SPEAKER EMBEDDINGS FOR DIARIZATION OF BROADCAST DATA IN THE ALLIES CHALLENGE

Anthony Larcher, Ambuj Mehrish, Marie Tahon, Sylvain Meignier, Jean Carrive, David Doukhan, Olivier Galibert, Nicholas Evans

SPEAKER EMBEDDINGS FOR DIARIZATION OF BROADCAST DATA IN THE ALLIES CHALLENGE

Anthony Larcher[1]*, Ambuj Mehrish*[1]*, Marie Tahon*[1]*, Sylvain Meignier*[1]*,*
Jean Carrive[2]*, David Doukhan*[2]*, Olivier Galibert*[3]*, Nicholas Evans*[4]

1.LIUM - EA4023, Le Mans Université
Avenue Olivier Messiaen, 72085 LE MANS CEDEX 9, France
firstname.lastname@univ-lemans.fr
2. French National Institute of Audiovisual, Paris, France {*jcarrive, ddoukhan*}*@ina.fr*
3. Laboratoire National d'Essais (LNE) 4. EURECOM, Sophia Antipolis, France

ABSTRACT

Diarization consists in the segmentation of speech signals and the clustering of homogeneous speaker segments. State-of-the-art systems typically operate upon speaker embeddings, such as i-vectors or neural x-vectors, extracted from mel cepstral coefficients (MFCCs) or spectrograms. The recent SincNet architecture extracts x-vectors directly from raw speech signals. The work reported in this paper compares the performance of different embeddings extracted from MFCCs or the raw signal for speaker diarization and broadcast media treated with compression and sub-sampling, operations which typically degrade performance. Experiments are performed with the new ALLIES database that was designed to complement existing, publicly available French corpora of broadcast radio and TV shows. Results show that, in adverse conditions, with compression and sampling mismatch, SincNet x-vectors outperform i-vectors and x-vectors by relative DERs of 43% and 73% respectively. Additionally we found that SincNet x-vectors are not the absolute best embeddings but are more robust to data mismatch than others.

Index Terms— Speaker diarization, x-vectors, i-vectors, SincNet, raw signal

1. INTRODUCTION

Most audiovisual broadcast data are archived, either for legal reasons, promotion or cultural preservation. In order to be exploitable, archived data is typically enriched with summaries, context or participant names. Due to the high cost, not all documents can be annotated manually, implying the need for automated annotation. Automatic speaker diarization goes some way to meeting this need by answering the question of "who speaks when?" in an audio track. Speaker diarization is an enabling technology, often being applied before subsequent speaker identification or speech recognition.

Speaker diarization solutions typically comprise three steps: 1) segmentation of audio recordings into speaker-homogeneous chunks; 2) characterization of each chunk with a compact representation; 3) clustering of same-speaker chunks. The question of how best to capture relevant speaker discriminative information is hence crucial (step 2). Speaker characteristics lie predominantly in the spectral magnitude domain and traditional systems typically extract speaker representations either from mel frequency cepstral coefficients (MFCCs) or magnitude spectrograms. These are used to extract speaker embeddings such as i-vectors [1] or x-vectors [2]. However, such features do not capture *all* relevant speaker information contained within the time domain signal. This is one reason why speaker diarization might benefit from the use of end-to-end neural approaches such as SincNet pre-processing that takes raw signal as input [3] for the extraction of speaker representations.

Speaker diarization was devised to support speaker recognition for conversational telephone speech and applied later to different domains such as meetings [4], broadcast media [5, 6, 7] and internet videos [8, 9]. While diarization performance can be reasonable for tasks where the number of speakers is limited or known (telephone speech), when the conversation is well structured in turn-taking, or when acoustic variability is modest, it degrades rapidly when the number of speakers is unpredictable, where there is substantial speaker overlap, or where acoustic variability is more significant. The latter typifies broadcast media. These broadcast media can mix several compression codecs. Remote speakers may use proprietary voice call or video chat softwares using heterogeneous lossy compression formats while media obtained from streaming platforms, using their own compression codecs may also be broadcast on TV or on Radio. Lastly, the use of additional compression and sub-sampling, applied to reduce storage and streaming overheads, compounds the difficulty. For these reasons, speaker diarization for broadcast media remains a significant challenge, as shown by the results from the recent DIHARD challenges [10].

