Interspeech 2025
17-21 August 2025, Rotterdam, The Netherlands

Do you read me? - flow of speech effect on speaker recognition systems

Alicja Martinek[1,2], *Joanna Gajewska*[1], *Ewelina Bartuzi-Trokielewicz*[1]

[1]NASK National Research Institute, Poland
[2]AGH University of Kraków, Poland

martinek@agh.edu.pl, joanna.gajewska@nask.pl, ewelina.bartuzi-trokielewicz@nask.pl

Abstract

Comparing two types of speech – read and spontaneous – poses a significant challenge for speaker verification models. This study examines the impact of these differences on the performance of advanced biometric systems. We conducted tests using two baseline speaker verification models and two spoof-aware approaches to assess their ability to handle variations between read and spontaneous speech. Additionally, we generated synthetic speech using two state-of-the-art Text-to-Speech methods, training the models either on spontaneous or read speech. The results indicate that mixing spontaneous and read speech compared with a uniform type of speech yields higher error rates in biometric verification. The situation is similar in both testing scenarios, comparing genuines with impostors and genuines with synthetic speech, regardless of the type of speaker recognition model — base or spoof-aware.

Index Terms: speaker recognition, spoof aware speaker verification, text-to-speech, spontaneous speech, read speech

1. Introduction

Voice biometrics and voice cloning technologies have become increasingly prominent in applications ranging from secure authentication to personalized speech synthesis. Speaker verification systems aim to authenticate individuals based on their voice, while voice cloning systems utilize models to synthesize speech resembling a specific speaker.

Spontaneous speech is characterized by greater variability in intonation, tempo, and phonetic structure, as well as the presence of disfluencies and emotional expressions [1]. In contrast, read speech is typically more structured, consistent, and predictable [2]. These differences can impact the robustness and accuracy of voice biometric systems, as well as the quality and realism of synthetic speech generated by TTS models. Systematic comparisons of these effects remain underexplored, particularly when both types of speech are evaluated under comparable environmental conditions.

The primary objective of this study is to compare the impact of spontaneous and read speech on the performance of speaker verification systems. Specifically, we evaluate two commonly used speaker verification models: Resemblyzer [3] and SpeechBrain [4]. Furthermore, we analyze synthetic speech samples generated by two TTS models ElevenLabs TTS (11labs) [5] and Coqui XTTS-V2 (Coqui) [6] using both spontaneous and read speech inputs. We assess the generated samples impact on speaker verification systems and spoof-aware models designed to detect synthetic speech: MFA-Conformer [7] and SKA-TDNN [8].

To the best of our knowledge, this is the first study to systematically compare spontaneous and read speech in the context of voice cloning. Our findings contribute to understanding the challenges posed by comparing read and spontaneous speech, paving the way for the development of more robust and adaptive biometric systems and voice synthesis technologies.

2. Related work

2.1. Forging voices

The field of synthetic voice generation has made remarkable progress in recent years, largely due to advances in Text-to-Speech (TTS) and Speech-to-Speech (STS) technologies. TTS models, such as Tacotron 2 [9] and FastSpeech [10], utilizing neural network architectures like sequence-to-sequence models and transformer-based frameworks. Several researchers proposed end-to-end solutions or reduced the length of reference audio needed to achieve realistic outputs. Lina-Speech [11] extends the work on the attention mechanism. The introduction of Gated Linear Attention allows the generation of benchmark-quality samples without fine-tuning. EmoKnob [12] or ParaE-VITS [13] allow to control emotional load of synthetic audio clips.

2.2. Speaker verification and spoof aware systems

Speaker recognition, encompassing both identification and verification, is a biometric technique that automatically identifies individuals based on their speech signal [14]. Common feature extraction methods include Mel-Frequency Cepstral Coefficients (MFCC), Linear Predictive Coding (LPC), and Perceptual Linear Prediction (PLP) [15]. Recent advancements in speaker recognition employ deep learning models such as Convolutional Neural Networks (CNN) with Long Short-Term Memory (LSTM) for improved performance [16]. Speaker verification architectures such as ECAPA-TDNN [17] and ResNet-based x-vectors [18] have demonstrated robust performance in conventional settings. Beyond model architectures, research has also focused on developing open-source toolkits to advance speaker verification. Ravanelli et al. [19] introduced SpeechBrain, a flexible framework achieving competitive results across multiple benchmarks. Resemblyzer [20, **?**] is another open-source speaker encoder that has gained widespread recognition in recent literature for its robust speaker verification and identification capabilities.

