IberSPEECH 2021
24-25 March 2021, Valladolid, Spain

Diarization and Identity Assignment Compatibility in the Albayzín 2020 Challenge

Ignacio Viñals, Pablo Gimeno, Alfonso Ortega, Antonio Miguel and Eduardo Lleida

University of Zaragoza, Spain
{ivinalsb, pablogj, ortega, amiguel, lleida}@unizar.es

Abstract

The current need to identify the speakers in a certain recording has evolved along time, requesting more and more information. While speaker recognition originally focused on determining whether a speaker talks in a certain audio with a single speaker, later diarization focused on differentiating speakers along the recording. The latest step is Identity Assignment (IA), which combines both of them, i.e., deciding whether a certain speaker is present in a given audio, as well as determining the periods of time when the speaker is active.

Our work presents and analyzes the ViVoLAB results for the Albayzín 2020 evaluation, focused on diarization and identity assignment. These challenges will be faced in the broadcast domain, with data coming from national Spanish TV Corporation RTVE. For this purpose we have developed a Bottom-Up diarization architecture based on the embedding-PLDA paradigm. On top of the diarization solution we have added an identity assignment block, based on the speaker verification approach.

Index Terms: diarization, identity assignment, neural networks.

1. Introduction

The recent increase of raw multimedia data has made capital the development of automatic techniques capable of labeling any input audio. Among these techniques diarization is the task dedicated to identify the time stamps defining when any speaker talks in a given audio. Diarization goal is the differentiation of the speakers rather than their real identification. For this purpose we can make use of generic labels.

Diarization has evolved simultaneously to other speech technologies along time, specially inheriting approaches coming from speaker recognition: From basic technologies [1] to Joint Factor Analysis (JFA) [2, 3], i-vectors [4] or Deep Neural Networks (DNNs) [5]. Furthermore, this development also allows us to obtain reasonable performances out of telephone domain, as in broadcast [6] or meetings domains [7].

Nevertheless, these labels become worthier as long as they become more and more informative. Thus, the generic labels obtained during diarization could be improved if the true identity of the speaker was inferred too. This task, also known as Identity Assignment (IA), can be considered a complementary task working on top of diarization [3]. Although diarization was a complementary task for IA during its early days, now it has gained more and more relevance as long as we need higher and higher quality information from an audio. The joint work of both blocks helps us determining whether a certain speaker is present in a given audio as well as determining when he is contributing.

Albayzín 2020 is the most recent edition in the ongoing series of Albayzín technological evaluations, seeking the improvement of speech technologies in Iberian languages (languages spoken in the Iberian peninsula), paying special attention on the broadcast domain. 2020 edition continues the line from previous evaluations, working with audios provided by Radio Televisión Española (RTVE), the national Spanish TV corporation. However, in addition to diarization, 2020 edition now includes the Identity Assignment problem as a complementary goal.

In this work we present the ViVoLAB submission to Albayzín 2020 evaluation, specifically to the diarization and identity assignment task. Our system considers the dual problem as a cascade of tasks. First we perform diarization considering a Bottom-Up approach where the original audio, once processed by the front-end, is segmented and then clustered to obtain the final speaker labels. Regarding the IA task, we manage it as a speaker verification problem where diarization clusters play the role of test audios to evaluate against the given enrollment recordings.

The rest of the document is structured as follows: In Section 2 we describe the ViVoLAB diarization system. The identity assignment block is explained in Section 3. Our experimental results are included in Section 4. Finally, our conclusions are expressed in Section 5.

2. ViVoLAB diarization system

The diarization system works according to the Bottom-Up philosophy, i.e., first identifying segments with a single speaker on them, later combined according to a clustering block. This clustering block makes use of the embedding-PLDA (Probabilistic Linear Discriminant Analysis) paradigm.

