Voice Activity Detection on Italian Language

Shibingfeng Zhang[1], Gloria Gagliardi[1] and Fabio Tamburini[1]

[1]FICLIT, Alma Mater Studiorum - University of Bologna, via Zamboni, 32, Bologna, Italy

Abstract

Voice Activity Detection (VAD) refers to the task of identifying human voice activity in noisy settings, playing a crucial role in fields like speech recognition and audio surveillance. However, most VAD research focuses on English, leaving other languages, such as Italian, under-explored. This study aims to evaluate and enhance VAD systems for Italian speech, with the goal of finding a solution for the speech segmentation component of the Digital Linguistic Biomarkers (DLBs) extraction pipeline for early mental disorder diagnosis. We experimented with various VAD systems and proposed an ensemble VAD system. Our ensemble system shows improvements in speech event detection. This advancement lays a robust foundation for more accurate early detection of mental health issues using DLBs in Italian.

Keywords

Voice Activity Detection, Digital Linguistic Biomarkers, Speech Processing, Speech Segmentation

1. Introduction

Voice Activity Detection (VAD) refers to the task of identifying the presence of human voice activity in noisy speech, classifying utterance segments as "speech" or "non-speech". Typically, it involves making binary decisions on each frame of a noisy signal [1]. VAD has a wide range of applications, serving as a crucial component in various fields such as telecommunications, speech recognition systems, and audio surveillance. Nevertheless, the great majority of current works focus on the application of VAD to English while there are many aspects that can affect the performance of transferring a VAD system from one language to another, potentially leading to suboptimal results. For instance, voice onset time may vary significantly between languages, affecting the system's ability to detect speech activity accurately [2]. Additionally, differences in phonetic structures can further complicate the system's effectiveness across languages. Given these factors, conducting research to evaluate various VAD systems on Italian speech would be highly valuable.

Digital Linguistic Biomarkers (DLBs) indicate linguistic features automatically extracted directly from patients' verbal productions that provide insights into their medical state [3]. Gagliardi and Tamburini [3] proposed the first DLBs extraction pipeline for the early diagnosis of mental disorders in Italian. The extraction of acoustic and rhythmic features relies heavily on the preprocessing step which consists of speech segmentation via VAD. The VAD system adopted by Gagliardi and Tamburini [3] is a statistical VAD system named "SSVAD v1.0" [4], which will be presented and compared to other VAD systems in Section 2.

In this project, we focus on VAD for the Italian language, an area that remains largely unexplored, aiming to find a VAD system that performs better and is more reliable than the one adopted in the original pipeline. The outcomes of this project will serve as a fundamental component in the pipeline for extracting DLBs and replacing the current VAD system. Moreover, our efforts will provide a robust foundation for future work in this domain, facilitating more accurate and early detection of mental health issues using linguistic biomarkers.

Our main contributions are as follows:

- Testing and evaluating various VAD systems on Italian speech.
- Proposing an ensemble VAD system that achieves superior results.

This paper is structured into five sections. Section 2 presents the data resources and VAD systems leveraged in this work. Section 3 details the experiments and resources for testing VAD systems. Section 4 presents and discusses the experimental results. Finally, Section 5 draws conclusions.

2. Background

This section outlines the background, state-of-the-art developments, and architectures of VAD systems.

The majority of Voice Activity Detection (VAD) systems approach the task as a binary classification for each frame of a noisy audio signal, with or without overlaps between frames. Based on their architecture, these systems

CLiC-it 2024: Tenth Italian Conference on Computational Linguistics, Dec 04 — 06, 2024, Pisa, Italy

shibingfeng.zhang@unibo.it (S. Zhang); gloria.gagliardi@unibo.it (G. Gagliardi); fabio.tamburini@unibo.it (F. Tamburini)

https://www.unibo.it/sitoweb/shibingfeng.zhang (S. Zhang); https://www.unibo.it/sitoweb/gloria.gagliardi (G. Gagliardi); https://www.unibo.it/sitoweb/fabio.tamburini (F. Tamburini)

0009-0005-7320-9088 (S. Zhang); 0000-0001-5257-1540 (G. Gagliardi); 0000-0001-7950-0347 (F. Tamburini)

can generally be divided into two categories: statistical VAD systems and deep neural network (DNN) VAD systems.