Recent research has focused on developing spoofing-aware speaker verification (SASV) systems that integrate speaker verification and anti-spoofing mechanisms to combat synthetic speech attacks. The SASV 2022 challenge [21] demonstated significant advancements, with systems achieving SASV-EER as low as 0.13% on ASVSpoof 2019 dataset [22]. SASV systems have introduced models like ECAPA-TDNN [17], MFA-

10.21437/Interspeech.2025-2629

Conformer [7] and SKA-TDNN [8], which incorporate countermeasures against synthetic speech attacks. Various techniques have been explored, including probabilistic fusion frameworks [23], compositional data analysis for score fusion [24], and jointly optimized embeddings [25].

Research on read and spontaneous speech has primarily focused on classification and speech pattern analysis. Studies have explored the differences between read and spontaneous speech in various contexts, such as prosodic variation, fluency, and phonetic structure. Kopparapu [26] introduced a DeepSpeech-based approach to distinguish between these two speech styles, which is particularly relevant for applications like remote interviews and automated speech profiling. There is a growing body of evidence that speaking style directly affects speaker verification performance [27, 28], spoofing risk [29], and voice anonymization robustness [30]. For instance, phoneme durations can leak speaker identity even in anonymized speech [30], and verification accuracy drops significantly when mismatched styles are used [27]. However, there is still little know about how different speech types affect modern biometric performance, including speaker verification and spoof detection. Furthermore, the impact of synthetic speech quality, when generated from read vs spontaneous speech, remains underexplored, despite its significance for spoof-aware speaker verification systems.

3. Methodology

The experiment was designed to evaluate the impact of different speech flows on the capabilities of speaker recognition systems, namely spontaneous, free speech (Free) and reading (Read). Biometric voice verification systems can use both types of speech for the registration of audio samples. Mixing different types of speech flow in the registration and verification phase can carry a potential risk of miss-classification. The main goal of this study is to examine whether speech expression and delivery style contain significant information for speaker recognition systems. In addition, the influence of synthetically generated voices created with respect to flow of speech was measured on common speaker recognition systems.

3.1. Dataset

To evaluate the impact of spontaneous and read speech on speaker verification and spoof detection, we utilize a proprietary dataset collected specifically for this research purposes and a publicly available dataset.

Our dataset Read-and-Tell (RaT)[1] comprises 100 speakers balanced across gender and age groups (50 female, 50 male) with a total of audio, recorded under controlled conditions to ensure comparability between speech styles. Each speaker contributed: **12 read speech samples** (60 seconds each) covering diverse text types, including fiction, phonetically rich sentences, and scientific content, and **spontaneous speech recordings** of 12 minutes per speaker, ensuring natural variability in prosody, tempo, and articulation.

Unlike existing datasets, which typically focus on either read or spontaneous speech, our dataset provides a balanced representation of both, making it uniquely suited for this investigation. Publicly available datasets often mix recordings from different sources, making direct comparisons unreliable due to variations in recording conditions and speaker demographics.

[1] A standardized subset of RaT available upon request at: `https://nask-biometria.github.io/RaT/`

For external validation, we include the ALLSSTAR Corpus [31], a publicly available dataset containing both scripted and spontaneous speech. While useful for broader generalization, its acquisition conditions differ from our controlled environment, making direct comparisons limited.

This dataset was limited to native english speakers, who took part in two types of recording sessions: **read speech** - including reading passages of Declaration of Human Rights (DHR), Little Prince (LPP) or NWS task, and **spontaneous speech** - containing Q&A session delivered by 26 speakers. The collection contains 134 audio files with spontaneous speech and 208 recordings with read speech.

Speakers in this dataset are not diverse in terms of age, which ranges between 18 and 26 with the average of 19.88. Gender-wise, the group of 26 talkers has 14 females and 12 males.

3.2. Synthetic audio

To evaluate speaker recognition systems, we tested them with synthetic voices. As Text-to-Speech methods capture the flow of speech present in the training data, they were used to further extend the study. Speech-to-Speech methods often copy the delivery style of the attacker rather than the victim and hence were not used in this study.