2.1. Voice Activity Detection

Our approach for voice activity detection (VAD) is based on a deep learning solution. We use a convolutional recurrent neural network (CRNN) consisting of 3 2D convolutional blocks (2D conv. layer with 64 filters of size 3x3, batch normalization and ReLU activation) followed by 3 Bidirectional Long Short Time Memory (BiLSTM) layers. Then, the final speech score is obtained through a linear layer. The neural network works in terms of streams of feature vectors, 300 to be specific, inferring a VAD label per feature in the input sequence. As input features, 64 Mel filter banks and the frame energy are extracted from the raw audio and fed to the neural network. These input features are normalized in mean and variance prior to any other calculation within the network.

Adam is chosen as the optimizer for the neural network, using a learning rate that decays exponentially from 10^{-3} to 10^{-4} in the 30 epochs that data is presented to the neural network. Cross entropy is the training objective, as usually done in classification tasks.

10.21437/IberSPEECH.2021-20

The CRNN is trained on a combination of different broadcast datasets. Specifically, we include data from the Albayzín 2010 dataset [8] (train and eval), Albayzín 2018 dataset [9] (dev2 and eval) and a selection of data from 2015 Multi-Genre Broadcast (MGB) Challenge [10] (train, dev. longitudinal and task3 eval). A 10% of all the data is reserved for training validation. Furthermore, audios are augmented with a variety of noises that can be usually found in broadcast emissions (sitcom noises, crowd and laughter noises, babble, street music and stadium noises). These noises are added in training time with a Signal to Noise ratio (SNR) that is sampled from an uniform distribution in the range (5, 25) dB.

2.2. Speaker Change Point Detection

The Speaker Change Point Detection block works in terms of Bayesian Information Criterion (BIC), according to its differential form (ΔBIC) [11]. We consider analysis windows of 6 seconds, modelling speakers with full- covariance Gaussian distributions. This block prioritizes those speech/non-speech boundaries given by VAD. As input features the system considers 20 MFCC [12] features vectors, over a 25 ms hamming window every 10 ms. Features are then normalized according to Cepstral Mean and Variance Normalization [13] to mitigate channel effects.

2.3. Embedding Extraction

Each one of the obtained segments will be transformed into a compact representation also known as embedding. For this purpose we have opted for an evolution of the extended x-vector [14] architecture, based on Time Delay Neural Networks (TDNNs). Compared to the original architecture, we have substituted the mean and standard deviation pooling block by a multi-head self-attention block [15]. This self-attention block simultaneously considers H different patterns to learn from the data themselves, also known as heads. For each pattern j this block weighs the stream of forwarded information $\mathbf{X} = \{\mathbf{x}_1, .., \mathbf{x}_i, ..\mathbf{x}_N\}$ by a set of weights $\boldsymbol{\alpha}_{ij}$ estimated as follows:

$$\boldsymbol{\alpha}_{ij} = \frac{\exp\left(\mathbf{W}_j \mathbf{x}_i + \boldsymbol{b_j}\right)}{\sum_i \exp\left(\mathbf{W}_j \mathbf{x}_i + \boldsymbol{b_j}\right)} \tag{1}$$

where $\mathbf{W}_j$ refers to the row j of the weight matrix $\mathbf{W}$, and $\boldsymbol{b}_j$ stands for the jth component of the bias vector $\boldsymbol{b}$. Both weight matrix and bias are the trainable parameters of an affine transformation of H neurons.

Given the obtained weights $\boldsymbol{\alpha}_{ij}$, the block estimates for each head j its weighted mean ($\boldsymbol{\mu}_j$) and weighted standard deviation ($\boldsymbol{\sigma}_j$). The output of the block consists of the concatenation of the estimated mean and standard deviation from each head. Experimentally we set the value of H up to 8. The final configuration for the network is shown in Table 1.

The neural network has been trained with data from VoxCeleb 1 [16] and 2 [17] The resulting neural network provides embeddings of dimension 512. These embeddings will be later centered, dimensionsionally reduced by means of LDA up to 200 as whitening and length-normalized [18].