Statistical VAD systems rely on probabilistic models and statistical signal processing techniques to distinguish between speech and non-speech segments. Common statistical methods include Gaussian Mixture Models (GMM), Hidden Markov Models (HMM), and Bayesian frameworks. For example, Sohn et al. [5] proposed a robust statistical VAD system that models the signal using a first-order two-state HMM. In this system, the VAD score of each frame is calculated based on the likelihood ratio between the probability density functions conditioned on two hypotheses: speech absent and speech present. Additionally, the state-transition probability is determined using the likelihood ratio from the previous frame, which helps in maintaining temporal coherence and improving the accuracy of voice activity detection.

On the other hand, VAD systems based on DNNs leverage the power of deep learning. These systems use neural network architectures, such as convolutional neural networks (CNNs), recurrent neural networks (RNNs), or more advanced structures with attention mechanism [6].

Below, we present the list of the VAD systems we experimented with in this project, along with a brief description of each system:

SSVAD v1.0 (Baseline) [4] is a statistical VAD system designed to handle low signal-to-noise-ratio (SNR), impulsive noise, and cross talks in interview-style speech files. The system enhances speech segments as a pre-processing step to improve SNR, thereby facilitating subsequent speech/non-speech decisions. SSVAD v1.0 was previously integrated into the older version of the DLBs extraction pipeline [7] for speech segmentation and serves as the baseline for comparison with other systems in this study.

rVAD [8] is an unsupervised model comprising two denoising steps followed by a final VAD stage. In the first denoising step, high-energy noise segments are identified and nullified. The second step utilizes a speech enhancement method to further denoise the signal.

Silero [9] is a pre-trained CNN systems with encoder-decoder architecture. Detailed information about this VAD system is limited, as it is closed source and undocumented.

WebRTC VAD is a system developed by Google for the WebRTC project[1]. Similar to the Silero VAD system, it is closed source and detailed information about its architecture are not publicly available.

GPVAD [10] is a 5-layer framework composed of CNN and RNN layers. The proposed model employs a data-driven teacher-student learning paradigm for VAD, where a teacher model is initially trained on a source dataset with weak labels to handle vast and noisy audio data. The trained teacher model then provides frame-level guidance to a student model trained on various unlabeled target datasets.

Context-aware VAD [11] is a self-attentive VAD system based on the Transformer architecture [12]. The proposed self-attentive VAD model processes acoustic features extracted from audio input, enhancing it with contextual information from surrounding frames.

Pyannote [13] is a pre-trained open-source toolkit for audio processing that involves a VAD model. Similar to GPVAD and Silero, it is a DNN-based model with CNN and RNN components.

3. Experiments

This section provides an overview of the experiments we conducted, the evaluation metrics applied, and the resources adopted for the experiments.

3.1. Evaluation Dataset

In this work, the CLIPS dataset (Corpora e Lessici dell'Italiano Parlato e Scritto, Italian for *Corpora and Lexicons of Spoken and Written Italian*)[2] [14] is adopted to evaluate different VAD systems.

CLIPS comprises approximately 100 hours of speech data, equally distributed between male and female voices. It includes a diverse range of regional and situational speech samples to ensure a comprehensive representation of the Italian language across different contexts. The CLIPS dataset is organized into five subsets, with the "DIALOGICO" and "LETTO" subsets offering complete temporal alignments between audio and textual transcription, totaling approximately 7.5 hours of test data. The "DIALOGICO" subset includes dialogues between two interlocutors, while the "LETTO" subset consists of recordings where words are read aloud from lists.

3.2. Experiment Settings & Evaluation

To thoroughly evaluate the performance of various VAD systems, we used two sets of metrics: segment-level metrics and event-level metrics. Segment-level metrics treat each 10ms segment of audio (a single frame) independently, calculating metrics such as F1 score, precision, recall, error rate, and accuracy. Event-level metrics, on the other hand, consider each speech segment as a unit. A prediction is deemed correct if its overlap with the ground truth exceeds 50%, and the same metrics are calculated accordingly.

[1] https://webrtc.org/

[2] http://www.clips.unina.it/it/

Experiments were conducted on CLIPS dataset using the VAD systems outlined in Section 2. To achieve optimal results, all systems were tested on their default frame size. Furthermore, we combined systems' predictions through different ensemble methods to enhance performance further. More details on these ensemble methods are provided in Section 4.2.