Analysis of openly available TTS algorithms with the addition of user-friendly commercial methods facilitated the selection of the group of TTS generators used in the experiments. ElevenLabs TTS, (Multilingual v2 with default hyperparameters: Stability: 50%, Similarity: 75%, Style Exaggeration: 0, Speaker boost: True) [5], Coqui XTTS-V2 [6]. Despite ElevenLabs being paid, it offers interfaces for generations as well as API access designed for large-scale, dump creations of audio files. These properties reflect the needs of potential scammers who are not necessarily technically-proficient and can value the availability of user interface, instead of digging through multiple GitHub repositories. On the other hand, accessibility via API and the possibility of automation are valuable for more tech-aware fraudsters. The selection of accessible languages within these algorithms is also a crucial factor, as the RaT database consists of Polish language, which can be considered an under-studied language in terms of TTS algorithms.

These two methods were further used to generate synthetic speech based on two reference datasets. As the main contribution of this study revolves around flow of speech, the input text for TTS methods was chosen in a way that carries both qualities of free and read speech in a casual description answering the question "How was your day?". It was generated in languages included in the used datasets. The text was designed to take around 30 seconds to read.

3.3. Comparisons explained

To systematically analyze the impact of read and spontaneous speech on speaker verification and spoof detection, we define comparisons based on intra-class and inter-class scenarios. These comparisons evaluate both natural speech (authentic) and synthetically generated speech (synthetic), allowing for a structured assessment of speaker variability and spoofing risks. We further categorize the comparisons into three main types to examine systems' behaviors under different modes: only **free** speech, focuses on robustness of purely spontaneous signals, **read** speech, and **mix** to assess the impact of cross-style variability by comparing read and spontaneous samples.

Comparison	Read-and-Tell				AALSSTAR			
	SpeechBrain	Resemblyzer	SKA-TDNN	MFA-Conformer	SpeechBrain	Resemblyzer	SKA-TDNN	MFA-Conformer
free: intra vs inter	0.00	0.46	1.52	4.19	0.00	0.00	0.79	3.09
read: intra vs inter	0.04	0.32	1.01	3.39	0.00	0.31	0.31	10.31
auth-mix (intra vs inter)	0.17	11.57	23.28	32.92	0.24	2.62	2.88	18.75
free: auth vs synth (11labs)	1.6	0.00	0.23	1.98	1.49	7.46	0.00	0.00
read: auth vs synth (11labs)	1.72	0.00	0.63	2.71	5.77	18.27	0.00	0.00
all combinations	13.23	10.89	13.93	20.67	10.92	26.47	0.00	5.04
free: auth vs synth (Coqui)	1.37	0.15	0.00	0.08	0.75	0.00	0.00	6.72
read: auth vs synth (Coqui)	1.36	0.36	0.00	0.00	0.00	4.81	0.96	17.31
all combinations	12.86	16.97	0.12	5.13	5.25	6.30	0.42	18.28
global	9.92	13.84	16.98	**22.82**	3.68	**15.99**	2.86	13.78

Table 1: *Equal Error Rate results for different models across Read-and-Tell and AALSSTAR datasets. The lowest values are highlighted in **bold**, and the highest errors are in red.*

Additionally, we introduced comparisons between synthetic speech generated from different speech types and their corresponding authentic samples, allowing us to investigate how spoofing risks vary with read versus spontaneous inputs. **All combinations** refer to the comparison of all combinations of genuine with synthetic samples from a given TTS method.

3.4. Speaker verification systems

To evaluate the impact of spontaneous and read speech on speaker verification and spoof detection, we employed two speaker verification systems **SpeechBrain** [19] and **Resemblyzer**, along with two spoof-aware models, **MFA-Conformer** [7] and **SKA-TDNN** [8, 32]. These models were selected due to their strong performance in prior evaluations and the availability of open-source implementations.

SpeechBrain was chosen for its modular design and state-of-the art performance in verification tasks. It utilizes X-vectors and ECAPA-TDNN, containing in total 22.150M parameters, producing 192-dimensional speaker embeddings.