Our first contribution consists in a new broadcast data collection that has been designed to complement existing publicly available corpora. This database will be made publicly available at the end of the ALLIES challenge [1] and is that used for all experimental work reported in this paper. The second contribution consists in a comparison of diarization performance using three different speaker embeddings, extracted from MFCCs (classical i-vectors or neural x-vectors) or from the raw signal (SincNet x-vectors). To the best of our knowledge, this is the first application of SincNet x-vectors to speaker diarization. The third contribution relates to an assessment of compression and sub-sampling mismatch which occurs in data archives and which degrades diarization performance. The goal of this work is then to determine which embedding offers the best robustness to compression and sub-sampling related mismatch.

This work has been funded by the CHIST-ERA project ALLIES (ARN-17-CHR2-0004-01), the French ANR Extensor (ANR-19-CE23-0001-01) and Antract (ANR-17-CE38-0010) projects.

[1] https://lium.univ-lemans.fr/en/allies-evaluation/

2. RELATED WORK

Many different corpora support research in speaker diarization for broadcast data. Albayzin [11], a series of evaluations organized by Zaragoza University, collected and released databases of Spanish broadcast data in different local languages. The MGB challenge [7] used data collected by the BBC. It includes data from a variety of TV shows with a large number of different speakers and acoustic conditions that make for an especially challenging dataset. Several related challenges organised in France used recordings of French radio and TV shows to support evaluations involving multiple tasks including speaker diarization, e.g. (REPERE [5], ESTER [12] and ETAPE [6]).

The use of compact speaker representations for diarization began to dominate the field [13] since the introduction of i-vectors for speaker recognition [14] in 2009. Embeddings computed with deep neural networks, known as x-vectors, have since been shown to outperform i-vectors [2, 15]. The extraction of both i-vectors and x-vectors begins with the extraction of spectro-temporal features such as MFCCs, filter-bank coefficients or other variants. An alternative, gaining traction in recent years, consists in the feeding of neural networks with raw waveforms. These techniques were developed in order to support optimization in an end-to-end manner and hence to design task-optimised front-ends without relying on empirically designed acoustic features [16, 17]. These approaches allows neural networks to determine automatically what information in the raw waveform is most beneficial, information that might sometimes lost in the extraction of hand-crafted acoustic features. Domain mismatch is a widely explored topic in speech processing [18], e.g. the degrading impact of data compression upon speaker verification using i-vectors reported in [19]. To the best of our knowledge, the only study of compression impacts upon the performance of neural speaker embeddings is reported in [20]. That study considers mainly the use of audio compression for data augmentation and speaker recognition rather than speaker diarization. [21] have shown that standard transform-based audio codecs (AAC and MP3) have a poor frequency resolution for timber-related tasks. These lossy transformations of the sound signal were shown to have impact on the resulting MFCC and chroma features, as well as on Music Information Retrieval tasks [22].

3. ALLIES DATA

The ALLIES data collection aims at extending corpora collected for a series of challenges (ESTER, REPERE, ETAPE) based on radio and TV shows from French channels that have been collected in high quality (16kHz, 16bits) during a precise period. One motivation of the ALLIES collection is to tackle the issue of diarization over a collection of shows across time and requires several time stamped series of shows with fine sampling across a precise time period. Since 1995, INA is in charge of the Legal Deposit for national TV and radio channels. At the time, the INA was faced with significant technical constraints regarding digital archiving, and the choice was made to calibrate video and audio compression to fit a full day's of programs on a 4.7Gb DVD. These constraints have since been lifted and INA now records 179 TV and radio channels 24/7 in a much better quality.

In this paper we only describe the ALLIES data and its acoustic specificity without addressing the question of diarization across time. However, the completion of previously collected dataset required to access archives of the French National Audiovisual Institute (INA) [2] to retrieve missing shows over the period covered by existing databases. Facing the challenged of processing both high quality data and archived data that suffered quality degradation along the archiving process.