4. Results

This section presents and analyses the experimental results of different VAD systems.

4.1. Single Systems Evaluation

Table 1 shows the experimental results obtained from the systems described in Section 2. The evaluation results are derived using the methods presented in Section 3.2.

Table 1
Results of VAD experiment on different systems. For segment-level results, each 10ms is considered one segment. For event-level results, a prediction is considered correct if its overlap with the ground truth exceeds 50%. The evaluation metric used is the F1 score.

Method	Segment-level	Event-level
Context-aware VAD	60.4	12.1
SSVAD (Baseline)	62.2	23.1
WebRTC	64.6	27.0
rVAD	69.5	72.2
GPVAD	89.5	72.3
Pyannote	92.3	**80.3**
Silero	**92.5**	80.1

As can be seen, the majority of the tested systems outperformed the baseline system SSVAD used in the current DLB pipeline at the segment level. A notable pattern from the experiment results is that DNN-based systems, such as Silero, GPVAD, and Pyannote, tend to achieve better results compared to traditional statistical systems like rVAD and SSVAD. However, context-aware VAD is an exception, with an F1 score of 60.4, which is lower than the baseline SSVAD score of 62.2. As for event-level results, similar to the segment-level results, almost all systems outperformed the baseline. DNN-based systems tend to perform better, with Context-aware VAD being again an exception, as its F1 score is the lowest among all systems. The poor performance of Context-aware VAD could be attributed to the fact that, unlike GPVAD and Pyannote, it is trained only on the TIMIT [15] dataset with additional background noise. The TIMIT dataset is a relatively small English speech dataset, containing only 5 hours of audio, likely causing the system to overfit on this dataset. Another possible reason for this relatively poor performance could be that, while Pyannote and GPVAD are trained on multilingual datasets like DIHARD III [16] and Audioset [17], Context-aware VAD is trained solely on English speech. When tested on Italian speech, the system could suffer a domain shift, resulting in diminished performance.

To gain a better understanding of the differences in system performance, a Kruskal-Wallis test was conducted. The results indicate that both the differences between segment-level results and event-level results are significant. A Dunn's test was then performed for post-hoc comparisons. The statistical analysis demonstrates that systems GPVAD, rVAD, Silero, and Pyannote exhibit similar performance at both the segment and event levels, while SSVAD, WebRTC, and Context-aware VAD show significantly lower performance at both levels.

After considering the performance at different levels, we tested all combination of three systems to form an ensemble prediction system to generate more accurate VAD results. The architectures of these ensemble systems and the corresponding experimental results are discussed in the following section.

4.2. Ensemble Systems Evaluation

This section details the ensemble methods that combine predictions of systems tested in Section 4.1. It subsequently presents the experimental results and analysis.

Of the systems presented in Section 2, Silero, Pyannote, GPVAD, and Context-aware VAD assign a score to each frame with a threshold used for making predictions. The other systems do not generate such scores, either due to differences in their architecture or because they are closed-source. This score can be interpreted as the probability of the frame being speech or not. We attempted to ensemble system's predictions using both the probability scores and their final predictions. The major challenge faced by these ensemble methods is that each system uses a different frame size, which complicates achieving alignment for the ensemble system.

We proposed and tested several ensemble strategies:

- **Probability Voting (PV)**: This method involves summing and averaging the probability scores from different predictions.
- **Probability Voting with Frame (PV_f)**: In this approach, each audio is first segmented into frames. For each frame, we identify all overlapping frames from all predictions, average their probability scores, and use this average as the probability score for the frame. The frame size of PV_f is 200 ms.
- **Simple Voting with Frame(SV_f)**: Similar to PV_f, this method segments audio into frames. However, instead of averaging probability scores, it performs simple majority voting based on the

predictions of overlapping frames. The frame size of SV_f is 200 ms.