Resamblyzer was selected for its robust speaker embedding space, specifically designed for speaker similarity. The model uses a lightweight deep neural network architecture (1.4M parameters) with generalized end-to-end loss [33], but its exact number of parameters and training details remain undocumented in the official release. Resemblyzer generates 256-dimensional embeddings optimized for speaker verification tasks.

For detecting synthetic speech, we employed two leading spoof-aware speaker verification models. MFA-Conformer incorporates multi-scale feature aggregation with self-attentionn to enhance robustness against spoofing attacks while maintaining high speaker verification accuracy, consisting of 21.9M parameters. SKA-TDNN (30.8M params) employees Selective Kernel Attention (SKA) mechanism, allowing the model to dynamically emphasize both local and global speaker-relevant information. To ensure consistency, all models were initialized with pre-trained weights and tested on the same set of authentic and synthetic speech samples. To assess the performance of the speaker verification and spoof detection models, we utilized two key evaluation metrics: Equal Error Rate (EER) and False Rejection Rate (FRR) at fixed False Acceptance Rate (FAR) thresholds: 5%, 1%, 0.1%, 0.01%, and 0.001%.

4. Experimental setup and results

4.1. Speech-type impact on speaker verification

In this experiment, we investigate how different speech types: read, spontaneous (free), and mixed speech affect speaker verification systems. The goal is to analyze how well speech-dependent verification systems perform when processing speech samples of varying structures and spontaneity.

First three rows in Table 1 present results across all analyzed models, showing EER for intra- (same speaker) and inter-class (different speakers) comparisons. Lower score indicates better speaker verification performance. The results for the Read-and-Tell dataset show that traditional speaker verification models maintain low EERs close to 0.00 across all speech types. However, for mixed speech, the performance decreases (0.17% - 11.57%), suggesting a higher degree of variability when crossing speech styles. Spoof-aware models show moderate robustness for uniform speech styles comparison, with errors in the range of 1.01-4.19%, while they struggle the most in mixed-speech scenarios, reaching an EER in the range of 23.28% - 32.92%.

The results for the AALSSTAR dataset are quite similar. Basic speaker verification systems achieve an EER close to zero for uniform types of speech, except for mixed speech. In this case, the EER increases slightly, reaching up to 2.62% for Resemblyzer. The SKA-TDNN model classifies spontaneous and read speech almost flawlessly. However, when comparing mixed speech, the EER rises to 2.88%. Unlike the other models, the MFA Conformer makes the most errors, not only when comparing mixed speech (18.75%) but also read speech (10.31%).

4.2. Evaluating speaker verification with synthetic voices

This experiment evaluates how speaker verification systems handle synthetic speech compared to authentic speech. The analysis includes synthetic samples generated from both read and spontaneous speech, allowing us to assess whether training on different speech styles influences system performance.

The authentic vs synthetic speech comparisons in Table 1 highlight key differences in speaker verification performance depending on the training speech style of TTS. Analyzing the results for the RaT dataset, comparisons of homogeneous speech styles yield EER values ranging from 0.0% to 2.71% for both TTS methods. However, when all possible combinations of genuine and synthetic speech are compared regardless of speech style, the EER increases significantly, reaching up to 20.67% for the MFA-Conformer. The best result in this category is achieved by SKA-TDNN for synthetic samples generated using the Coqui model, with an EER of 0.12%. This is most likely due to the poor quality of the generated samples and the robustness of the spoof-aware model against deepfakes.

In the case of the AALSSTAR dataset, the situation is different. Baseline speaker verification models analyzing the 11Labs samples achieve higher EER values, especially for read

speech comparisons and all combinations comparisons (ranging from 5.77% to 26.47%). The worst-performing model is Resemblyzer, with the EER from 7.46% to 26.47%. For spoof-aware models, the results differ notably. SKA-TDNN achieves an EER of 0.0% across all types of comparisons. Similarly, MFA Conformer maintains low EER values, except in the case of all-combination comparisons, where the EER increases to 5.04%. Analyzing the results for the Coqui samples, all models perform well, with the exception of the MFA Conformer, which shows EERs ranging from 6.72% to 18.28%.. The SKA-TDNN model achieves the best classification performance, with a maximum EER of only 0.96%. In the case of typical speaker verification models, the EER ranges from 0.0% to 6.30%.

