ODESSA at Albayzin Speaker Diarization Challenge 2018

Jose Patino[1], *Héctor Delgado*[1], *Ruiqing Yin*[2], *Hervé Bredin*[2], *Claude Barras*[3] *and Nicholas Evans*[1]

[1]EURECOM, Sophia Antipolis, France
[2]LIMSI, CNRS, Université Paris-Saclay, Orsay, France
[3]LIMSI, CNRS, Univ. Paris-Sud, Université Paris-Saclay, Orsay, France

[1]`firstname.lastname@eurecom.fr`, [2,3]`firstname.lastname@limsi.fr`

Abstract

This paper describes the ODESSA submissions to the Albayzin Speaker Diarization Challenge 2018. The challenge addresses the diarization of TV shows. This work explores three different techniques to represent speech segments, namely binary key, x-vector and triplet-loss based embeddings. While training-free methods such as the binary key technique can be applied easily to a scenario where training data is limited, the training of robust neural-embedding extractors is considerably more challenging. However, when training data is plentiful (open-set condition), neural embeddings provide more robust segmentations, giving speaker representations which lead to better diarization performance. The paper also reports our efforts to improve speaker diarization performance through system combination. For systems with a common temporal resolution, fusion is performed at segment level during clustering. When the systems under fusion produce segmentations with an arbitrary resolution, they are combined at diarization hypothesis level. Both approaches to fusion are shown to improve diarization performance.

Index Terms: speaker diarization, diarization fusion, neural embeddings, binary key

1. Introduction

Albayzin evaluations cover a range of speech processing related tasks that include search on speech, audio segmentation, speech-to-text transcription, or speaker diarization. The latter motivates the work reported on this paper. Speaker diarization is the task of processing an audio stream into speaker homogeneous clusters. Traditionally considered as an enabling technology, a number of potential applications can benefit from speaker diarization as a pre-processing step, such as automatic speech recognition [1], speaker recognition and identification [2], or spoken document retrieval. The increasing maturity of these technologies calls for continuous improvement in speaker diarization that has accordingly been the objective of numerous evaluations and campaigns, be it in the context of the NIST RT evaluations [3], the more recent multi-domain DIHARD challenge [4], or in the well-established Albayzin Speaker Diarization evaluations [5]. The current edition of the Albayzin Speaker Diarization Challenge [6] includes audio content from the recently released RTVE2018 database [7], composed of TV shows from a range of topics broadcast on the Spanish TV public network. Further details can be found in [6, 7].

This work has been produced in the context of the ODESSA[1] project, which is focused on improving speaker diarization performance by leveraging recent developments in the task of text-independent speaker recognition. Efforts were made by the different members of the consortium to improve the reliability of their own speaker diarization systems. In doing so, different speaker modelling techniques were used, on both the closed- and open-set conditions of the evaluation. For the closed-set condition, binary key speaker modelling [8] offers a training-free option that has produced competitive performance in previous editions of the challenge [9]. Triplet-loss neural embeddings [10] trained on the provided data were also explored. Experiments in the open-set condition include embeddings in the form of state-of-the-art text-independent speaker recognition x-vector [11]. Different clustering techniques were also explored across the training conditions and speaker modelling techniques.

Finally, and motivated by the access to significantly different approaches to speaker modelling and diarization that could potentially offer complementary solutions, the main contribution of this work lies in the exploration of different fusion techniques for speaker diarization. Whereas fusion, for example at score level, is often applied as a mean of increasing robustness in closely related tasks such as speaker recognition, the problem of merging clustering solutions for speaker diarization scenarios remains challenging. Diarization hypothesis level fusion is applied following a label merging approach [12]. A segment-level fusion technique similar to that employed in [13] is also tested.

The remainder of this paper is structured as follows. Section 2 details the processing blocks which compose the different diarization solutions. Section 3 describes the two fusion approaches. Sections 4 and 5 report the experimental setup, and submitted systems with results, respectively. Finally, Sections 6 and 7 provide discussions and conclusions.