Table 1. Statistics of the three partitions of the ALLIES corpus (all timestamps are in hh:mm:ss format).

	TRAIN	DEV		EVAL
		LCP + BFM	LCP	
Total duration	223:37:17	105:51:06	47:04:23	275:24:46
Annotated duration	175:32:48	33:54:46	18:37:02	98:05:51
Total # of speakers	6037	1790	725	3759
Number of shows	475	200	85	324
min # of speaker per show	1	2	3	3
max # of speaker per show	74	38	33	43
average # of speaker per show	12	9	8	12
min duration per show	00:04:31	00:04:05	00:19:59	00:40:02
max duration per show	01:18:07	1:15:58	1:15:58	01:29:59
average duration per show	00:28:14	00:31:45	00:33:13	00:51:00

The ALLIES corpus comprises the usual three partitions: training (TRAIN), development (DEV) and evaluation (EVAL). The TRAIN and DEV partitions consist of data used previously for the REPERE, ESTER and ETAPE challenges. Collected between 1998 and 2013, this data is of relatively high-quality (16kHz 16bit).

Having been extracted from INA's archive, the third partition: ALLIES EVAL has been compressed using an AAC codec[3] and subsampled to 11,025Hz. There is hence substantial mismatch between the EVAL data on one side and the TRAIN and DEV data on the other side. In terms of content, the DEV set includes TV shows from two TV channels (BFM and LCP) while the EVAL set contains only additional episodes from LCP shows, thereby creating an additional mismatch in terms of content. To evaluate the impact of content mismatch on the different embeddings this paper reports experiments conducted with two different versions of the DEV set: the complete version and a balanced version containing only LCP shows. Statistics for the three partitions of the ALLIES corpus (including the two versions of the DEV set) are illustrated in Table 1.

4. SYSTEMS DESCRIPTION

In this work, a single system, developed using Sidekit/S4D [23, 24], is used to compare speaker embeddings and analyse the effect of compression and sub-sampling on broadcast data. This system is made publicly available for the ALLIES challenge [4]. All results are evaluated using the standard diarization error rate (DER).

4.1. BIC segmentation and clustering

In order to focus on the performance of speaker embeddings, the initial voice activity detection is taken from the reference labels. Speaker segmentation is then applied, followed by an initial, weak hierarchical agglomerative clustering (HAC) to group same-speaker segments. Both use the Bayesian Information Criterion (BIC) as a dissimilarity measure and stopping criterion and 12 MFCC features with energy coefficients. More details are given in [24].

[2] https://institut.ina.fr/en

[3] ffmeg with the following parameters `-acodec libfdk_aac -b:a 64k`

[4] https://git-lium.univ-lemans.fr/Larcher/allies-evalution

4.2. Speaker clustering

While the initial clustering produces speaker-homogeneous clusters, speakers are not necessarily represented by a single cluster. The initial clustering is hence refined through a second speaker clustering step. It operates on segments, each modeled by an embedding, while the set of embeddings for each cluster are averaged to produce a more robust representation. A PLDA model is trained for each type of embedding using the exact same configuration with a full rank speaker factor matrix and a pseudo distance between clusters is obtained from the PLDA scores. This distance is the measure employed to select the clusters to be grouped as well as to stop the clustering process. Cluster embeddings are grouped with hierarchical agglomerative clustering (HAC) using a complete linkage criteria.

4.3. Speaker embeddings

Three speaker embeddings were investigated. All models used to extract embeddings have been trained using all sessions from the 659 speakers that appear more than 15 times in the training corpus. All embeddings have 100 dimensions.

I-vectors For i-vector extraction, acoustic parameters are normalized (centered/reduced over a sliding window of up to 3 seconds) and the 12 MFCC and energy features are augmented by first- and second-order derivatives. The UBM-GMM is composed of 256 Gaussian distributions and the i-vector dimension is set to 100.