Crucially, these results emphasize the importance of threshold selection, as EER values vary based on speech type. In a global analysis, considering all authentic intra-class comparisons and impostors, including inter-class comparisons and synthetic samples, the EER increases from 9.92% to 22.82% for RaT dataset, and 2.86% - 15.99% for AALSSTAR.

To better illustrate these variations, Figures 1 and 2 present the score distributions for selected base and spoof-aware speaker verification systems on the RaT dataset, visualizing how read, spontaneous, and mixed speech types influence system behavior.

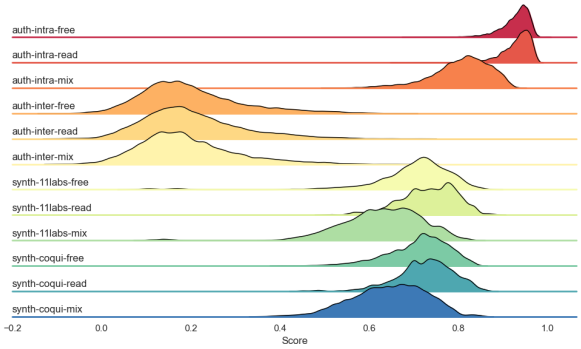

Figure 1: *Speechbrain score distributions for RaT dataset.*

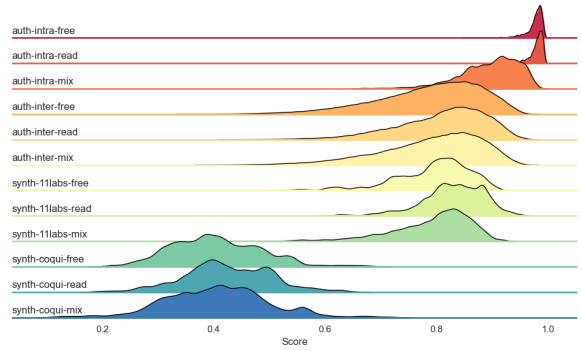

Figure 2: *SKA-TDNN score distributions for RaT dataset.*

Additionally, Figure 3 presents the FRR measured at various FAR thresholds, providing an estimate of system security across different speaker verification models and speech types. This evaluation helps assess the trade-off between strict authentication and user convenience, emphasizing the impact of speech variability on system robustness

5. Conclusion

Our findings highlight that comparing two different speech styles presents unique challenges for biometric systems,

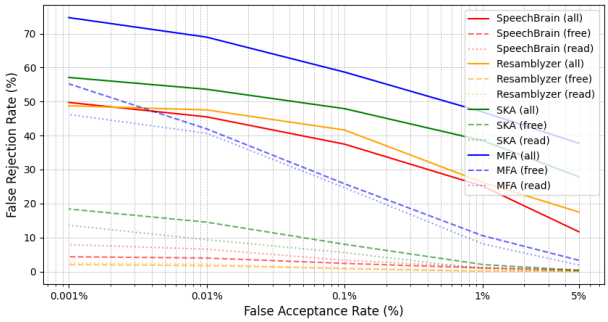

Figure 3: *False Rejection Rate at differet False Acceptance Rate threshold for system security evaluation, across different speech types and speaker verification models*

whether dealing with genuine and impostors or genuine and generated samples comparison. Addressing this issue is critical to ensuring fairness and inclusivity in voice biometrics.

Analyzing the baseline models, SpeechBrain delivers the best performance on both datasets. Among the spoof-aware models, SKA-TDNN stands out, achieving an EER close to 0.0% in nearly all cases—except for the 11labs all-combination comparisons in the RaT dataset, where the EER reaches 13.93%. The MFA-Conformer model exhibits the largest performance degradation, particularly in read and mixed speech conditions, highlighting its difficulty in adapting to different speech styles.

Future research should focus on training models on diverse speech styles, including both read and spontaneous speech, to improve resilience.

Ensuring robust and fair speaker verification across different speech styles is crucial for societal adopting of voice biometrics, particularly in authentication, and forensic analysis. The observed biases in mixed speech conditions raise concerns about unequal performance across different user groups, potentially disadvantaging individuals with more natural, unstructured speaking patterns. Addressing these issues is essential to ensure that voice-based security systems remain reliable, inclusive and resistant to adversarial attacks in real-world applications.