2. Processing modules

This section reviews the different processing modules composing our diarization systems. These include feature extraction, speech activity detection, segmentation, segment and cluster representation, clustering and re-segmentation. As shown in Figure 1, one or more techniques are proposed for each processing module.

2.1. Feature extraction

Two different acoustic frontends were used. They include (i) a standard Mel-frequency cepstral coefficients (**MFCC**) [14] frontend and (ii) an *infinite impulse response - constant Q, Mel-frequency cepstral coefficients* (**ICMC**) [15] frontend. The latter has been applied successfully to tasks including speaker recognition, utterance verification [15] and speaker diarization [9, 16]. These features are similar to MFCC, but they replace the short-time Fourier transform by an infinite-impulse re-

[1] This work was partly supported by ANR through the ODESSA project (ANR-15-CE39-0010).

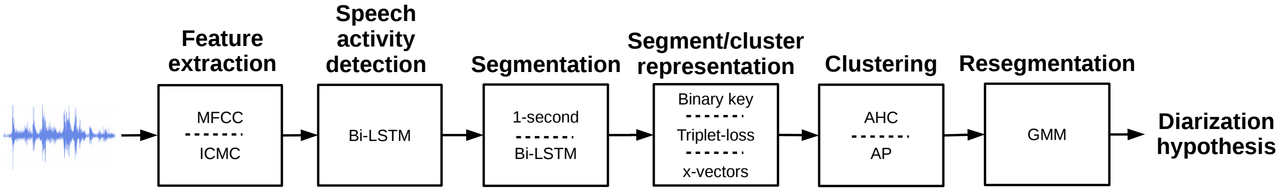

Figure 1: *Diarization pipeline adopted by the proposed individual systems.*

sponse, constant Q transform (IIR-CQT) [17]. This is a richer, multi-resolution time-frequency representation for audio signals, which provides a greater frequency resolution at lower frequencies and a higher time resolution at higher frequencies.

2.2. Speech activity detection and segmentation

All submissions share a common speech activity detection (SAD) module [18], where SAD is modelled as a supervised binary classification task (speech vs. non-speech), and addressed as a frame-wise sequence labelling task using a bi-directional long short-term memory (LSTM) network operating on MFCC features. As for segmentation, two systems were explored: (i) a straightforward uniform segmentation which splits speech content into 1 second segments and (ii) segmentation via the detection of speaker change points. The speaker change detection (SCD) module is that proposed in [19]. Similarly to the SAD module, SCD is also modelled here as a supervised binary sequence labelling task (change vs. non-change).

2.3. Segment/cluster representation

Binary key. This technique was initially proposed for speaker recognition [8, 20] and applied to speaker diarization [21, 9, 16]. It represents speech segments as low-dimensional, speaker-discriminative binary or integer vectors, which can be clustered using some sort of similarity measure. The core model to perform this mapping is a binary key background model (KBM) which is trained in the test segment before diarization. The KBM is actually a collection of diagonal-covariance Gaussian models selected from a pool of Gaussians learned on a sliding window over the test data. The window rate is adjusted dynamically to assure a minimum number of Gaussians. Then, a selection process is performed to keep a percentage p of the Gaussians in the pool to ensure sufficient coverage of all the speakers in the test audio stream. The KBM is then used to binarise an input sequence of acoustic features, which are then accumulated to obtain a cumulative vector, which is the final representation. Refer to [21] for more details.

Triplet-loss neural embedding. The embedding architecture used is the one introduced in [10] and further improved in [22]. In the embedding space, using the triplet loss paradigm, two sequences $\mathbf{x}_i$ and $\mathbf{x}_j$ of the same speaker (resp. two different speakers) are expected to be close to (resp. far from) each other according to their angular distance.

x-vector. This method [11] uses a deep neural network (DNN) which maps variable length utterances to fixed-dimensional embeddings. The network consists of three main blocks. The first is a set of layers which implements a time-delay neural network (TDNN) [23] which operates at the frame level. The second is a statistics pooling layer that collects statistics (mean and variance) at the utterance level. Finally a number of fully connected layers are followed by the output layer with as many outputs as speakers in the training data. Neurons of all layers use ReLu activations except the output layer neurons which use soft-max. The network is trained to discriminate between speakers in the training set. Once trained, the network is used to extract utterance-level embeddings for utterances from unseen speakers. The embedding is just the output of one of the fully connected layers after the statistics pooling layer.

