5th Symposium on Security and Privacy in Speech Communication
16 August 2025, Delft, The Netherlands

The influence of non-timbral cues in voice anonymisation and evaluation

Rayane Bakari[1,2], *Olivier Le Blouch*[1], *Nicholas Evans*[2], *Nicolas Gengembre*[1], *Michele Panariello*[2], *Massimiliano Todisco*[2]

[1]Orange innovation, France
[2]Eurecom, Sophia Antipolis, France

{mohamedrayane.bakari, olivier.leblouch, nicolas.gengembre}@orange.com, {evans, michele.panariello, todisco}@eurecom.fr

Abstract

Most approaches to voice anonymisation focus predominantly upon the obfuscation of timbral attributes. Approaches to evaluation which use traditional automatic speaker verification (ASV) systems such as ECAPA-TDNN can result in the overestimation of anonymisation performance since they too focus on timbral cues. In this paper, we show that the use of residual non-timbral attributes, e.g. related to prosody, rhythm, style and accent which also carry information related to the voice identity, can still be used to re-identify the speaker. When timbral cues are compromised, non-timbral cues can provide more reliable estimates of anonymisation performance. We also show that, when trained to focus on non-timbral attributes, a WavLM-based model outperforms the baseline ECAPA-TDNN model when operating upon anonymised speech. Using the latter, the equal error rate for the best 2024 VoicePrivacy Challenge baseline is overestimated by 32% relative. Ultimately, we hope to provide a fresh perspective, laying the foundation for more robust and comprehensive evaluations of voice anonymisation and highlighting the importance to future anonymisation systems of obfuscating non-timbral information.

Index Terms: voice privacy, anonymisation, speech disentanglement, non-timbral, voice attributes

1. Introduction

Given that voice recordings contain sensitive information which can be used to identify individuals or infer personal traits [1], the rapid growth of voice-activated services has amplified privacy concerns. Voice anonymisation techniques have hence been developed to sanitise speech recordings of the voice identity thereby preventing remaining sensitive content from being linked to the original speaker. At the same time, specific voice services and applications demand the preservation of other linguistic and paralinguistic attributes [2].

Most anonymisation systems are based upon the use of distinct representations of speech [2], one usually a form of speaker or voice representation and at least one more used as an auxiliary representation of other speech attributes, e.g. the linguistic content, prosody, rhythm, accent etc. The general approach to anonymisation then involves the manipulation or substitution of the speaker representation followed by its recombination with the auxiliary representation to generate a voice-anonymised output.

The common underlying assumption is that voice identity information is contained predominantly in the speaker representation. Upon its substitution with the representation of another speaker, the output speech signal should be fully sanitised of any cues which could be utilised to re-identify the original speaker. If this was the case, then such approaches would result in perfect voice anonymisation. In reality, however, some voice identity information resides also in the auxiliary representation. Anonymisation performance is then suboptimal and residual voice information can still be used to re-identify the original speaker, albeit to a lesser degree.

What precisely constitutes residual voice information is difficult to determine. If one assumes that the dominant source of voice information in the speaker representation is related to timbre, it is not a stretch to assume that voice information contained in the auxiliary representation is predominantly non-timbral, e.g. prosody, rhythm, style and accent, etc. Setting aside for now the debatable distinction between the respective timbral and non-timbral content of speaker and auxiliary representations [3, 4, 5], the key to improving anonymisation performance lies in the sanitisation of voice information contained in the auxiliary representation. With this longer-term goal in mind, we have hence set out to determine the degree to which non-timbral and other sources of voice information can be used to re-identify the original speaker post-anonymisation and what is the potential to improve anonymisation performance through their obfuscation.

We propose the use of variants of the popular ECAPA-TDNN [6] and WavLM [7] automatic speaker verification (ASV) models which are retrained specifically to focus upon the use of non-timbral information and report their use for the evaluation of anonymisation performance. Our study explicitly focuses on disentangled embeddings in order to isolate and analyse the contribution of non-timbral cues to speaker identity. We show that current approaches to evaluation exaggerate anonymisation performance and that the degree of voice identity information residing in non-timbral cues is considerably higher than some might think. Our retrained systems designed to focus on non-timbral information can be used to re-identify the speaker when the usual timbral cues are unreliable. The findings provide new insights for future research in anonymisation in addition to evaluation.