6. Limitations

The work presented in this study has several limitations. As every experiment is framed within the boundaries of the data, our research was conducted on constrained number of languages. However, selection of English, as the most studied language, having wide assortment of algorithms allows placing the research on the international area and makes it more relatable. On the other hand, Polish language and its unique in size and nature RaT database allows to conclude in statistically significant way. Polish remains underrepresented in biometric research despite being morphologically rich and phonetically diverse, which makes it particularly relevant.

Both datasets used in the experimentation, RaT and ALLSSTAR Corpus, operate on limited number of speakers. Another limitation lies in the number of speaker recognition systems tested.

Despite the aforementioned limitations, the experimental setup was carefully selected in a manner that facilitates good research practices and makes conclusions meaningful, relevant to everyone, and support generalization.

7. References

[1] E. Shriberg, "Spontaneous speech: How people really talk and why engineers should care," in *Proceedings of EUROSPEECH*, 2001.

[2] T. Asami, R. Masumura, H. Masataki, and S. Sakauchi, "Read and spontaneous speech classification based on variance of gmm supervectors," in *Proceedings of the Fifteenth Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2014. [Online]. Available: https://doi.org/10.21437/Interspeech.2014-516

[3] Resemble AI, "Resemblyzer Github Repository," https://github.com/resemble-ai/Resemblyzer, 2020, accessed: 2023-10-21.

[4] M. Ravanelli, T. Parcollet, P. Plantinga, A. Rouhe, S. Cornell, L. Lugosch, C. Subakan, N. Dawalatabad, A. Heba, J. Zhong, J.-C. Chou, S.-L. Yeh, S.-W. Fu, C.-F. Liao, E. Rastorgueva, F. Grondin, W. Aris, H. Na, Y. Gao, R. D. Mori, and Y. Bengio, "SpeechBrain: A general-purpose speech toolkit," 2021, arXiv:2106.04624.

[5] ElevenLabs, "Elevenlabs - webpage," accessed: January 2025. [Online]. Available: https://elevenlabs.io/

[6] E. Casanova, K. Davis, E. Gölge, G. Göknar, I. Gulea, L. Hart, A. Aljafari, J. Meyer, R. Morais, S. Olayemi, and J. Weber, "Xtts: a massively multilingual zero-shot text-to-speech model," in *Interspeech 2024*, 2024, pp. 4978–4982.

[7] Y. Zhang, Z. Lv, H. Wu, S. Zhang, P. Hu, Z. Wu, H. yi Lee, and H. M. Meng, "Mfa-conformer: Multi-scale feature aggregation conformer for automatic speaker verification," *ArXiv*, vol. abs/2203.15249, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:247779244

[8] S. H. Mun, J.-w. Jung, M. H. Han, and N. S. Kim, "Frequency and multi-scale selective kernel attention for speaker verification," in *2022 IEEE Spoken Language Technology Workshop (SLT)*, 2023, pp. 548–554.

[9] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan, R. A. Saurous, Y. Agiomvrgiannakis, and Y. Wu, "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," pp. 4779–4783, 2018.

[10] Y. Ren, Y. Ruan, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "Fastspeech: fast, robust and controllable text to speech," in *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates Inc., 2019.

[11] T. Lemerle, H. Vanderbyl, V. Srivastav, N. Obin, and A. Roebel, "Lina-speech: Gated linear attention is a fast and parameter-efficient learner for text-to-speech synthesis," 2024. [Online]. Available: https://arxiv.org/abs/2410.23320

[12] H. Chen, R. Chen, and J. Hirschberg, "Emoknob: Enhance voice cloning with fine-grained emotion control," 2024. [Online]. Available: https://arxiv.org/abs/2410.00316

[13] X. Jing, K. Zhou, A. Triantafyllopoulos, and B. W. Schuller, "Enhancing emotional text-to-speech controllability with natural language guidance through contrastive learning and diffusion models," 2024. [Online]. Available: https://arxiv.org/abs/2409.06451

[14] H. Beigi, *Speaker Recognition*, 06 2011.

