Exploiting Context-dependent Duration Features for Voice Anonymization Attack Systems

Natalia Tomashenko, Emmanuel Vincent, Marc Tommasi

▶ **To cite this version:**

Natalia Tomashenko, Emmanuel Vincent, Marc Tommasi. Exploiting Context-dependent Duration Features for Voice Anonymization Attack Systems. Interspeech 2025, Aug 2025, Rotterdam, Netherlands. ⟨hal-05099074⟩

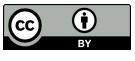
HAL Id: hal-05099074

https://hal.science/hal-05099074v1

Submitted on 5 Jun 2025

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons CC BY 4.0 - Attribution - International License

Exploiting Context-dependent Duration Features for Voice Anonymization Attack Systems

Natalia Tomashenko[1], *Emmanuel Vincent*[1], *Marc Tommasi*[2]

[1]Université de Lorraine, CNRS, Inria, LORIA, F-54000, Nancy, France
[2]Université de Lille, CNRS, Inria, Centrale Lille, UMR 9189 - CRIStAL, Lille, France
natalia.tomashenko@inria.fr, emmanuel.vincent@inria.fr, marc@tommasi@inria.fr

Abstract

The temporal dynamics of speech, encompassing variations in rhythm, intonation, and speaking rate, contain important and unique information about speaker identity. This paper proposes a new method for representing speaker characteristics by extracting context-dependent duration embeddings from speech temporal dynamics. We develop novel attack models using these representations and analyze the potential vulnerabilities in speaker verification and voice anonymization systems.The experimental results show that the developed attack models provide a significant improvement in speaker verification performance for both original and anonymized data in comparison with simpler representations of speech temporal dynamics reported in the literature.

Index Terms: automatic speaker verification, anonymization, attack model, context-dependent duration features.

1. Introduction

The widespread use of speech data today raises major privacy concerns, leading to its protection under the scope of privacy regulations like the European General Data Protection Regulation (GDPR) [1]. Voice recordings contain a large amount of personal information, and beyond revealing a speaker's identity, speech data can disclose other sensitive characteristics including age, gender, health conditions, emotional state, personality traits, ethnic background, and socioeconomic status. Understanding and protecting this vast amount of personal information is becoming increasingly important as voice technology becomes more prevalent in our daily lives.

A common method for preserving the privacy of speech data is voice anonymization which aims to suppress personally identifiable characteristics of the speaker, while preserving the linguistic and paralinguistic content [2]. Voice anonymization techniques have evolved significantly. They include two broad categories of methods.

The first category, signal processing-based methods, employs simple signal transformations to alter voice characteristics. These include spectral warping utilizing the McAdams coefficient [3], pitch shifting through time-scale modification [4], and others [5, 6], offering straightforward but limited anonymization capabilities.

The second and more complex category comprises neural voice conversion based methods [7–15], that operate by disentangling various speech attributes — including content, speaker characteristics, pitch, and emotion — before selectively anonymizing specific attributes and reconstructing the speech signal using speech synthesis models. Most state-of-the-art voice conversion based anonymization systems typically leverage large-scale pre-trained models for attribute extraction, demonstrating superior performance in both content preservation and privacy protection compared to signal processing-based methods.

While researchers have made significant advances in voice anonymization techniques and conducted several studies on the speaker information carried by pitch [6, 9, 10], little attention has been paid to the role of speech temporal dynamics in voice anonymization. Many state-of-the-art anonymization systems keep speech rate and phoneme durations unchanged [7, 9, 12, 16]. Cascaded automatic speech recognition (ASR) and text-to-speech (TTS) systems [17, 18] present rare exceptions, where word-level or phoneme-level transcripts from ASR are used by TTS to synthesize linguistic content with a new target voice. Although these systems likely avoid retaining speaker identity, they typically fail to preserve the paralinguistic attributes crucial for real-life voice anonymization applications.

Prior research relevant to our study [19] involves using speed perturbation with a constant factor as an anonymization technique, either independently or combined with anonymization based on a cycle consistent generative adversarial network (CycleGAN). These studies demonstrated that speech rate perturbation with a constant factor reduces the effectiveness of automatic speaker verification systems against both *ignorant* and *lazy-informed* attackers. However, they did not account for the more robust *semi-informed* attack model that is now the standard [20].

The studies in [21, 22] proposed speech rhythm-based speaker embeddings for duration modeling in multi-speaker speech synthesis. Another recent study [23], using a simple approach based on comparison of average phoneme duration vectors, has revealed the importance of speech temporal dynamics analysis in voice anonymization research.