2.4. Clustering

The obtained embeddings are modeled in a generative manner according to [19], where a tree-based PLDA clustering is proposed. This solution proposes a Maximum A Posteriori (MAP) estimation of the speaker labels Θ given the set of embeddings

#	Component Type	Context	Size
1	TDNN-ReLU-BN	[t-2:t+2]	512
2	FFN-ReLU-BN	t	512
3	TDNN-ReLU-BN	[t-2,t,t+2]	512
4	FFN-ReLU-BN	t	512
5	TDNN-ReLU-BN	[t-3,t,t+3]	512
6	FFN-ReLU-BN	t	512
7	TDNN-ReLU-BN	[t-4,t,t+4]	512
8	FFN-ReLU-BN	t	512
9	FFN-ReLU-BN	t	1536
10	Multi-Head Self-Att. Pool.	Full Seq.	H*1536*2
11	FFN-ReLU-BN	Full Seq.	512
12	FFN-ReLU-BN	Full Seq.	512
13	Softmax	Full Seq.	N_{spk}

Table 1: *Architecture for the embedding extractor. Involved elements are Time Delay Neural Network (TDNN), Rectified Linear Unit (ReLU), Batch Normalization (BN) and Feed Forward Network (FFN). Context explains which frames at the input of the layer are taken into account to build the t-th output frame. Layers with full sequence (Full Seq.) context work at utterance level.*

$\boldsymbol{\Phi}$ in order to obtain the diarization labels Θ_{DIAR}:

$$\Theta_{\mathrm{DIAR}} = \arg\max_{\theta} P\left(\Theta|\boldsymbol{\Phi}\right) = \arg\max_{\theta} P\left(\boldsymbol{\Phi}|\Theta\right) P\left(\Theta\right) \tag{2}$$

The model considers a Fully Bayesian PLDA [20] of dimension 100 to explain $P\left(\boldsymbol{\Phi}|\Theta\right)$, while the priors $P\left(\Theta\right)$ follow [21] making use of a modification of the Distance Dependent Chinese Restaurant (ddCR) process. Additionally, we interpret $P\left(\boldsymbol{\Phi}, \Theta\right)$ as a tree structure by means of the product rule of probability. Hence, we opt for an optimization of the model according to a sequential manner making use of the M-algorithm [22] to find the best possible path along the tree. Moreover, prior to any clustering evaluation, the PLDA model is adapted thanks to unsupervised adaptation approaches as described in [6].

3. Identity Assignment

The Identity Assignment (IA) block in ViVoLAB submission follows the schematic of a speaker verification task based on the standard embedding-PLDA paradigm. Hence, as preparation each one of the enrollment recordings is converted into its corresponding embedding as well as the obtained segments from diarization. For the speaker verification evaluation, enrollment models are built according to the corresponding given audios while test models represent the clusters obtained during diarization. Each test model is made in terms of all segments assigned to the cluster. For simplicity reasons we make use of the same embedding extractor and PLDA trained for diarization purposes.

The obtained scores are then normalized by means of adaptive S-normalization

$$s' = \frac{s - \mu_t}{\sigma_t} + \frac{s - \mu_e}{\sigma_e} \tag{3}$$

where the score s is transformed into the score s' in terms of the the means μ_t and μ_e and standard deviations σ_t and σ_e. While μ_t and σ_t are computed on the scores of the cohort versus the test segments, μ_e and σ_e are computed on the scores of

Table 2: *Results of the ViVoLAB system for the two subtasks, diarization and identity assignment. Results obtained for two subsets, development and test.*

Metric	Results (%)	
	Dev.	Test
Diarization		
DER	16.96	15.24
Identity Assignment		
AER	47.90	72.63
ASE	128.11	505.82

the enrollment segments versus the cohort. The chosen normalization cohort in our experiments consisted of the MGB 2015 dataset. The score normalization was adaptive. For each segment, we select cohorts similar to the test segment to compute the normalization values. For each trial, we selected 25% of the total segments in the cohort. The selection is based on the own PLDA scores.

The final labels are built according to a threshold adjusted during calibration. This adjustment was obtained experimentally with the development set. Furthermore, as a design choice we do not exclude the possibility of multiple clusters assigned to the same enrollment. This decision was made in order to allow the correction of diarization errors.