2.4. Clustering

Agglomerative hierarchical clustering. The AHC clustering uses a bottom-up agglomerative clustering algorithm as follows. First, and assuming that the input audio stream is represented as a matrix of segment-level embeddings, a number of clusters M_{init} are initialised by a uniform splitting of the segment-level embedding matrix. Cluster embeddings are estimated as the mean segments embeddings. An iterative process including: (i) segment to-cluster assignment, (ii) closest cluster pair merging and (iii) cluster embedding re-estimation by averaging embeddings of cluster members is then applied. All comparisons are performed using the cosine similarity between embeddings. The clustering solutions generated after (i) are stored at every iteration. The output solution is selected by finding a trade-off between the number of clusters and the within-class sum of squares (WCSS) among all solutions. This is accomplished through an elbow criterion, as described in [21].

Affinity propagation. As proposed in [24], an affinity propagation (AP) algorithm [25] is our second clustering method. In contrast to other approaches, AP does not require a prior choice of the number of clusters contrary to other clustering methods. All speech segments are potential cluster centres (exemplars). Taking as input the pair-wise similarities between all pairs of speech segments, AP will select the exemplars and associate all other speech segments to an exemplar. In our case, the similarity between the i^{th} and j^{th} speech segments is the negative angular distance between their embeddings.

2.5. Re-segmentation

A resegmentation process is performed to refine time boundaries of the segments generated in the clustering step. It uses Gaussian mixture models (GMM) to model the clusters, and maximum likelihood scoring at feature level. Since the log-likelihoods at frame level are noisy, an average smoothing within a sliding window is applied to the log-likelihood curves obtained with each cluster GMM. Then, each frame is assigned to the cluster which provides the highest smoothed log-likelihood.

3. System fusion

Two approaches to fusion were explored. The first operates at the similarity matrix level suited to combine speaker diarization systems that are aligned at the segment level. The second operates at the hypothesis level and can be applied to systems with

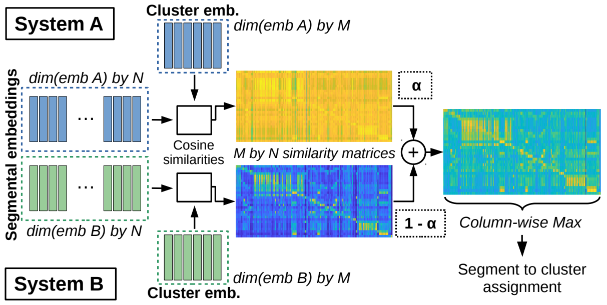

Figure 2: *Illustration of the segment-to-cluster similarity matrix fusion.*

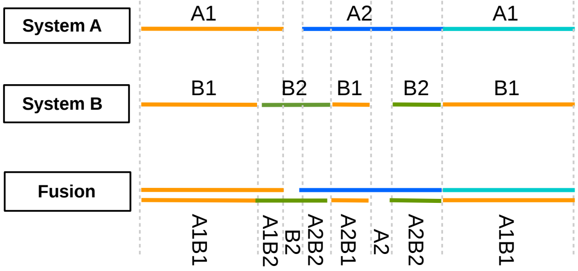

Figure 3: *Illustration of the fusion of two diarization hypotheses.*

arbitrary segment resolutions.

Fusion at similarity matrix level. Systems sharing the same segmentation can be combined at the similarity level. In [13] fusion is performed by the weighted sum of the similarity matrices of two segment-aligned systems before a linkage agglomerative clustering. This approach was adapted to our AHC algorithm by combining segment-to-cluster and cluster-to-cluster similarity matrices at every iteration. Similarities are also combined in the WCSS computation for the best clustering selection procedure. In this way, the full process takes into account the influence of the systems being fused. An example of combination of two M-cluster to N-segments similarity matrices using weights α and $1-\alpha$ is depicted in Figure 2.