X-vectors X-vectors are extracted using a standard neural network architecture [2]. This network consists of five 1D-convolutional layers, followed by a temporal pooling layer that computes mean and standard deviation over the whole sequence. The x-vectors are extracted after a linear layer while two additional layers are used for network training as depicted in Table 2. The network is fed with vectors of 30 MFCCs and a mean variance normalization is applied over the whole speech segments. Batches of 64 segments of 2 seconds are used for training and data are augmented 9 times with additive noise taken from the MUSAN database [25].

SincNet x-vectors The last type of embeddings is computed using the same architecture as the previous x-vectors, except that MFCCs are replaced by a SincNet layer that directly processes the raw waveform so that the entire processing chain is optimized for the speaker characterization task during training. The neural network is trained using the same configuration, input data, and augmentation parameters that are used to train the x-vector extractor. The architecture of the entire network is illustrated in Table 2.

5. EXPERIMENTS AND RESULTS

5.1. Impact of the archiving process

TRAIN and DEV data share similar bandwidth and quality while EVAL data has been sub-sampled and compressed when archived. One obvious solution to mitigate the effects of this mismatch would be to train a system using compressed and sub-sampled data. This would limit its application to the processing of similarly treated data and every change to the archiving pipeline would require system retraining. Our motivation is thus twofold: i) evaluate the impact of the archiving process to guide archivers such as INA to optimize this process and ii) find robust embeddings that would protect robustness in the case of future pipeline changes.

Table 2. Architecture of the x-vector extractors. Dropout is used for all layers except the Linear layers. The activation function for Convolutional and Fully Connected layers is LeakyReLu.

x-vector	SincNet X-vector
MFCC	SincNet [80, 251, 1]
	1D-Conv [60, 5, 1]
	1D-Conv [60, 5, 1]
1D-Conv [512, 5, 1]	
1D-Conv [512, 3, 2]	
1D-Conv [512, 3, 3]	
1D-Conv [512, 1, 1]	
1D-Conv [1536, 1, 1]	
StatPooling	
Linear [3072, 100]	
Fully Connected [100, 512]	
Fully Connected [512, 512]	
Linear [512, 659]	
SoftMax	

Table 3. Performance of different embeddings on the ALLIES datasets for four mismatch conditions in terms of DER. First and fourth column indicate the treatment applied to TRAIN+DEV and EVAL respectively (*comp* for compression, *sub* for sub-sampling).

Embedding	Dev		Eval	
	Process	DER	DER	Process
i-vectors	*None*	7.10	28.55	
	comp	7.36	21.04	*comp + sub*
	sub	7.36	17.17	
	comp + sub	8.85	17.71	
x-vectors	*None*	7.13	43.41	
	comp	7.41	37.88	*comp + sub*
	sub	7.16	32.28	
	comp + sub	8.02	20.27	
SincNet	*None*	7.91	16.35	
	comp	8.22	15.98	*comp + sub*
	sub	8.00	15.66	
	comp + sub	9.36	15.40	

In order to evaluate the robustness of speaker embeddings, we produce three additional versions of TRAIN and DEV by applying compression, sub-sampling or both, to reduce or suppress the mismatch with EVAL data and determine which processing impacts most each of the different embeddings. Table 3 presents the DER obtained for different embeddings and for different processing mismatch between TRAIN+DEV and EVAL data. Note that in all experiments, TRAIN and DEV are processed the same way.

Focusing on DEV results, where data exactly matches TRAIN data, we see how much sub-sampling or/and compression hurts the performance for all types of embeddings. We also observe that the degradation resulting from compression and sub-sampling are similar in terms of DER for i-vectors while x-vectors SincNet x-vectors suffer more from compression than from sub-sampling.

Moving to EVAL results, we observe immediately the substantial impact of compression and sub-sampling mismatch. Compared to results for DEV data, DERs increases by 300% relative for i-vectors, 500% for x-vectors but only 106% for SincNet x-vectors. Of course, reducing the mismatch to only sub-sampling (all data are compressed but only EVAL is sub-sampled) improves the DER by a

Table 4. Performance of different systems on the ALLIES datasets for four mismatch conditions in terms of DER. DEV set is reduced to LCP data. First and fourth column indicate the treatment applied to TRAIN+DEV and EVAL respectively (*comp* for compression, *sub* for sub-sampling.