4. Results

Due to the fact that 2020 evaluation considers two different tasks, we need metrics for both of them. Regarding diarization, the evaluation takes into account the commonly used Diarization Error Rate (DER) , which determines the ratio of mislabeled audio to the total audio to analyze. With respect to identity assignment two metrics were proposed: On the one hand we have Assignment Error Rate (AER), similar to DER, although only matching clusters from hypothesis and reference when sharing the same speaker label. On the other hand we have Average Speaker Error (ASE), an average of the ratio of error per speaker along the subset of interest. Once introduced the three metrics, the results for ViVoLAB system are shown in Table 2.

The results given in Table 2 evidence multiple trends. First, diarization results offer a good performance, not suffering from great mismatches between development and test. This is very important when involved shows within both subsets may not match. Besides, these results follow the trend from previous evaluations [9], specially considering the addition of more complex audio. With respect to Identity Assignment results, we observe a high degradation compared to DER results in both types of metric (AER and ASE). In fact, this degradation is much more severe in the second metric (ASE). Moreover, this degradation does not affect development and test in a similar way but specially harms evaluation scores.

The first analysis of interest studies the decomposition of DER into its three decoupled terms: miss (speech not considered to contain human voice), false alarm (audio mistakenly labeled to contain speech) and speaker (speech misclassified among the speakers). While the first two are related to the VAD stage, the latest one is only influenced by the SCPD as well as clustering. The results for this analysis are illustrated in Table 3:

Table 3: *Decomposition of DER into its three terms, Miss, False Alarm (F.A.) and Speaker (SPK). Analysis performed for both development and test subsets.*

Subset	Errors(%)		
	MISS	F.A.	SPK
Development	3.61	2.74	10.61
Test	3.65	1.97	9.62

Table 4: *Decomposition of AER in False Alarm (F.A.), Miss (MISS) and Speaker (SPK) errors for development and test subsets*

Subset	Errors(%)		
	MISS	F.A.	SPK
DEV	14.0	29.6	4.3
TEST	5.2	57.0	14.5

The results in Table 3 evidence a reasonable good performance of the VAD block, being the posterior blocks responsible for most of the diarization error. This VAD performance is specially relevant when bearing in mind that its labels work as anchors for posterior stages and its errors cannot be compensated afterwards. Moreover we also want to highlight the robustness of this VAD block, working similarly with both subsets thanks to its generalization capabilities.

Apart from the traditional diarization subtask, the interest for identity assignment as well as its poor results encourages to take a deeper look into the new subtask. Our first analysis is a decomposition of AER into its three composite terms:

- E_{MISS}, or miss error, determines how much speech is lost.
- E_{FA}, also known as false alarm error, illustrates how much non-desired audio is considered as speech.
- E_{SPK}, named as speaker error, indicates how much speech is mistakenly assigned among the speakers.

This analysis is carried out for both development and test subsets. The obtained results are included in Table 4.

According to those results shown given in Table 4 we can conclude that the main cause of error is the False Alarm term. In fact, this error implies at least a relative 60% of the whole error in both subsets. Consequently too much undesired audio is considered as coming from the target speakers. This error also explains the high values for ASE, highly affecting speakers of interest with limited speech contributions. Besides, the differences in this error term between development and test subsets are responsible for the great difference in performance between both subsets.

Nevertheless, we still must take into account the way ViVoLAB system faced the IA task, i.e. by means of a speaker verification evaluation. In this evaluation enrollment models are created according to the given audios while test models are built according to diarization labels. Our next analysis studies DET curves for both development and test subsets. These curves are illustrated in Fig. 1, also including the EER (Equal Error Rate) value. Please notice that this analysis omits the amount of speech contained in each cluster, treating them evenly for scoring purposes.

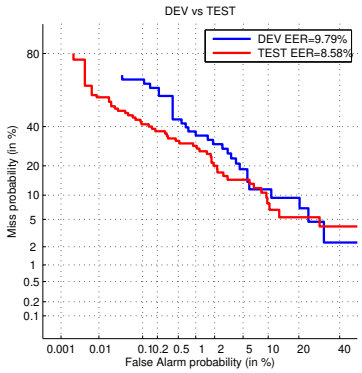

Figure 1: *DET curves for development (blue) and test (red) subsets, also indicating the EER.*

According to the curves shown in Fig. 1 we see high degradation values for both subsets. Moreover, along the whole curves the test subset is always outperforming its development counterpart regardless of the operation point. These results seem to disagree with the previously obtained results.