The main contributions of this work are as follows.

- We demonstrate that non-timbral cues (e.g., prosody, rhythm, speaking style, and accent) can still be used to re-identify speakers after anonymisation.
- We propose ASV models retrained specifically to focus on non-timbral cues.
- We show that these models outperform baseline ECAPA-TDNN systems in all attack scenarios, revealing weaknesses in existing anonymisation evaluations.
- Our WavLM-based systems outperform the top VoicePrivacy Attacker Challenge systems on all anonymisation systems except one.
- We highlight the importance of obfuscating non-timbral cues

10.21437/SPSC.2025-6

in future anonymisation systems and advocate for the consideration of non-timbral cues in future evaluations.

The remainder of this paper is organised as follows. In Section 2 we provide a review of relevant work in the literature and context for our research. In Section 3 we describe the design of ASV systems which focus upon the use of non-timbral cues to infer the voice identity and in Section 4, the experimental setup. We present results for some anonymisation systems in Section 5 and an analysis of the findings and implications in Section 6. Conclusions and the directions of our ongoing work are presented in Section 7.

2. Related Work

2.1. VoicePrivacy Challenge and Evaluation

The bulk of related work has been performed within the scope of the VoicePrivacy initiative founded in 2020 [8]. The first challenge was held the same year [9], with subsequent editions being held biennially since. Six baseline anonymisation systems were made available to the community for the most recent, third edition, with three among them (B3-B5) being competitive with the state of the art at the time. Baseline B3, a combination of automatic speech recognition (ASR) and text-to-speech (TTS) synthesis, uses a generative adversarial network (GAN) to generate pseudonymised speaker representations in the form of speaker embeddings [10]. B4 leverages a neural audio codec to obfuscate the voice identity [11], while B5 incorporates vector quantization to enhance the disentanglement of linguistic and voice representations [12].

Through the VoicePrivacy Challenge (VPC), different metrics have been proposed for evaluation [2]. The *word error rate* (WER) and the *unweighted average recall* (UAR) are used to evaluate utility preservation, reflecting respectively the preservation of linguistic content and emotional state. That for assessing privacy, which is the primary consideration in our work, is the *equal error rate* (EER) which reflects the potential to re-identify the original speaker using an ASV system. The default evaluation scenario involves the so-called *semi-informed* attack model: in order to reduce domain-mismatch, and hence to strengthen the attack, the ASV system used for EER estimation is itself trained using data which is anonymised using the same anonymisation system under test [2].

2.2. Limitations of Existing Evaluation and ASV Models

Even if current anonymisation systems achieve high EERs, it is acknowledged in the community that EER estimates are themselves of questionable reliability. In the words of Bäckström [13], '*privacy* (here, anonymisation) *is only as good as the adversarial model*'. Use of a weak adversarial/attack model will result in an exaggerated estimate of anonymisation performance; reliable estimates can only be made using the strongest possible model [14].

The ASV system provided by the VPC organisers is based upon the adaptation of the SpeechBrain [15] implementation of the long-established ECAPA-TDNN model [6]. Based on time-delay neural networks and channel attention mechanisms, it is used for the extraction of robust speaker embeddings which are scored using cosine similarity. It is however acknowledged that the ECAPA-TDNN model is not the strongest possible [7]. Even though it is retrained using anonymised data, the architecture was designed to operate upon unprotected speech data, not data where the usual cues might be compromised. An adversary seeking to re-identify the original speaker post-anonymisation might fare well to use alternative cues, e.g. non-timbral cues and those extracted using a stronger ASV model.

WavLM [7] is one such stronger model. Built on a self-supervised learning paradigm, WavLM is built on a Transformer backbone, making it particularly interesting for analyzing non-timbral characteristics in speaker privacy evaluations. Thanks to its architecture, WavLM effectively captures long-range dependencies, contextual interactions, and non-timbral features such as prosody, accent, and speaking style, all attributes of the voice identity [16]. Notably, it has been shown to outperform ECAPA-TDNN using benchmarks such as the VoxCeleb1 database [7] and is hence a stronger candidate ASV system in our work for the evaluation of anonymisation performance.