- **Probability Voting with Weight (PV_w)**: This method is akin to PV_f but with a twist: probability scores of overlapping frames from the three predictions are weighted according to their overlap percentage. These weighted scores are then summed to determine the probability score for each frame.
- **Probability Voting with Sampling (PV_s)**: For a given audio, this method samples timestamps. For each timestamp, it calculates the mean of the probability scores from the three systems, using this mean as the probability score for the timestamp. The sampling rate of PV_s is approximately 33.33 Hz, meaning that one point is sampled every 0.03 seconds.
- **Probability Voting with Bézier curve modelling (PV_b)**: For each prediction from each system, a Bézier curve is generated using control points sampled from the prediction. This approach aims to use a smooth curve to model the prediction and address the alignment issues caused by different frame sizes of the systems. Similar to PV_f, each audio segment is divided into frames, and the probability score for each frame is the average of the scores estimated by the Bézier curves. The sampling rate of control points that are used to generate Bézier curve in PV_b is 5 Hz (0.2 seconds).

We experimented with all possible system combinations using the SV_f ensemble method, as well as all possible combinations of Silero, Pyannote, GPVAD, and Context-aware VAD using other probability-based ensemble methods, as these are the only systems that generate probability scores. For all probability-based methods, the "speech/non-speech" prediction for each frame is determined by applying a threshold of 0.5 to the probability score.

Table 2 presents results of all possible combinations to compose the ensemble system using SV_f method. Table 3 presents results of all possible combinations to compose the ensemble systems using probability score related methods. The evaluation results are derived using the methods presented in Section 3.2.

As shown in Table 2, the ensemble created using the SV_f method did not yield better results than the individual systems at the segment level. The highest segment-level score of 91.5 was achieved by the combination of GPVAD, Silero, and Pyannote, which is still 0.6 lower than the best performance of the Silero system alone. However, at the event level, the same combination achieved the highest score among all ensemble systems, with an F1 score of 84.0, which is higher than the best score achieved by a single system. Meanwhile, all other combinations yielded scores lower than the best performance of the individual systems.

As shown in Table 3, the ensemble systems related to probability score did not achieve results that are prominently better than single systems at the segment level either, with PV_s and PV_b systems of the combination Pyannote, GPVAD, Silero being only slightly higher by a small margin of 0.6 compared to Silero. However, at the event level, several evident improvements can be observed in the performance of the ensemble systems. Probability-based ensemble systems combining Pyannote, GPVAD, Silero, except for PV_b and PV, outperformed the simple systems at event level, with PV_f achieving an F1 score of 85.9, which is 5.6 points higher than that of Pyannote. This result demonstrates that the ensemble approach can lead to substantial performance gains in detecting the temporal interval in which speech takes place. It is worth noticing that the ensemble system PV_b consistently shows great disparity between its performance at segment level and event level across all combinations. Despite its good performance on segment level, PV_b achieves rather F1 score on event level, far lower than all other systems. The disparity of performance at different levels is likely to be caused by the insufficient number of control points adopted for generating the Bézier curve. However, increasing the number of control points is infeasible due to the computational complexity of the curve, which is $O(n^2)$, with n being the number of control points.

Given that the ensemble systems composed of GPVAD, Silero, and Pyannote consistently outperformed other combinations across all ensemble methods, a Kruskal-Wallis test, followed by Dunn's post-hoc test, was conducted to assess the differences in performance between the ensemble methods and the individual systems of GPVAD, Silero, Pyannote. At the segment level, the Kruskal-Wallis test indicates that the differences are not significant. However, at the event level, the results reveal that PV_b's performance is significantly lower compared to the other systems.

In summary, given the performance of the systems, we plan to adopt PV_f as the speech segmentation component of the DLBs extraction pipeline, leveraging the combined predictions of Pyannote, Silero, and GPVAD. While PV_f shows slightly lower segment-level performance compared to the top-performing individual system, it enhances the accuracy in identifying speech intervals. This trade-off is justified by the substantial improvement in speech event detection performance.

Table 2
Results of VAD experiments on using SV_f ensemble method. For comparison, results from individual systems that achieved the best performance, Silero and Pyannote, are also included. S stands for segment-level result. E stands for event-level result. C-a stands for Context-aware VAD system. For segment-level results, each 10ms is considered one segment. For event-level results, a prediction is considered correct if its overlap with the ground truth exceeds 50%. The evaluation metric used is the F1 score.