This paper builds upon [22, 23] and makes a further step in this direction, performing a more systematic and advanced analysis of the speaker information conveyed in speech temporal dynamics. We consider two levels of information regarding speech temporal dynamics: statistics of average phoneme durations in utterances and complete set of phoneme durations in the given utterances. Based on the corresponding representations, we propose speaker verification models that can be used to attack voice anonymization systems. A key contribution is to propose a new method to encode the complete set of phoneme durations in the form of learned context-dependent duration embeddings and to show the strong improvement in attack performance achieved with this representation.

Section 2 describes the raw duration feature sequences and Section 3 explains how context-dependent duration embeddings can be extracted and used to construct attacks. Sections 4 and 5 describe the experimental setup and the results. Section 6 concludes the paper.

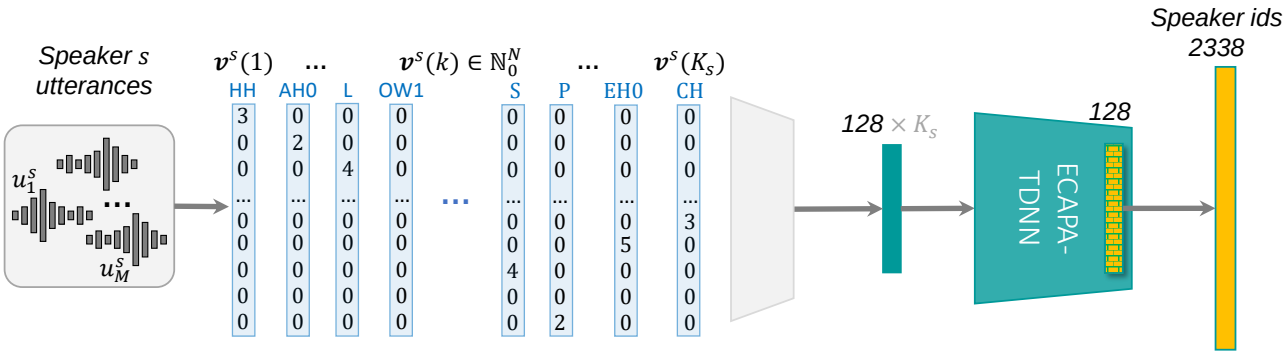


Figure 1: *Attack model training on raw duration feature sequences, yielding a 128-dimensional context-dependent duration embedding (yellow).*

2. Raw duration feature sequences

Previous works [23, 24] show that speaking rate and average durations of phonemes contain a significant amount of speaker information. In this study, our objective is to discover representations that more efficiently capture speaker information conveyed by speech temporal dynamics. We also aim to better understand how much speaker information can be retrieved from these different representations.

To do so, we consider raw duration feature sequences that are more informative than average phoneme durations and include all information about individual phoneme occurrences in the given utterance(s), including their context. Let us denote by N the number of phoneme classes $ph_1, \ldots, ph_N$, and by K_s the total number of phones $phone(1), \ldots, phone(K_s)$ in the considered utterance(s) of speaker s. For $k \in [1, K_s]$, each phone $phone(k)$ in the utterance(s) of speaker s is represented by a corresponding N-dimensional vector $\boldsymbol{v}^s(k) \in \mathbb{N}_0^N$: $\boldsymbol{v}^s(k) = [v_1^s(k), \ldots, v_N^s(k)]$ – a one-hot vector multiplied by the phone length (in acoustic frames) $len(phone(k))$, where the index i of a non-zero component $v_i^s(k)$ corresponds to the actual phoneme class ph_i: $phone(k) \in ph_i$. In other words,

$$v_i^s(k) = \begin{cases} len(phone(k)), phone(k) \in ph_i \\ 0, phone(k) \notin ph_i \end{cases} . \quad (1)$$

The phone lengths are extracted from the speech utterances using phonetic alignment.

3. Context-dependent duration embeddings and attack model

The raw duration feature sequences proposed in Section 2 allow us to build an efficient speaker verification model that learns a context-dependent duration embedding for the given utterance(s) of each speaker. This model is expected to be robust to modifications of the speech signal during anonymization other than changes in the temporal dynamics. Therefore, it can be used by attackers against voice anonymization.