Our final study consists of analyzing the score distribution per subset, development and test, as well as type of trial, target and non-target. This analysis is carried out considering score histograms. In Fig. 2 we reproduce our analysis, illustrating in thick and dashed lines the scores for non-target and target trials respectively. In the same figure we have included the scores for both development (blue line) and test (red line) subsets.

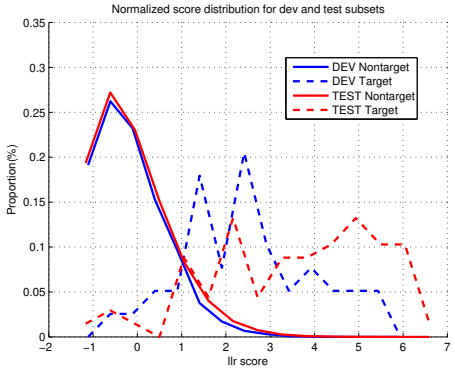

Figure 2: *Score histograms for development (blue) and test (red) subsets, differentiating between non-target (thick lines) and target (dashed lines) trials*

The histograms in Fig. 2 explain the behaviour of the system. While non-target trials offer similar score distributions for development and test subsets, target trials do not. Target distributions for development and test subsets are slightly shifted from each other. Hence, the calibration task performed during development is not valid anymore with test data. This circumstance also explains why a better DET curve provides worse IA results. This mismatch may have many reasons, such as quality of the clusters as well as the domain mismatch between enrollment and test data. This latest factor is specially relevant when comparing diarization with IA tasks. While diarization does only deal with a single domain, given by the test audio, IA may have few of them, including the test audio as well as those present in the enrollment data. In addition to this factor IA must also fit a threshold, ideally robust against any of these mismatches.

5. Conclusions

Along the previous lines we have seen the performance of ViVoLAB system in two different tasks, diarization and identity assignment. While its performance in diarization seems promising with good results, we notice a great degradation when evaluating IA.

Diarization results follow the trend of performance from previous evaluations despite the higher complexity of 2020 data. Moreover, unsupervised adaptation techniques help minimizing the mismatch between development and test subsets, offering similar results.

By contrast, IA results suffer from a significant degradation. According to the carried out analysis we see that the main source of degradation is the large amount of false alarm errors, i.e. undesired speakers considered as part of the target ones. This type of degradation specially influences ASE metric. In addition to this factor, the speaker verification treatment of the I problem suffers from mismatch in score histograms between enrollment and test, causing a calibration issue. This can be partially caused by the quality of clusters, given by diarization, as well as acoustic domain mismatches of data, both enrollment and test. Regardless of the cause, its solution is capital for the robustness of this type of systems and its addition to real world applications.

6. References

[1] S. E. Tranter and D. Reynolds, "An Overview of Automatic Speaker Diarization Systems," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 5, pp. 1557–1565, 2006.

[2] F. Valente, P. Motlicek, and D. Vijayasenan, "Variational Bayesian Speaker Diarization of Meeting Recordings," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2010, pp. 4954–4957.

[3] C. Vaquero, A. Ortega, A. Miguel, and E. Lleida, "Quality Assessment of Speaker Diarization for Speaker Characterization," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 21, no. 4, pp. 816–827, 2013.

[4] J. Villalba, A. Ortega, A. Miguel, and E. Lleida, "Variational Bayesian PLDA for Speaker Diarization in the MGB Challenge," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2015, pp. 667–674.

[5] D. Garcia-Romero, D. Snyder, G. Sell, D. Povey, and A. McCree, "Speaker Diarization Using Deep Neural Network Embeddings," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 4930 – 4934.