Fusion at hypothesis level. The combination of systems with totally diarization pipelines is generally only possible at hypothesis level. In this work we explored hypothesis level combination using the approach described in [12]. Given a set of diarization hypotheses, every frame-level decision can be merged to assign a new frame-level cluster label which is the concatenation of all labels of the individual hypotheses. An example of this strategy is illustrated in Figure 3. This process will result in a large set of potential speaker clusters. Clusters shorter than 15 seconds are excluded and a final resegmentation is applied on the merged diarization to obtain the final diarization hypothesis.

4. Experimental setup

This section gives details of the training data and the configuration of the different modules.

4.1. Training data

For the closed-set condition, the *3/24 channel* database of around 87 hours TV broadcast programmes in Catalan language provided by the organisers was used. For the open-set condition, two popular datasets were used:

SRE-data. It includes several datasets released over the years in the context of the NIST speaker recognition evaluations (SRE), namely SRE 2004, 2005, 2006, 2008 and 2010, Switchboard, and Mixer 6. This dataset contains mostly telephone speech sampled at 8 kHz.

VoxCeleb. The VoxCeleb1 dataset [26] consists of videos containing more than 100,000 utterances for 1,251 celebrities extracted from YouTube videos. The speakers represent a wide range of different ethnicities, accents, professions and ages, and a large range of acoustic environments. This dataset is sampled at 16 kHz.

4.2. System configuration

Feature extraction. MFCCs are extracted with different numbers of coefficients depending on the subsequent segment representation: 23 static coefficients for x-vector, and 19 plus energy augmented with their first and second derivatives for triplet-loss embeddings. The binary key system uses 19 static ICMC features. Finally, the re-segmentation stage uses 19 static MFCC features.

Segment representation. For BK, the cumulative vector dimension is set to $p = 40\%$ of the size of the initial pool of Gaussian components in the KBM, leading to different representation dimensions which depend on the length of the test audio file. Gaussians are learned on a sliding window of 2 seconds to conform a pool with a minimum size set to 1024. The x-vector system uses the configuration employed in the Kaldi recipe for the SRE 2016 task[2]. Data augmentation by means of additive and convolutive noise is performed for training. The dimension of the embeddings is 512, which was later reduced to 170 using LDA. For triplet-loss embeddings, and because of the lack of global identities in the Albayzin dataset, triplets are only sampled from intra-files for the closed-set condition. Thanks to the given speaker names in Voxceleb, triplets are also sampled from inter-files for the open-set condition.

Clustering. AHC is initialised to a number N_{init} of cluster higher than the number of expected clusters in the test sessions. We set $N_{init} = 30$. The parameters of AF clustering such as preference and damping factor are tuned on the development set with the chocolate toolkit[3].

Resegmentation. It is performed with GMMs with 128 diagonal-covariance components. Likelihoods are smoothed by a sliding window of 1s.

4.3. Evaluation

Performance is assessed and optimised using the diarization error rate (DER), a standard metric for this task. It is defined as $DER = E_{spkr} + E_{FA} + E_{MS}$, where E_{FA} and E_{MS} are factors mainly associated to the SAD module, namely the false alarm and miss speech error rates. E_{MS} is also incremented by the percentage of overlapping speech present in the audio sessions that is insufficiently assigned to a single speaker. Our systems do not contain any purpose-specific module to detect such segments. Finally, E_{spkr} is the error related to speech segments that are associated to the wrong speaker identities. It is

[2] https://github.com/kaldi-asr/kaldi/tree/master/egs/sre16/v2

[3] https://chocolate.readthedocs.io/

Table 1: *Summary of ODESSA Primary (P) and contrastive (C1/C2) submissions for the closed- and open-set (denoted by c and o subscript, respectively) conditions, including feature extraction, segmentation and training data used, segment representation, clustering and fusion. Performance (DER, %) is shown in the last column.*