The training process is illustrated in Figure 1. To reduce overfitting, instead of training the model on utterances, we train it on chunks of variable length ($32 - 256$ phones) with a random shift in the first utterance u_1 in every chunk, i.e. we remove a random number of phones r from the first utterance of each chunk. This random variable follows a uniform distribution $r \sim U(0, \min\{len(u_1), len(chunk)\})$, where $len(u_1)$ represents the number of phones in the first utterance u_1, and $len(chunk)$ is number of phones in the entire chunk. We adopt the standard ECAPA-TDNN (emphasized channel attention, propagation and aggregation in the time delay neural network) architecture [25] and use 128-dimensional projected versions of the raw duration feature vectors as inputs. A speaker embedding which encodes the context-dependent duration information is extracted from the last fully connected layer of the ECAPA-TDNN shown in yellow. The output layer performs speaker classification. The projection layer and the ECAPA-TDNN are jointly trained with a cross-entropy loss.

At test time, speaker embeddings are computed for the trial utterance(s) and the enrollment speaker using the trained network, and automatic speaker verification (ASV) scores are obtained using the cosine distance between embeddings.

4. Experimental setup

4.1. Data

Experiments were conducted on the *LibriSpeech*[1] [26] corpus of read English audiobooks. It contains approximately 1,000 hours of speech from 2,484 speakers sampled at 16 kHz. The training data is the *LibriSpeech-train-960* subset with 2,338 speakers.

For development and evaluation we used the *dev-clean* and *test-clean* subsets of *LibriSpeech*. For experiments on anonymized data, we used the *dev-clean* and *test-clean* subsets anonymized by a voice anonymization system as described in Section 4.2.

To perform phonetic alignment, two triphone Gaussian mixture model - hidden Markov model (GMM-HMM) acoustic models were trained using the Kaldi speech recognition toolkit [27] on the following training data: (1) original *LibriSpeech-train-960*; and (2) *LibriSpeech-train-clean-360* anonymized by the voice anonymization system. The second model was used to perform segmentation of the anonymized development and evaluation datasets.

In our experiments, we utilize exact text transcripts to obtain phonetic alignment. However, in real-world application scenarios, these alignments can be automatically obtained by ASR. The aim of this paper is to prove the importance of speech temporal dynamics for voice anonymization against a strong attacker, which is a necessary step before considering real-world application scenarios possibly involving weaker attacks. The motivation for using exact transcripts is therefore to eliminate,

[1] *LibriSpeech*: `http://www.openslr.org/12`

as much as possible, all factors that may negatively impact the attacker's performance. These include transcription errors caused by ASR.

4.2. Anonymization system

To investigate the impact of voice anonymization, as an example, we consider a state-of-the-art speaker voice anonymization system that keeps the original temporal phoneme dynamics unchanged, but modifies speaker identity and some prosodical characteristics such as pitch and energy. This voice anonymization system, proposed in [28] and used as baseline **B3** in the VoicePrivacy 2024 Challenge [29], is uses phonetic transcription and a generative adversarial network (GAN) that generates artificial pseudo-speaker embeddings. Anonymization is performed in three steps:

1. extraction of the speaker embedding, phonetic transcription, pitch, energy, and phone duration from the original audio waveform;
2. speaker embedding anonymization, pitch and energy modification;
3. synthesis of an anonymized speech waveform from the anonymized speaker embedding, modified pitch and energy features, original phonetic transcripts and original phone durations.

The automatic speaker verification results in terms of equal error rate (EER) on the *LibriSpeech* test set for anonymized data, according to [29], are around $27 - 28\%$ for the strongest *semi-informed* attacker trained on *utterance-level* anonymized data.

4.3. Attack model configuration

We perform experiments with the attack models described in Section 3. In all the reported experiments, we use the ARPAbet symbol set corresponding to the Carnegie Mellon University pronunciation dictionary[2] with $N = 336$ phoneme classes that take into account position in the word and stress.

The configuration and parameters of the ECAPA-TDNN are the same as those used in the *SpeechBrain* recipe[3]. The size of the speaker embedding extracted from the ECAPA-TDNN models and used for ASV score computation is 128. The model is trained using the cross-entropy criterion.

4.4. Baseline models

In the following, we use two models for comparison.