Embedding	DEV		EVAL	
	Process	DER	DER	Process
i-vectors	*None*	4.98	27.67	*comp + sub*
	comp	5.12	23.02	
	sub	4.78	17.17	
	comp + sub	6.51	22.59	
x-vectors	*None*	5.22	39.12	*comp + sub*
	comp	5.29	34.09	
	sub	5.37	32.28	
	comp + sub	6.30	19.10	
SincNet	*None*	7.91	16.35	*comp + sub*
	comp	6.23	15.98	
	sub	5.69	15.73	
	comp + sub	7.23	15.40	

relative 27% for i-vectors, but by only 13% and 3% for the classical and SincNet x-vectors respectively. When the mismatch is reduced to compression alone (all data sets are sub-sampled but only EVAL data is compressed), the relative improvement compared to a complete mismatch is 40% for i-vectors, 26% for x-vectors and only 4% for SincNet x-vectors. This can be explained by the fact that signal representation in AAC has a low frequency resolution and that MFCCs are not able to capture as much information as SincNet. Eventually, removing the mismatch by compressing and sub-sampling all data sets provides the best result for all embeddings. Suppressing the compression and sub-sampling mismatch improves the DER on the EVAL data by a relative 38% for i-vectors, 54% for x-vectors and 6% for SincNet x-vectors. These series of experiments show that sub-sampling mismatch hurts more than AAC compression for all embeddings and that x-vectors are much more sensitive to this mismatch than i-vectors. In all cases, SincNet x-vectors are more robust to archiving mismatch.

5.2. Impact of the TV content in the development set

The poor performance on the EVAL set can be due to the TV shows mismatch between DEV and EVAL that would impact the optimal clustering threshold set on the DEV set. In order to observe any impacts of content mismatch, we hence conducted a second set of experiments using the reduced DEV set containing only LCP TV shows (see Table 1). Results are illustrated in Table 4.

Trends are similar to the previous experiments regarding the effect of compression and sub-sampling. When the threshold is optimised for the reduced DEV set and LPC TV shows only, we observe better results for EVAL data, except for the i-vectors system for which performance degrades when TRAIN+DEV data is compressed. While a general improvement is expected some interesting observations can nonetheless be made from results plotted in Figure 1. It shows DER for EVAL data for the three types of embeddings as a function of the clustering threshold that is set using the complete DEV set in high quality (with no compression nor sub-sampling). Although the optimal performance on EVAL data is quite similar for the three types of embeddings (14.39% for i-vectors, 19.38% for x-vectors and 15.64% for SincNet x-vectors), the threshold obtained from optimisation on DEV data (highlighted by crosses on each curve) is sub-optimal for i-vectors and x-vectors. This is due to substantial differences in the PLDA score distributions for DEV and EVAL data. By optimising the clustering threshold using the reduced DEV data (denoted by solid circles on each curve) the DER improves for i-vector and x-vector embeddings but doesn't change for SincNet x-vectors. Hence, while SincNet x-vectors do not give the best DER, the flatter curve in Figure 1 shows that they are the most robust to data mismatch. The fact that the optimal threshold for SincNet x-vectors doesn't vary when modifying the DEV set is also encouraging as it means that the optimal threshold for DEV data is less sensitive to TV show mismatch.

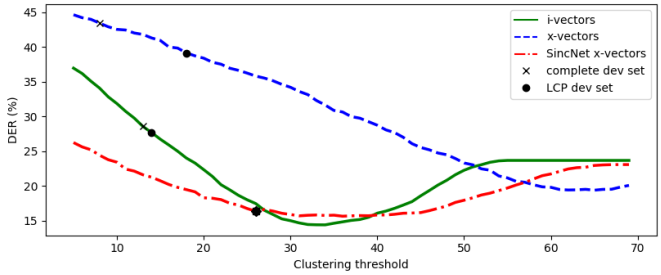

Fig. 1. Diarization Error Rate on the EVAL data for three types of embeddings. On each curve, the cross and circle indicate respectively the DER for the optimal threshold set with the complete development set or the development set reduced to LCP data.