Involved Systems	S	E
Silero	**92.5**	80.1
Pyannote	92.3	80.3
GPVAD, Silero, Pyannote	91.5	**84.0**
GPVAD, C-a, WebRTC	58.4	62.0
GPVAD, SSVAD, C-a	66.0	17.6
GPVAD, SSVAD, WebRTC	58.9	76.6
Pyannote, C-a, WebRTC	60.6	70.1
Pyannote, GPVAD, C-a	81.5	42.1
Pyannote, GPVAD, SSVAD	83.3	58.1
Pyannote, GPVAD, WebRTC	61.3	55.3
Pyannote, SSVAD, C-a	68.6	17.7
Pyannote, SSVAD, WebRTC	60.9	72.6
SSVAD, C-a, WebRTC	47.0	29.8
Silero, C-a, WebRTC	60.7	70.0
Silero, GPVAD, C-a	81.8	43.1
Silero, GPVAD, SSVAD	83.6	57.7
Silero, GPVAD, WebRTC	61.4	59.9
Silero, Pyannote, C-a	84.4	52.5
Silero, Pyannote, SSVAD	85.9	68.7
Silero, Pyannote, WebRTC	62.0	47.9
Silero, SSVAD, C-a	68.8	17.5
Silero, SSVAD, WebRTC	60.8	73.0
rVAD, C-a, WebRTC	52.2	41.4
rVAD, C-a, WebRTC	52.2	41.4
rVAD, GPVAD, C-a	71.1	29.0
rVAD, GPVAD, SSVAD	74.3	42.5
rVAD, GPVAD, WebRTC	58.4	79.3
rVAD, Pyannote, C-a	73.4	27.5
rVAD, Pyannote, GPVAD	83.5	75.1
rVAD, Pyannote, SSVAD	76.7	43.2
rVAD, Pyannote, WebRTC	60.8	58.7
rVAD, SSVAD, C-a	56.8	18.1
rVAD, SSVAD, WebRTC	54.0	63.0
rVAD, Silero, C-a	73.5	27.1
rVAD, Silero, GPVAD	83.6	73.5
rVAD, Silero, Pyannote	86.3	82.4
rVAD, Silero, SSVAD	76.8	42.2
rVAD, Silero, WebRTC	61.0	63.3

Table 3
Results of VAD experiments on using probability score related ensemble methods. For comparison, results from individual systems that achieved the best performance, Silero and Pyannote, are also included. Method stands for ensemble method adopted. S stands for segment-level result. E stands for event-level result. C-a stands for Context-aware VAD system. For segment-level results, each 10ms is considered one segment. For event-level results, a prediction is considered correct if its overlap with the ground truth exceeds 50%. The evaluation metric used is the F1 score.

Involved Systems	Method	S	E
Silero	-	92.5	80.1
Pyannote	-	92.3	80.3
Pyannote, GPVAD, Silero	PV	91.5	67.9
Pyannote, GPVAD,Silero	PV_f	91.9	**85.9**
Pyannote, GPVAD, Silero	PV_s	**93.1**	81.8
Pyannote, GPVAD, Silero	PV_w	91.8	85.6
Pyannote, GPVAD,Silero	PV_b	93.0	9.5
Pyannote, GPVAD, C-a	PV	87.2	60.4
Pyannote, GPVAD, C-a	PV_f	87.6	80.0
Pyannote, GPVAD, C-a	PV_s	89.3	79.4
Pyannote, GPVAD, C-a	PV_w	87.5	79.2
Pyannote, GPVAD, C-a	PV_b	89.2	10.5
Silero, GPVAD, C-a	PV	85.4	50.6
Silero, GPVAD, C-a	PV_f	85.7	72.7
Silero, GPVAD, C-a	PV_s	84.2	67.3
Silero, GPVAD, C-a	PV_w	85.6	71.6
Silero, GPVAD, C-a	PV_b	88.8	11.0
Silero, Pyannote, C-a	PV	89.4	70.4
Silero, Pyannote, C-a	PV_f	89.6	81.2
Silero, Pyannote, C-a	PV_s	89.5	77.7
Silero, Pyannote, C-a	PV_w	89.6	81.5
Silero, Pyannote, C-a	PV_b	89.6	9.3

5. Conclusions

In this study, we explored and enhanced Voice Activity Detection systems for the Italian language, a relatively under-explored area in speech processing. We experimented with various systems and integrated systems into an ensemble to improve detection accuracy. Our findings indicate that combining predictions from multiple models can lead to better results in detecting speech temporal intervals. This effective ensemble method will be used as a component of a Digital Linguistic Biomarkers extraction pipeline.