As the first baseline model, we consider the simple metric-based approach in [23]. We shortly review it in this section. For a speaker s with utterances $u_1^s, \ldots, u_M^s$, we define an N-dimensional vector $\boldsymbol{\mu}^s = [\mu_1^s, \ldots, \mu_N^s] \in \mathbb{R}_{\geq 0}^N$. This vector represents the average durations of phonemes $ph_1, \ldots, ph_N$ calculated across all utterances from speaker s. In cases where a particular phoneme is absent from the speaker's utterances, we fill in the vector with approximated values to ensure completeness. These approximations are derived from the mean duration of all phonemes present in the considered utterances for that speaker. Through this approach, any utterance or set of utterances from speaker s can be represented by the vector $\boldsymbol{\mu}^s$. The ASV score is computed using the metric proposed in [23]:

$$\rho(s_i, s_j) = 1 - \frac{1}{N} \sum_{n=1}^{N} \min \left\{ \frac{\mu_n^{s_i}}{\mu_n^{s_j}}, \frac{\mu_n^{s_j}}{\mu_n^{s_i}} \right\}, \qquad (2)$$

where $\mu_n^{s_i}$, $\mu_n^{s_j}$ are the n-th coordinates of mean duration vectors $\boldsymbol{\mu}^{s_i}$, $\boldsymbol{\mu}^{s_j}$ for speakers s_i, s_j.

As a second baseline model, we use an ASV model trained on anonymized data on conventional 80-dimensional filter bank features. This model is similar to the *semi-informed* attack model used in the VoicePrivacy 2024 Challenge evaluation setup [29]. The configuration and parameters of the ECAPA-TDNN are the same as those used in the *SpeechBrain VoxCeleb* recipe.

5. Results

The results on the original evaluation data for different attack models are summarized in Table 1 in terms of equal error rate (EER, %). The EER values are reported with 95% confidence intervals, calculated as suggested in [30]. In this table, the results are given for development and evaluation data, on the original unprocessed (the third and fourth columns) and anonymized (the last three columns) data for three different attack models:

1. **A1. metric** – the baseline metric-based approach [23] described in Section 4.4
2. **A2. duration features** – the proposed attack model relying on learned context-dependent duration embeddings
3. **A3. fbanks** (only for anonymized data) – a *semi-informed* attack model trained on anonymized data using conventional filter bank features.

It's important to note that models trained using only temporal dynamics information do not require retraining when used with anonymized data. This is one advantage of the proposed model with context-dependent duration embeddings which, unlike the conventional *semi-informed* attack model, is trained on original data and does not require retraining on anonymized data. Thus, a single universal attack model can be applied against different anonymization algorithms.

Table 1 shows the results for different numbers of utterances used to compute the speaker embeddings. In these experiments we use the same number of utterances for enrollment and trials and consider three different setups: 1, 4, and 8 utterances. The ASV performance improves significantly with increasing number of utterances, especially for the model using on context-dependent duration embeddings and for the *semi-informed* attack model trained on anonymized data on filter bank features. On original test data, for **A2**, the EER reduces from 24.8% to 2.8% when the number of utterances increases from 1 to 8.

The new attack model **A2** significantly outperforms the baseline metric-based attack system **A1**. On the original test data, the EER reduction for **A1** w.r.t. **A2** is $8 - 26\%$ absolute ($42 - 93\%$ relative), while on the anonymized test data the EER reduction is $12 - 22\%$ absolute and $28 - 85\%$ relative.

We can observe some degradation of the results for **A2** on anonymized data w.r.t. the results on original data in all cases, unlike for the baseline model **A1**, for which the results on the original and anonymized data are similar. However, **A2** substantially outperforms the baseline **A1** on the anonymized data as well.

Many automatic speaker verification systems typically utilize multiple utterances per speaker for enrollment and a sin-

[2] http://www.speech.cs.cmu.edu/cgi-bin/cmudict

[3] https://github.com/speechbrain/speechbrain/tree/develop/recipes/VoxCeleb/SpeakerRec

Table 1: *Automatic speaker verification results (EER,%) on original and anonymized data depending on the number of utterances (#utter) used to compute speaker embeddings, for three different attack models.*

		original data		anonymized data		
		A1	**A2**	**A1**	**A2**	**A3**
data	**# utter**	**metric**	**duration features**	**metric**	**duration features**	**fbanks**
dev-clean	1	38.2 ±0.7	24.8 ±0.6	41.7 ±0.7	32.4 ±0.6	25.9 ±0.6
	4	32.2 ±0.6	7.6 ±0.4	32.1 ±0.6	12.5 ±0.5	7.8 ±0.4
	8	25.9 ±0.6	2.8 ±0.2	26.1 ±0.6	6.0 ±0.3	3.2 ±0.2
test-clean	1	41.9 ±0.7	24.2 ±0.6	41.8 ±0.7	30.0 ±0.6	22.9 ±0.6
	4	32.2 ±0.6	5.9 ±0.3	31.8 ±0.6	10.4 ±0.4	5.4 ±0.3
	8	26.0 ±0.6	1.8 ±0.2	25.9 ±0.6	3.9 ±0.3	2.0 ±0.2

gle trial utterance for each test comparison. Table 2 demonstrates the performance of the models close to this scenario: for the **A2** model on original and anonymized data, and for **A3** on anonymized data. In these experiments, the number of enrollment utterances is either 8 or 16, and a single test trial utterance is used.