6. CONCLUSION AND DISCUSSIONS

In this work we've extended existing corpora for speaker diarization on broadcast data. The resulting corpus includes 1,010 TV and radio shows. The new 324 shows have been precisely annotated for overlap speech. This corpus will be used in the future to evaluate lifelong learning speaker diarization systems in the ALLIES challenge that will take place in the coming year. After this challenge, the corpus and associated protocols will be publicly released for research purpose. While collecting this corpus from INA's archives, which are dedicated to storage and streaming of the TV and radio data, we observed that this process can induce severe compression and sub-sampling mismatches that affect state-of-the-art speaker embeddings relying on cepstral features. We proposed a combination of standard x-vectors with a SincNet pre-processing that takes raw signal as input and has shown to be robust to compression, sub-sampling and even TV content mismatches where i-vectors and especially x-vectors suffer from this mismatch. In adverse conditions, when compression and sampling mismatch affects the EVAL data, SincNet x-vectors out-perform i-vectors and x-vectors by a relative 43% and 73% DER respectively. Additionally, we found that SincNet x-vectors are more robust to TV show mismatch. In the future, this work will be extended in two directions. First the ALLIES corpus will be used to evaluate lifelong learning speaker diarization systems in the ALLIES challenge before being publicly released for research purpose together with associated metrics and protocols. The systems described in this work will be released to participants as baseline systems for the challenge. Note the code is already available in the Sidekit and S4D platforms. Second we will analyze the bias introduced by the compression and sub-sampling mismatch in the x-vector and SincNet x-vectors networks in order to understand why the former one is more robust.

7. REFERENCES

[1] Gregory Sell and Daniel Garcia-Romero, "Speaker diarization with plda i-vector scoring and unsupervised calibration," in *2014 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2014, pp. 413–417.

[2] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur, "X-VECTORS: ROBUST DNN EMBEDDINGS FOR SPEAKER RECOGNITION," in *ICASSP*, 2018.

[3] Mirco Ravanelli and Yoshua Bengio, "Speaker recognition from raw waveform with sincnet," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 1021–1028.

[4] Jonathan G Fiscus, Jerome Ajot, Martial Michel, and John S Garofolo, "The rich transcription 2006 spring meeting recognition evaluation," in *International Workshop on Machine Learning for Multimodal Interaction*. Springer, 2006, pp. 309–322.

[5] Aude Giraudel, Matthieu Carré, Valérie Mapelli, Juliette Kahn, Olivier Galibert, and Ludovic Quintard, "The REPERE Corpus : a Multimodal Corpus for Person Recognition," in *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, 2012.

[6] Guillaume Gravier, Gilles Adda, Niklas Paulsson, Matthieu Carré, Aude Giraudel, and Olivier Galibert, "The ETAPE Corpus for the Evaluation of Speech-based TV Content Processing in the French Language," in *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, 2012.

[7] Peter Bell, Mark JF Gales, Thomas Hain, Jonathan Kilgour, Pierre Lanchantin, Xunying Liu, Andrew McParland, Steve Renals, Oscar Saz, Mirjam Wester, et al., "The mgb challenge: Evaluating multi-genre broadcast media recognition," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, 2015, pp. 687–693.

[8] Joon Son Chung, Jaesung Huh, Arsha Nagrani, Triantafyllos Afouras, and Andrew Zisserman, "Spot the conversation: speaker diarisation in the wild," *arXiv preprint arXiv:2007.01216*, 2020.

[9] Lei Sun, Jun Du, Chao Jiang, Xueyang Zhang, Shan He, Bing Yin, and Chin-Hui Lee, "Speaker diarization with enhancing speech for the first dihard challenge.," in *Interspeech*, 2018, pp. 2793–2797.