2.3. Disentanglement and Vulnerabilities in Anonymisation

Previous studies have revealed the vulnerabilities in anonymisation systems, especially through analysis of disentangled speech representations. These representations separate speaker identity from other speech aspects, allowing targeted manipulation or analysis. Early foundational work by Williams and King [17] introduced methods to disentangle style factors from speaker embeddings, enabling more interpretable representations. This was extended by Williams et al. [18] using a VQ-VAE to separate phone and speaker information. These studies suggest that disentangled speech representations could be used for the re-identification of anonymised speakers. While disentanglement remains challenging [19], [20] introduced the Binary-Attribute-based LR estimation approach (BA-LR) to disentangle speaker voice characteristic into so-called BA-vectors, a vector of n binary attributes. Voice conversion systems such as SpeechSplit [21] and AutoVC [22] enable aspect-specific voice manipulation by disentangling speech into components such as content, rhythm, pitch, and timbre. Other works focus on reversibility or identity leakage. Champion et al. [23] demonstrated the vulnerabilities of voice anonymisation systems using alignment techniques like Wasserstein-Procrustes analysis to match x-vectors [24] before and after anonymisation. Panariello et al. [25] highlighted weaknesses stemming from vocoder drift, which refers to the substantial difference between an input x-vector and the x-vector extracted from the vocoder output, indicating that even pseudo-anonymised x-vectors can leak voice information.

2.4. VoicePrivacy Attacker Challenge

The VoicePrivacy Attacker Challenge [26] was launched in late 2024 to foster further progress in attack models and evaluation. The objective was to develop stronger attacker systems against voice anonymisation techniques as a means to improve evaluation reliability. A summary of these systems is presented in [27]. A set of anonymisation systems were selected as attack targets including the three competitive baselines (B3, B4, B5) in addition to four systems developed by the VoicePrivacy 2024 Challenge participants (T8-5, T10-2, T12-5, and T25-1).

Xinyuan et al. [28] developed the T8-5 system, an admixture based technique enabling flexible adjustments to the privacy-utility trade-off, also achieving an EER of over 40%. Yao et al. [29] proposed the T10-2, a serial distillation approach to process distinct representations of spoken content, voice, and emotion, achieving an EER exceeding 40%. T12-5 [30] is based on B5, with additional pitch smoothing. T25-1 [31] proposed a disentanglement of content (VQ-BN as in B5) and style (global style token - GST) features and emotion transfer from target speaker utterances. Despite these impressive results, EERs re-

main below 50%, the target which would suggest that no voice-dependent information remains [11, 12, 29].

Diverse attacker systems emerged from the VoicePrivacy Attacker Challenge. One of the best attacker systems proposed by Zhang et al. [32] combines data augmentation enhanced feature representations and a speaker identity difference enhanced classifier to improve ASV performance. They showed that the EER for the best baseline (B5) falls from 34% to 27% for LibriSpeech test data, a reduction of 21% relative. Another top-performing system proposed by Lyu et al. [33] adapts a pre-trained *ResNet34* ASV model [34] with a *LoRA* technique [35] using anonymised data. This approach reduces the EER for B5 by 26% relative. Li et al. [36] proposed SpecWav-Attack, a system which leverages *wav2vec2.0* for feature extraction, integrates spectrogram resizing and incremental training to improve attack performance. The approach proposed by Mawalim et al. [37] is based upon the fine-tuning of the *TitaNet-Large* model [38] with anonymised data achieves a reduction of 21% relative EER for B5. [39] used alternative distance metrics and voice kNN-VC-based voice normalisation to attack anonymisation sytems. Tomashenko et al. demonstrate the potential of leveraging phoneme duration, a non-timbral attribute estimated from the analysis of speech temporal dynamics to re-identify anonymised speakers [40].