By enhancing the accuracy of speech segmentation, this method provides a more reliable foundation for extracting meaningful linguistic features for the diagnosis of cognitive impairment. Future research could focus on refining the ensemble method by incorporating additional linguistic features into VAD systems and exploring their synergistic effects. Additionally, investigating the application of this approach to other languages and dialects could expand its utility.

Acknowledgements

This study was funded by the European Union – NextGenerationEU programme through the Italian National Re-

covery and Resilience Plan – NRRP (Mission 4 – Education and research), as a part of the project ReMind: an ecological, costeffective AI platform for early detection of prodromal stages of cognitive impairment (PRIN 2022, 2022YKJ8FP – CUP J53D23008380006).

CRediT Author Statement

SZ: Investigation, Software, Formal analysis, Visualization, Writing - Original Draft. GG: Writing - Review & Editing, Project administration, Funding acquisition. FT: Conceptualization, Methodology, Supervision, Writing - Review & Editing.

References

[1] S. Graf, T. Herbig, M. Buck, G. Schmidt, Features for voice activity detection: a comparative analysis, EURASIP Journal on Advances in Signal Processing 2015 (2015) 1–15.

[2] T. Cho, D. H. Whalen, G. Docherty, Voice onset time and beyond: Exploring laryngeal contrast in 19 languages, Journal of Phonetics 72 (2019) 52–65.

[3] G. Gagliardi, F. Tamburini, The automatic extraction of linguistic biomarkers as a viable solution for the early diagnosis of mental disorders, in: Proceedings of the Thirteenth Language Resources and Evaluation Conference, European Language Resources Association, 2022, pp. 5234–5242.

[4] M.-W. Mak, H.-B. Yu, A study of voice activity detection techniques for nist speaker recognition evaluations, Computer Speech & Language 28 (2014) 295–313.

[5] J. Sohn, N. S. Kim, W. Sung, A statistical model-based voice activity detection, IEEE signal processing letters 6 (1999) 1–3.

[6] A. Sehgal, N. Kehtarnavaz, A convolutional neural network smartphone app for real-time voice activity detection, IEEE access 6 (2018) 9017–9026.

[7] L. Calzà, G. Gagliardi, R. R. Favretti, F. Tamburini, Linguistic features and automatic classifiers for identifying mild cognitive impairment and dementia, Computer Speech & Language 65 (2021) 101113.

[8] Z.-H. Tan, N. Dehak, et al., rvad: An unsupervised segment-based robust voice activity detection method, Computer speech & language 59 (2020) 1–21.

[9] Silero Team, Silero vad: pre-trained enterprise-grade voice activity detector (vad), number detector and language classifier, https://github.com/snakers4/silero-vad, 2021.

[10] H. Dinkel, S. Wang, X. Xu, M. Wu, K. Yu, Voice activity detection in the wild: A data-driven approach using teacher-student training, IEEE/ACM Transactions on Audio, Speech, and Language Processing 29 (2021) 1542–1555.

[11] Y. R. Jo, Y. K. Moon, W. I. Cho, G. S. Jo, Self-attentive vad: Context-aware detection of voice from noise, in: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2021, pp. 6808–6812.

[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30 (2017).

[13] H. Bredin, R. Yin, J. M. Coria, G. Gelly, P. Korshunov, M. Lavechin, D. Fustes, H. Titeux, W. Bouaziz, M.-P. Gill, Pyannote. audio: neural building blocks for speaker diarization, in: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2020, pp. 7124–7128.

[14] F. A. Leoni, F. Cutugno, R. Savy, V. Caniparoli, L. D'Anna, E. Paone, R. Giordano, O. Manfrellotti, M. Petrillo, A. De Rosa, Corpora e lessici dell'italiano parlato e scritto, 2007.

[15] J. S. Garofolo, Timit acoustic phonetic continuous speech corpus, Linguistic Data Consortium, 1993 (1993).

[16] N. Ryant, P. Singh, V. Krishnamohan, R. Varma, K. Church, C. Cieri, J. Du, S. Ganapathy, M. Liberman, The third dihard diarization challenge, arXiv preprint arXiv:2012.01477 (2020).

[17] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, M. Ritter, Audio set: An ontology and human-labeled dataset for audio events, in: Proceedings of the 2017 IEEE international conference on acoustics, speech and signal processing (ICASSP), IEEE, 2017, pp. 776–780.