The results for both the **A2** and **A3** models demonstrate that we can achieve a relatively low EER with a small number of utterances for enrollment and trials.

Table 2: *Automatic speaker verification results (EER,%) on original and anonymized data for different attack models and varying number of utterances for enrollment (enr), with one utterance for trial (trl).*

		original data	anonymized data	
		A2	**A2**	**A3**
data	**# utter (enr,trl)**	**duration features**	**duration features**	**fbanks**
dev-clean	(8, 1)	15.3	21.5	17.1
	(16,1)	14.5	20.7	16.5
test-clean	(8, 1)	14.5	18.2	14.1
	(16,1)	14.0	16.6	13.1

6. Conclusions

Our introduction of context-dependent duration embeddings provides new possibilities for robust speaker verification and privacy-preserving voice technologies. The proposed context-dependent duration embeddings and attack model allowed us to significantly outperform the results reported in the literature for simpler representations of speech temporal dynamics [23], providing $8-26\%$ absolute ($42-93\%$ relative) EER reduction on original unprocessed data, and $12-22\%$ absolute ($28-85\%$ relative) EER reduction on anonymized data. This model offers a key advantage: unlike traditional *semi-informed* attack models [31], it can be trained directly on original data, eliminating the need for retraining on anonymized data. The results of this study can provide guidance for improving current voice anonymization algorithms.

Future research directions include extension of the study to other speech corpora, investigation of the complementarity of the proposed context-dependent duration embeddings and conventional speaker embeddings, and development of more advanced attack models and countermeasures that can conceal speaker temporal patterns effectively and withstand new attack scenarios.

7. Acknowledgements

This work was supported by the French National Research Agency under project Speech Privacy and project IPoP of the Cybersecurity PEPR. Experiments were carried out using the Grid'5000 testbed.

8. References

[1] A. Nautsch, C. Jasserand, E. Kindt, M. Todisco, I. Trancoso, and N. Evans, "The GDPR & speech data: Reflections of legal and technology communities, first steps towards a common understanding," in *Interspeech*, 2019, pp. 3695–3699.

[2] N. Tomashenko, B. M. L. Srivastava, X. Wang, E. Vincent, A. Nautsch, J. Yamagishi, N. Evans, J. Patino, J.-F. Bonastre, P.-G. Noé, and M. Todisco, "Introducing the VoicePrivacy Initiative," in *Interspeech*, 2020, pp. 1693–1697.

[3] J. Patino, N. Tomashenko, M. Todisco, A. Nautsch, and N. Evans, "Speaker anonymisation using the McAdams coefficient," in *Interspeech*, 2021, pp. 1099–1103.

[4] C. O. Mawalim, S. Okada, and M. Unoki, "Speaker anonymization by pitch shifting based on time-scale modification," in *2nd Symposium on Security and Privacy in Speech Communication*, 2022, pp. 35–42.

[5] P. Gupta, G. P. Prajapati, S. Singh, M. R. Kamble, and H. A. Patil, "Design of voice privacy system using linear prediction," in *2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2020, pp. 543–549.

[6] L. Tavi, T. Kinnunen, and R. G. Hautamäki, "Improving speaker de-identification with functional data analysis of f0 trajectories," *Speech Communication*, vol. 140, pp. 1–10, 2022.

[7] F. Fang, X. Wang, J. Yamagishi, I. Echizen, M. Todisco, N. Evans, and J.-F. Bonastre, "Speaker anonymization using x-vector and neural waveform models," in *Speech Synthesis Workshop*, 2019, pp. 155–160.

[8] X. Miao, X. Wang, E. Cooper, J. Yamagishi, and N. Tomashenko, "Speaker anonymization using orthogonal Householder neural network," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 3681–3695, 2023.

[9] B. M. L. Srivastava, M. Maouche, M. Sahidullah, E. Vincent, A. Bellet, M. Tommasi, N. Tomashenko, X. Wang, and J. Yamagishi, "Privacy and utility of x-vector based speaker anonymization," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 30, pp. 2383–2395, 2022.