Condition	Sys.	Features	Segmentation	Segment rep. / train data	Clustering	Fusion	DER
Closed	P_c	-	-	-	-	C1_c,C2_c, Hyp-level	10.17
	C1_c	ICMC	1-second	BK / -	AHC	-	12.33
	C2_c	MFCC	BiLSTM	EMB / 3/24 data	AP	-	14.10
Open	P_o	-	1-second	-	AHC	C1_c,C1_o,C2_o, Sim-level[5]	7.21
	C1_o	MFCC	1-second	x-vector / SRE-data	AHC	-	9.29
	C2_o	MFCC	BiLSTM	EMB / Voxceleb	AP	-	11.46

generally the main source of error in a speaker diarization system. To compute DER, a standard forgiveness collar of 0.25s is applied around all reference segment boundaries.

Systems were tuned on approximately 15 hours of audio available at the 12 sessions that compose the partition *dev2* of the RTVE2018 database. These sessions belong to the Spanish TV shows "La noche en 24h" and "Millenium", of roughly 1h of duration and an average of 14 speakers per session. For further details please refer to [7].

5. Submitted systems and results

Table 1 summarises the ODESSA submissions and reports their performance on the RTVE2018 development set[4]. All systems share the same speech activity detection module, implying that the speech/non-speech segmentation is identical for each system. The segmentation error rate was 1.9%, composed of a missed speech rate of 0.3% and a 1.6% false alarm speech rate. ODESSA submitted systems to both closed- and open-set conditions.

Closed-set condition. C1_c uses ICMC features, 1-second uniform segmentation, BK representation and AHC clustering, while C2_c uses MFCC features, BiLSTM based SCD, triplet-loss neural embedding representation (EMB) and AP clustering. The DERs on the development set were 12.33% and 14.10%, for systems C1_c C2_c, respectively. Our primary system P_c is the fusion at diarization hypothesis level of C1_c and C2_c. Since these two systems use different segmentations this combination method was found to be the most convenient. The combined hypothesis decreases the DER to 10.17%.

Open-set condition. In the open-set condition, ODESSA aimed to analyse how systems can benefit from greater amounts of training data. The SRE and Voxceleb databases were used for this purpose. Contrastive system C1_o used MFCC features, 1-second uniform segmentation, x-vector representation trained on SRE data and AHC clustering. System C2_o aligns with the closed-set C2_c system, but where the training data was replaced with the Voxceleb data. The DERs on the development set are 9.29% and 11.46%, respectively. The primary submission P_o is the combination at similarity matrix level of three systems, namely C1_c (closed-set condition), C1_o and C2_o[5]. To speed up the tuning of optimal weights for three systems, we first tuned α for the fusion of C1_o and C1_c, and then β for the fusion of previously fused C1_o-C2_o and C2_o. Weights were set to $\alpha = \beta = 0.98$. This combined system led to the best performance on the development set, with a DER of 7.21%.

6. Discussion

Open- vs. closed-set conditions. As shown in Table 1, open-set systems outperform closed-set ones on the development set. This is somehow expected since neural approaches usually benefit from large amounts of data. Apart from differences in the amount of training data, the absence of global speaker IDs on the closed-set training data are also likely to influence performance. Our attempt to train an x-vector extractor for the closed-set condition was unfruitful, and performance of the triplet-loss embedding extractor was significantly worse than when using external training data (system C2_c vs. C2_o). In this situation, the simpler, training-free binary key approach turned out to be the single best performing system (C1_c), showing the potential of such techniques when training data is sparse or unavailable.

Segmentation. It is not clear if a dedicated module to speaker turn detection brings significant benefits compared to simple, straightforward uniform segmentation of the input stream. A volume of work reports the relative success of both explicit approaches to SCD [27, 28, 24] and uniform segmentation approaches [13, 29, 16]. In this work we found that some of the speaker representations work better with either one of the two strategies, with the uniform approach being better suited to BK and x-vector, and SCD to the triplet-loss embeddings.