[10] Neville Ryant, Kenneth Church, Christopher Cieri, Alejandrina Cristia, Jun Du, Sriram Ganapathy, and Mark Liberman, "The Second DIHARD Diarization Challenge: Dataset, Task, and Baselines," in *Proc. Interspeech 2019*, 2019, pp. 978–982.

[11] Diego Castán, David Tavarez, Paula Lopez-Otero, Javier Franco-Pedroso, Héctor Delgado, Eva Navas, Laura Docio-Fernández, Daniel Ramos, Javier Serrano, Alfonso Ortega, et al., "Albayzín-2014 evaluation: audio segmentation and classification in broadcast news domains," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2015, no. 1, pp. 1–9, 2015.

[12] Sylvain Galliano, Guillaume Gravier, and Laura Chaubard, "The ester 2 evaluation campaign for the rich transcription of french radio broadcasts," in *Tenth Annual Conference of the International Speech Communication Association*, 2009.

[13] J. Villalba, A. Ortega, A. Miguel, and E. Lleida, "Variational bayesian plda for speaker diarization in the mgb challenge," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2015, pp. 667–674.

[14] Najim Dehak, Patrick J Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2010.

[15] Hervé Bredin, Ruiqing Yin, Juan Manuel Coria, Gregory Gelly, Pavel Korshunov, Marvin Lavechin, Diego Fustes, Hadrien Titeux, Wassim Bouaziz, and Marie-Philippe Gill, "pyannote.audio: neural building blocks for speaker diarization," in *ICASSP 2020, IEEE International Conference on Acoustics, Speech, and Signal Processing*, Barcelona, Spain, May 2020.

[16] M. Ravanelli and Y. Bengio, "Speaker recognition from raw waveform with sincnet," in *2018 IEEE Spoken Language Technology Workshop (SLT)*, 2018, pp. 1021–1028.

[17] Jee-weon Jung, Seung-bin Kim, Hye-jin Shim, Ju-ho Kim, and Ha-Jin Yu, "Improved rawnet with feature map scaling for text-independent speaker verification using raw waveforms," *Proc. Interspeech 2020(to appear)*, pp. 3583–3587, 2020.

[18] Qing Wang, Wei Rao, Sining Sun, Leib Xie, Eng Siong Chng, and Haizhou Li, "Unsupervised domain adaptation via domain adversarial training for speaker recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4889–4893.

[19] Mitchell McLaren, Victor Abrash, Martin Graciarena, Yun Lei, and Jan Pesán, "Improving robustness to compressed speech in speaker recognition.," in *INTERSPEECH*, 2013, pp. 3698–3702.

[20] ML McLaren, Diego Castan, Mahesh Kumar Nandwana, Luciana Ferrer, and Emre Yilmaz, "How to train your speaker embeddings extractor," in *Odyssey*. 2018, Les Sables d'Olonne, France: ISCA.

[21] E. Ravelli, G. Richard, and L. Daudet, "Audio signal representations for indexing in the transform domain," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 3, pp. 434–446, 2010.

[22] Julián Urbano, Dmitry Bogdanov, Herrera Boyer, Emilia Gómez Gutiérrez, and Xavier Serra, "What is the effect of audio quality on the robustness of mfccs and chroma features?," in *Proc. of the 15th Conference of the International Society for Music Information Retrieval (ISMIR)*, 2014, pp. 573–578.

[23] Anthony Larcher, Kong Aik Lee, and Sylvain Meignier, "An Extensible Speaker Identification SIDEKIT in Python," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2016)*, Shanghai, China, 2016, Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on, pp. 5095–5099.

[24] Pierre-Alexandre Broux, Florent Desnous, Anthony Larcher, Simon Petitrenaud, Jean Carrive, and Sylvain Meignier, "S4D: Speaker Diarization Toolkit in Python," in *Interspeech*, Hyderabad, India, Sept. 2018.

[25] David Snyder, Guoguo Chen, and Daniel Povey, "Musan: A music, speech, and noise corpus," *arXiv preprint arXiv:1510.08484*, 2015.