[10] P. Champion, "Anonymizing speech: evaluating and designing speaker anonymization techniques," Ph.D. dissertation, Université de Lorraine, 2023.

[11] J. Yao, Q. Wang, P. Guo, Z. Ning, Y. Yang, Y. Pan, and L. Xie, "MUSA: Multi-lingual speaker anonymization via serial disentanglement," *arXiv preprint arXiv:2407.11629*, 2024.

[12] X. Miao, X. Wang, E. Cooper, J. Yamagishi, and N. Tomashenko, "Language-independent speaker anonymization approach using self-supervised pre-trained models," *arXiv preprint arXiv:2202.13097*, 2022.

[13] J. Yao, N. Kuzmin, Q. Wang, P. Guo, Z. Ning, D. Guo, K. A. Lee, E.-S. Chng, and L. Xie, "NPU-NTU system for Voice Privacy 2024 challenge," *arXiv preprint arXiv:2409.04173*, 2024.

[14] S. Saini and N. Saxena, "Speaker anonymity and voice conversion vulnerability: A speaker recognition analysis," in *2023 IEEE Conference on Communications and Network Security (CNS)*. IEEE, 2023, pp. 1–9.

[15] J. J. Webber, O. Watts, G. E. Henter, J. Williams, and S. King, "Voice conversion-based privacy through adversarial information hiding," *arXiv preprint arXiv:2409.14919*, 2024.

[16] M. Panariello, F. Nespoli, M. Todisco, and N. Evans, "Speaker anonymization using neural audio codec language models," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 4725–4729.

[17] Y. Sinha, J. Hintz, M. Busch, T. Polzehl, M. Haase, A. Wendemuth, and I. Siegert, "Why Eli Roth should not use TTS-systems for anonymization," in *Proceedings of the 2nd Symposium on Security and Privacy in Speech Communication*, 2022, pp. 17–22.

[18] H. L. Xinyuan, Z. Cai, A. Garg, K. Duh, L. P. García-Perera, S. Khudanpur, N. Andrews, and M. Wiesner, "Hltcoe jhu submission to the voice privacy challenge 2024," *arXiv preprint arXiv:2409.08913*, 2024.

[19] G. P. Prajapati, D. K. Singh, P. P. Amin, and H. A. Patil, "Voice privacy using CycleGAN and time-scale modification," *Computer Speech & Language*, vol. 74, 2022.

[20] N. Tomashenko, X. Miao, E. Vincent, and J. Yamagishi, "The first VoicePrivacy Attacker Challenge evaluation plan," *arXiv preprint arXiv:2410.07428*, 2024.

[21] K. Fujita, A. Ando, and Y. Ijima, "Phoneme duration modeling using speech rhythm-based speaker embeddings for multi-speaker speech synthesis," in *Interspeech*, 2021, pp. 3141–3145.

[22] ——, "Speech rhythm-based speaker embeddings extraction from phonemes and phoneme duration for multi-speaker speech synthesis," *IEICE Transcations on Information and Systems*, vol. 107, no. 1, pp. 93–104, 2024.

[23] 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, 2025, pp. 1–5.

[24] E. Bulgakova, A. Sholohov, N. Tomashenko, and Y. Matveev, "Speaker verification using spectral and durational segmental characteristics," in *17th International Conference on Speech and Computer*, 2015, pp. 397–404.

[25] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification," in *Interspeech*, 2020, pp. 3830–3834.

[26] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "LibriSpeech: an ASR corpus based on public domain audio books," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.

[27] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlíček *et al.*, "The Kaldi speech recognition toolkit," in *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2011.

[28] S. Meyer, F. Lux, J. Koch, P. Denisov, P. Tilli, and N. T. Vu, "Prosody is not identity: A speaker anonymization approach using prosody cloning," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[29] N. Tomashenko, X. Miao, P. Champion, S. Meyer, X. Wang, E. Vincent *et al.*, "The VoicePrivacy 2024 challenge evaluation plan," *arXiv preprint arXiv:2404.02677*, 2024.

[30] S. Bengio and J. Mariéthoz, "A statistical significance test for person authentication," in *Proceedings of Odyssey 2004: The Speaker and Language Recognition Workshop*, no. CONF, 2004.

[31] N. Tomashenko, X. Miao, E. Vincent, and J. Yamagishi, "The First VoicePrivacy Attacker Challenge," in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025, pp. 1–2.