System combination. In the closed-set condition, and because of the segmentation mismatch of our best performing single systems, they could not be combined at the similarity matrix level. Hence, the most obvious combination was at hypothesis level. The label combination procedure followed by a re-segmentation resulted in a lower DER than those of the two individual systems. In the open-set condition three subsystems used the same segmentation, and hence they could be fused at the similarity matrix level. This enabled the clustering to be performed jointly by considering the contributions of the three subsystems in parallel along the complete diarization process.

7. Conclusions

This paper reports the participation of the ODESSA team to the Albayzin Speaker Diarization Challenge 2018. As a consortium, our main interest was on the combination of multiple approaches to diarization. We assessed the effectiveness of two fusion strategies, namely at similarity matrix level, and at diarization hypothesis level, which allowed the combination of segment-time-aligned and arbitrary time-aligned diarization algorithms, respectively. We also found the use of appropriate and abundant training data was critical to the learning of robust embeddings, while training-free approaches are demonstrated to be adequate in the absence of suitable training data. For future work, and together with other classical challenges such as the problem of overlapping speakers, the impact of speaker change detection should be further investigated.

[4] At time of submission, neither results on the evaluation set, nor the ground-truth labels were available.

[5] In the fused system, C2_o was modified to use the 1-second segmentation and AHC clustering approaches to enable the alignment with C1_o and C1_c.

8. References

[1] A. Stolcke, G. Friedland, and D. Imseng, “Leveraging speaker diarization for meeting recognition from distant microphones,” in *Proc. ICASSP*. IEEE, 2010, pp. 4390–4393.

[2] J. Patino, R. Yin, H. Delgado, H. Bredin, A. Komaty, G. Wisniewski, C. Barras, N. Evans, and S. Marcel, “Low-latency speaker spotting with online diarization and detection,” in *Proc. Odyssey 2018*, 2018, pp. 140–146.

[3] “NIST Rich Transcription Evaluation,” 2009. [Online]. Available: https://www.nist.gov/itl/iad/mig/rich-transcription-evaluation

[4] N. Ryant, K. Church, C. Cieri, A. Cristia, J. Du, S. Ganapathy, and M. Liberman, “First DIHARD Challenge Evaluation Plan,” 2018, https://zenodo.org/record/1199638.

[5] A. Ortega, I. Viñals, A. Miguel, and E. Lleida, “The Albayzin 2016 Speaker Diarization Evaluation,” in *Proc. IberSPEECH*, 2016.

[6] A. Ortega, I. Viñals, A. Miguel, E. Lleida, V. Bazán, C. Pérez, M. Zotano, and A. de Prada, “Albayzin Evaluation: IberSPEECH-RTVE 2018 Speaker Diarization Challenge,” 2018. [Online]. Available: http://catedrartve.unizar.es/reto2018/EvalPlan-SpeakerDiarization-v1.3.pdf

[7] E. Lleida, A. Ortega, A. Miguel, V. Bazán, C. Pérez, M. Zotano, and A. de Prada, “RTVE2018 Database Description,” 2018. [Online]. Available: http://catedrartve.unizar.es/reto2018/RTVE2018DB.pdf

[8] X. Anguera and J.-F. Bonastre, “A novel speaker binary key derived from anchor models,” in *Proc. INTERSPEECH*, 2010, pp. 2118–2121.

[9] J. Patino, H. Delgado, N. Evans, and X. Anguera, “EURECOM submission to the Albayzin 2016 Speaker Diarization Evaluation,” in *Proc. IberSPEECH*, Nov 2016.

[10] H. Bredin, “TristouNet: Triplet Loss for Speaker Turn Embedding,” in *Proc. ICASSP*, New Orleans, USA, March 2017.

[11] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, “X-Vectors: Robust DNN Embeddings for Speaker Recognition,” in *Proc. ICASSP*, April 2018, pp. 5329–5333.

[12] S. Meignier, D. Moraru, C. Fredouille, J.-F. Bonastre, and L. Besacier, “Step-by-step and integrated approaches in broadcast news speaker diarization,” *Computer Speech & Language*, vol. 20, no. 2-3, pp. 303–330, 2006.