[6] I. Viñals, A. Ortega, J. Villalba, A. Miguel, and E. Lleida, "Unsupervised adaptation of PLDA models for broadcast diarization," *Eurasip Journal on Audio, Speech, and Music Processing*, vol. 2019, no. 1, 2019.

[7] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, L. McCowan, W. Post, D. Reidsma, and P. Wellner, "The AMI Meeting Corpus: A pre-announcement," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 3869 LNCS, pp. 28–39, 2006.

[8] T. Butko and C. Nadeu, "Audio segmentation of broadcast news in the albayzin-2010 evaluation: overview, results, and discussion,"

EURASIP Journal on Audio, Speech, and Music Processing, vol. 2011, no. 1, pp. 1–10, 2011.

[9] E. Lleida, A. Ortega, A. Miguel, V. Bazán, C. Pérez, M. Gómez, and A. de Prada, "Albayzin 2018 evaluation: The IberSpeech-RTVE challenge on speech technologies for Spanish broadcast media," *Applied Sciences*, vol. 9, no. 24, pp. 1–22, 2019.

[10] P. Bell, M. J. F. Gales, T. Hain, J. Kilgour, P. Lanchantin, X. Liu, A. McParland, S. Renals, O. Saz, M. Wester, and P. C. Woodland, "The MGB Challenge: Evaluating Multi-Genre Broadcast Media Recognition," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2015, pp. 687–693.

[11] S. S. Chen and P. Gopalakrishnan, "Speaker, Environment and Channel Change Detection and Clustering Via the Bayesian Information Criterion," *Proc. DARPA Broadcast News Transcription and Understanding Workshop*, vol. 6, pp. 127–132, 1998.

[12] S. B. Davis and P. Mermelstein, "Comparison of Parametric Representations for," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.

[13] M. J. Alam, P. Ouellet, P. Kenny, and D. O'Shaughnessy, "Comparative evaluation of feature normalization techniques for speaker verification," *Advances in Nonlinear Speech Processing. NOLISP 2011. Lecture Notes in Computer Science*, vol. 7015, no. 2011, pp. 246–253, 2011.

[14] J. Villalba, N. Chen, D. Snyder, D. Garcia-Romero, A. McCree, G. Sell, J. Borgstrom, F. Richardson, S. Shon, F. Grondin, R. Dehak, L. P. Garcia-Perera, D. Povey, P. Torres-Carrasquillo, S. Khudanpur, and N. Dehak, "State-of-the-art Speaker Recognition for Telephone and Video Speech: the JHU-MIT Submission for NIST SRE18," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pp. 1488–1492, 2019.

[15] D. Bahdanau, K. Cho, and Y. Bengio, "NEURAL MACHINE TRANSLATION BY JOINTLY LEARNING TO ALIGN AND TRANSLATE," in *International Conference on Learning Representations (ICLR)*, 2015, pp. 1–15.

[16] A. Nagrani, J. S. Chung, and A. Zisserman, "VoxCeleb: A large-scale speaker identification dataset," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2017-Augus, pp. 2616–2620, 2017.

[17] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep Speaker Recognition," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2018-Septe, no. ii, pp. 1086–1090, 2018.

[18] D. Garcia-Romero and C. Y. Espy-Wilson, "Analysis of I-vector Length Normalization in Speaker Recognition Systems," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2011, pp. 249–252.

[19] I. Viñals, P. Gimeno, A. Ortega, A. Miguel, and E. Lleida, "ViVoLAB Speaker Diarization System for the DIHARD 2019 Challenge," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pp. 988–992, 2019.

[20] J. Villalba and E. Lleida, "Unsupervised Adaptation of PLDA By Using Variational Bayes Methods," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 744–748.

[21] A. Zhang, Q. Wang, Z. Zhu, J. Paisley, and C. Wang, "Fully Supervised Speaker Diarization," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6301–6305.

[22] A. J. Viterbi, "Error Bounds for Convolutional Codes and an Asymptotically Optimum Decoding Algorithm," *IEEE Transactions on Information Theory*, vol. 13, no. 2, pp. 260–269, 1967.