddings are then clustered until the number of clusters is two. Figure 2 illustrates the process to refine the speaker boundaries.

5. Identity Assignment

If there is prior knowledge about the identity of the people involved in an audio, it can be used to assign a name to each diarization label output by the speaker diarization system.

The steps for the identity assignment process in our submitted system are the following:

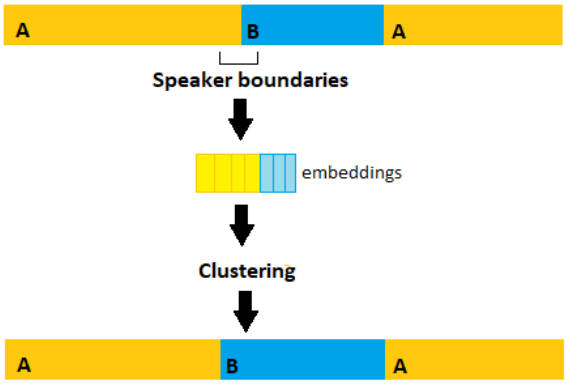

Figure 2: *Speaker boundaries refinement*

1. Creating speaker models. To that end, DNN x-vector embeddings are extracted from the audio files provided for each person in the enrollment set and averaged so that each speaker model is the average of the embedding for all enrollment files available for that speaker.
2. Performing speaker diarization with our primary system.
3. Extracting an x-vector embedding for each final segment obtained as a result of the diarization process.
4. Comparing each segment's x-vector with the speaker models by computing log-likelihoods ratios using the same back-end described in Section 3.3.
5. Assigning the identity of the best-matching speaker model to the segment only if it exceeds a threshold that was tuned on the development set.

6. Results

Table 3 shows the results for our two systems on the development and test portions of the RTVE2018DB.

Table 3: *DER (%) on the RTVE2018DB. MISS column is for the missed speaker time, FA for false alarm speaker time and SPK for speaker error time.*

System	MISS	FA	SPK	DER
Dev2				
Primary	2.1	1.4	4.1	7.68
Contrastive	1.9	1.7	7.0	10.59
Test				
Primary	2.7	3.5	8.7	14.96
Contrastive	0.6	3.9	14.6	19.11

The evaluation results provided by the organizers for the 2020 Speaker Diarization Challenge's evaluation set are shown in Table 4. Table 5 presents the 2020 evaluation results by TV show.

The best results are obtained for the shows *Millennium* (ML) and *Los desayunos de TVE* (LD), which revolve around news content and long-turn conversation almost without interruptions between speakers. For the entertainment domain, on the other hand, results suggest that improving the speech detection and music/noise removal stage could reduce the miss

speech and false alarm. This could also help to obtain a better segmentation and reduce the speaker error time.

For the identity assignment task, two additional metrics are used to evaluate the system performance: Assignment Error Rate (AER) and Average Speaker Error (ASE). For our system, the AER on the test data in the RTVE2018 dataset was 37.39% and for the ASE, 39.17%. The DER for the identity assignment evaluation set provided by the organizers is shown in Table 6. These results by show are consistent with those obtained for the speaker diarization task.

The processing time for the subset dev2 was 6 hours and 31 minutes on an Intel(R) Core(TM) i7-5820K CPU @ 3.30GHz with 6 cores.

Table 4: *DER (%) RTVE2020 Evaluation results. MISS column is for the missed speaker time, FA for false alarm speaker time and SPK for speaker error time.*

System	MISS	FA	SPK	DER
Primary	4.0	4.8	12.5	21.35
Contrastive	2.7	10.3	18.5	31.57

Table 5: *DER (%) RTVE2020 Evaluation results by TV show on our primary and contrastive systems for the Speaker Diarization task. TV shows: Aquí la Tierra (AT), Boca Norte (BN), Bajo la Red (BR), Comando Actualidad (CA), Ese Programa del que Usted me Habla (EP), Los desayunos de TVE (LD), Millennium (ML), Never Films Mira Ya (NFMY), Si Fueras Tu (SFT) and Wake-Up (WU)*

TV Show	Primary	Contrastive
AT	18.70	34.51
BN	100.70	137.54
BR	70.08	109.17
CA	35.55	55.84
EP	24.26	39.38
LD	13.60	12.56
ML	10.93	12.99
NFMY	64.63	146.97
WU	78.83	133.52

Table 6: *DER (%) RTVE2020 Evaluation results by TV show for the Identity Assigment task.*

TV Show	Primary
AT	97.33
BN	153.52
BR	123.18
CA	100.67
EP	64.00
LD	49.32
ML	80.28
NFMY	126.38
WU	112.03
Global	65.09

7. Conclusions

We have presented our diarization systems submitted to the Albayzin Speaker Diarization and Identity Assignment Challenge and reported the results on the development and evaluation sets. This challenge focuses on grouping together all speech segments belonging to the same speaker on TV shows, and assigning the identity in case the speaker is known to the system, a highly challenging task especially in the detection of the number of speakers and dealing with a wide range of realistic conditions such as background noise/music and overlapped speakers.

We submitted two systems for the speaker diarization task and one for the identity assignment subtask. The primary system focuses on performing a better music removal stage, which helps on the clustering stage to obtain better results. The contrastive system focuses on refining the boundaries between speakers. Unsurprisingly, this latter process, geared towards conversations with long turns, a condition that is not present in most of the shows under consideration, obtained a higher DER than our primary, more general-purpose, system.

8. Acknowledgements

The authors would like to thank the organizers of the Albayzin Challenge.

9. References

[1] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," 04 2018, pp. 5329–5333.

[2] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: A large-scale speaker identification dataset," in *Proc. Interspeech 2017*, 2017, pp. 2616–2620.

[3] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," in *Proc. Interspeech 2018*, 2018, pp. 1086–1090.

[4] D. Snyder, G. Chen, and D. Povey, "MUSAN: A music, speech, and noise corpus," 2015.

[5] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, Dec. 2011, iEEE Catalog No.: CFP11SRW-USB.