2.5. Limitations and Gaps Addressed in This Work

In summary, current anonymisation evaluation relies on ASV models not optimally suited for protected data, and often underestimates residual speaker information embedded in non-timbral cues. Although recent attacker systems have demonstrated substantial progress, few explicitly target disentangled representations or non-timbral characteristics. Our work builds on these insights by using stronger ASV systems retrained to exploit such cues, thereby providing a more realistic evaluation of anonymisation robustness.

3. Influence of non-timbral cues

In this section we describe a set of ASV systems which focus by design on *global* speaker characteristics, i.e. entangled timbral and non-timbral attributes, and their adaptation to focus more specifically on non-timbral cues. We also describe the experimental setup and the three different attack models used in our experiments.

3.1. ASV baselines

To enable comparisons to other results reported in the literature, we use the reference ECAPA-TDNN system provided by the VPC organisers. It is henceforth referred to as *E-VPC*. We also trained two additional ASV baselines, *E-SPK* and *W-SPK* based respectively on ECAPA-TDNN and WavLM architectures. The suffix *SPK* denotes systems which model global *speaker* characteristics. In contrast to *E-VPC*, which is trained using the LibriSpeech train-clean-360 (LS360) dataset, *E-SPK* is trained using the Voxceleb1&2 datasets and with the modest adjustments to the training parameters described in Section 4. Drawing inspiration from [16], we also explored the potential of the WavLM model [7]. Its use *is* authorised in the evaluation plans for both the 2024 VoicePrivacy Challenge [2] and the 2024 VoicePrivacy Attacker Challenge [26]. These baselines provide reference points to evaluate the added value of focusing on non-timbral cues.

3.2. Adaptation to non-timbral characteristics

We adapted both ASV baselines to focus on the use of non-timbral (NT) characteristics giving *E-NT*, based on ECAPA-TDNN and *W-NT*, based on WavLM. Both systems are trained using datasets converted by a voice conversion (VC) process in order to mask the timbral characteristics of the original speaker while preserving, to varying degrees, non-timbral characteristics such as prosody, rhythm, style, and accent. The VC process is implemented using the Retrieval-based Voice Conversion (RVC) framework[1] depicted in Figure 1. This framework is designed to retain the temporal structure of the source speech: by using SSL features that preserve the number of frames and a training process that minimises misalignment, the output signal is naturally synchronised with the input. This alignment ensures that non-timbral aspects such as rhythm and speaking rate are preserved in the converted signal. To visually confirm this synchronisation, Figure 2 shows an example of voice conversion, where a source utterance from VoxCeleb2 is converted to a target speaker identity from the LS360 dataset. The spectrograms of the source and converted waveforms are closely aligned in time, demonstrating the preservation of temporal structure.

Training is nonetheless performed using original speaker labels. *NT* variants hence provide a controlled environment to evaluate the influence of non-timbral characteristics.

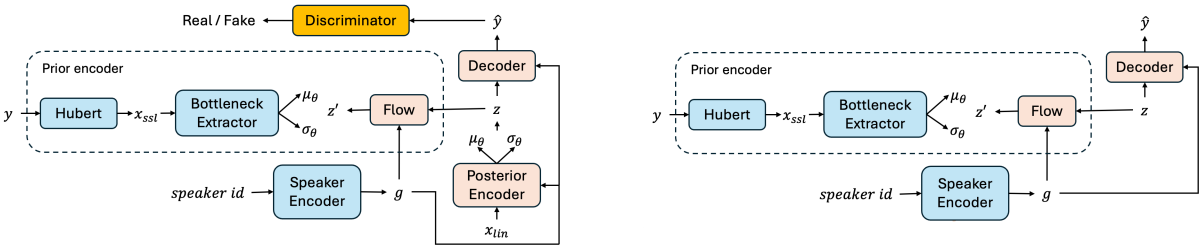
3.3. Attack models

We report experiments performed using the usual three VPC-defined attack models described in [2, 9, 41], also illustrated in Figure 3. All three models involve comparisons between a pair of utterances, an original or anonymised enrolment utterance and an anonymised trial utterance. Anonymisation performance is evaluated in terms of the ASV EER.