[15] N. Chauhan, T. Isshiki, and D. Li, "Speaker recognition using fusion of features with feedforward artificial neural network and support vector machine," *2020 International Conference on Intelligent Engineering and Management (ICIEM)*, pp. 170–176, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:225847893

[16] B. Alkhatib and M. M. K. Eddin, "Deep learning model cnn with lstm for speaker recognition," *Journal of Digital Information Management*, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:259773122

[17] B. Desplanques, J. Thienpondt, and K. Demuynck, "Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification," 10 2020.

[18] T. Zhou, Y. Zhao, and J. Wu, "Resnext and res2net structures for speaker verification," *2021 IEEE Spoken Language Technology Workshop (SLT)*, pp. 301–307, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:263863922

[19] M. Ravanelli, T. Parcollet, P. W. V. Plantinga, A. Rouhe, S. Cornell, L. Lugosch, C. Subakan, N. Dawalatabad, A. Heba, J. Zhong, J.-C. Chou, S.-L. Yeh, S.-W. Fu, C.-F. Liao, E. N. Rastorgueva, F. Grondin, W. Aris, H. Na, Y. Gao, R. D. Mori, and Y. Bengio, "Speechbrain: A general-purpose speech toolkit," *ArXiv*, vol. abs/2106.04624, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:235377273

[20] L. Wan, Q. Wang, A. Papir, and I. L. Moreno, "Generalized end-to-end loss for speaker verification," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 4879–4883.

[21] S. C. Team, "Sasv2 baseline: Spoof-aware speaker verification benchmark," https://github.com/sasv-challenge/SASV2_Baseline, 2023, accessed: February 12, 2025.

[22] J. weon Jung, H. Tak, H. jin Shim, H.-S. Heo, B.-J. Lee, S.-W. Chung, H. jin Yu, N. W. D. Evans, and T. H. Kinnunen, "Sasv 2022: The first spoofing-aware speaker verification challenge," in *Interspeech*, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:247762170

[23] Y. Zhang, G. Zhu, and Z. Duan, "A probabilistic fusion framework for spoofing aware speaker verification," 02 2022.

[24] X. Wang, T. Kinnunen, K. A. Lee, P.-G. Noé, and J. Yamagishi, "Revisiting and improving scoring fusion for spoofing-aware speaker verification using compositional data analysis," 06 2024.

[25] S. H. Mun, H.-J. Shim, H. Tak, X. Wang, X. Liu, M. Sahidullah, M. Jeong, M. Han, M. Todisco, K. A. Lee, J. Yamagishi, N. Evans, T. Kinnunen, N. Kim, and J.-W. Jung, "Towards single integrated spoofing-aware speaker verification embeddings," 05 2023.

[26] S. K. Kopparapu, "A novel scheme to classify read and spontaneous speech," in *International Conference on Speech and Computer*, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:259164809

[27] R. Javali, S. Kanaki, S. Pujar, and S. Chikkamath, "Comparative evaluation of deep learning approaches in speaker recognition under varying speech styles: Read versus spontaneous," in *Proceedings of the International Conference on Recent Trends in Machine Learning, IoT, Smart Cities and Applications*. Singapore: Springer Nature Singapore, 2024, pp. 223–233.

[28] S. J. Park, G. Yeung, N. Vesselinova, J. Kreiman, P. A. Keating, and A. Alwan, "Towards understanding speaker discrimination abilities in humans and machines for text-independent short utterances of different speech styles," *The Journal of the Acoustical Society of America*, vol. 144, no. 1, pp. 375–386, 2018.

[29] A. Afshan and A. Alwan, "Learning from human perception to improve automatic speaker verification in style-mismatched conditions," in *Proceedings of Interspeech 2022*, 2022, pp. 2338–2342.

[30] N. Tomashenko, E. Vincent, and M. Tommasi, "Analysis of speech temporal dynamics in the context of speaker verification and voice anonymization," in *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, Apr. 2025, pp. 1–5.

[31] A. R. Bradlow, "ALLSSTAR: Archive of L1 and L2 Scripted and Spontaneous Transcripts And Recordings," retrieved from https://speechbox.linguistics.northwestern.edu/allsstar.

[32] M.-S. Han, "SKA-TDNN: Selective Kernel Attention TDNN for Speaker Verification," https://github.com/msh9184/ska-tdnn, 2022, accessed: December 12, 2024.

[33] L. Wan, Q. Wang, A. Papir, and I. L. Moreno, "Generalized end-to-end loss for speaker verification," 2020.