[13] G. Sell, D. Snyder, A. McCree, D. Garcia-Romero, J. Villalba, M. Maciejewski, V. Manohar, N. Dehak, D. Povey, S. Watanabe *et al.*, “Diarization is Hard: Some Experiences and Lessons Learned for the JHU Team in the Inaugural DIHARD Challenge,” *Proc. INTERSPEECH*, pp. 2808–2812, 2018.

[14] S. Davis and P. Mermelstein, “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357–366, Aug 1980.

[15] H. Delgado, M. Todisco, M. Sahidullah, A. K. Sarkar, N. Evans, T. Kinnunen, and Z. H. Tan, “Further optimisations of constant Q cepstral processing for integrated utterance and text-dependent speaker verification,” in *2016 IEEE Spoken Language Technology Workshop (SLT)*, Dec 2016, pp. 179–185.

[16] J. Patino, H. Delgado, and N. Evans, “The EURECOM Submission to the First DIHARD Challenge,” in *Proc. INTERSPEECH 2018*, 2018, pp. 2813–2817. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2018-2172

[17] P. Cancela, M. Rocamora, and E. López, “An efficient multiresolution spectral transform for music analysis,” in *Proc. ISMIR*, 2009, pp. 309–314.

[18] G. Gelly and J.-L. Gauvain, “Minimum Word Error Training of RNN-based Voice Activity Detection.” in *Proc. INTERSPEECH*, 2015, pp. 2650–2654.

[19] R. Yin, H. Bredin, and C. Barras, “Speaker Change Detection in Broadcast TV using Bidirectional Long Short-Term Memory Networks,” in *Proc. INTERSPEECH*, Stockholm, Sweden, August 2017. [Online]. Available: https://github.com/yinruiqing/change_detection

[20] G. Hernandez-Sierra, J. R. Calvo, J.-F. Bonastre, and P.-M. Bousquet, “Session compensation using binary speech representation for speaker recognition,” *Pattern Recognition Letters*, vol. 49, pp. 17 – 23, 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167865514001779

[21] H. Delgado, X. Anguera, C. Fredouille, and J. Serrano, “Fast single-and cross-show speaker diarization using binary key speaker modeling,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 23, no. 12, pp. 2286–2297, 2015.

[22] G. Gelly and J.-L. Gauvain, “Spoken Language Identification using LSTM-based Angular Proximity,” in *Proc. INTERSPEECH*, August 2017.

[23] V. Peddinti, D. Povey, and S. Khudanpur, “A time delay neural network architecture for efficient modeling of long temporal contexts,” in *Proc. INTERSPEECH*, September 2015, pp. 3214–3218.

[24] Y. Ruiqing, B. Hervé, and B. Claude, “Neural Speech Turn Segmentation and Affinity Propagation for Speaker Diarization,” in *Proc. INTERSPEECH*, Hyderabad, India, September 2018.

[25] B. J. Frey and D. Dueck, “Clustering by Passing Messages Between Data Points,” *science*, vol. 315, no. 5814, pp. 972–976, 2007.

[26] A. Nagrani, J. S. Chung, and A. Zisserman, “VoxCeleb: a large-scale speaker identification dataset,” in *Proc. INTERSPEECH*, August 2017.

[27] Z. Zajíc, M. Kunešová, J. Zelinka, and M. Hrúz, “ZCU-NTIS Speaker Diarization System for the DIHARD 2018 Challenge,” in *Proc. INTERSPEECH*, 2018, pp. 2788–2792.

[28] I. Viñals, P. Gimeno, A. Ortega, A. Miguel, and E. Lleida, “Estimation of the Number of Speakers with Variational Bayesian PLDA in the DIHARD Diarization Challenge,” in *Proc. INTERSPEECH*, 2018, pp. 2803–2807.

[29] M. Diez, F. Landini, L. Burget, J. Rohdin, A. Silnova, K. Žmolíková, O. Novotný, K. Veselý, O. Glembek, O. Plchot, L. Mošner, and P. Matějka, “BUT System for DIHARD Speech Diarization Challenge 2018,” in *Proc. INTERSPEECH*, 2018, pp. 2798–2802.