For the *ignorant* (**I**) attack model [9], the adversary compares original enrolment and anonymised trial utterances without any compensation for anonymisation. Under the *lazy-informed* (**L**) attack model, the adversary makes a nominal effort to compensate for the use of anonymisation. To reduce domain mismatch, the adversary anonymises the enrolment utterance so that comparisons are made between anonymised enrolment and anonymised trial utterances. The adversary is assumed to have access to the same anonymisation system, but not the specific configuration [2]. The third, strongest model is the *semi-informed* (**S**) attack. In an effort to further reduce domain mismatch, the adversary now uses the same anonymisation system to produce anonymised data with which to retrain the ASV system, referred as $ASV_{\text{eval}}^{\text{anon}}$ in [2].

Results for **I** and **L** attack models are included to show the benefit to the adversary of focusing *specifically* on non-timbral cues, features that remain after anonymisation removes identifiable timbral information. These two attack models are particularly informative because they test model robustness without requiring retraining on anonymised data, unlike the semi-informed **S** model. In contrast, the results for the semi-informed model show the extent to which informative cues (including non-timbral) can be learned automatically when the $ASV_{\text{eval}}^{\text{anon}}$ system is retrained using anonymised data. These models help to reveal the extent to which non-timbral cues remain exploitable after anonymisation.

[1] We used the implementation available at `https://github.com/RVC-Project`



(a) *Training procedure*

(b) *Inference procedure*

Figure 1: *An illustration of the RVC training and inference procedures where y denotes the input source waveform, $\hat{y}$ denotes the output converted waveform, x_{lin} denotes a linear spectrogram, x_{ssl} denotes SSL features, and g denotes a speaker embedding. By design, RVC preserves the number of SSL frames and produces a converted waveform which is perfectly synchronised with the source waveform.*

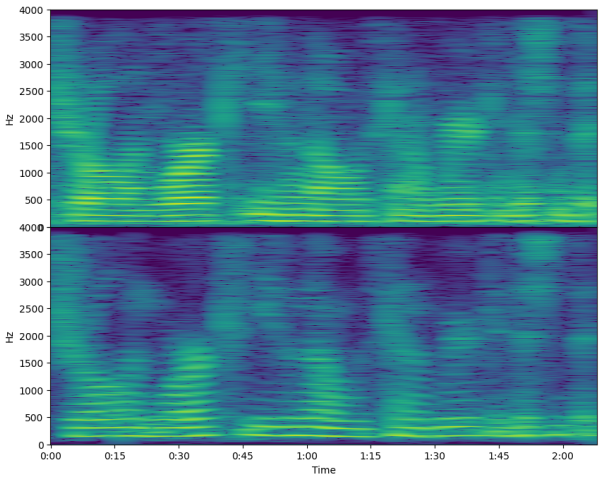


Figure 2: *Example of voice conversion using RVC. Top: source audio extracted from VoxCeleb2, id00012. Bottom: source audio converted to the LS360 "7416" female voice. The signals are perfectly temporally synchronized.*

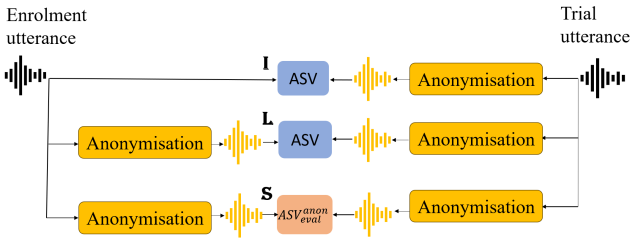


Figure 3: *Privacy evaluation in ignorant (I), lazy-informed (L) and semi-informed (S) attack models. In the S attack model, the attacker uses a retrained ASV system, ASV^{anon}_{eval} which is fine-tuned on anonymised data.*

4. Experimental setup

RVC is trained using data collected from the first 99 speakers of the VCTK dataset [42]. Data for the remaining 10 speakers was set aside for testing as in [16]. VoxCeleb1&2 datasets were used as source data for RVC-based conversions. Each utterance is converted to a target voice chosen at random from among the 99 VCTK speakers, thereby resulting in *converted VoxCeleb1&2* datasets with +1M utterances. Figure 4 illustrates our full experimental pipeline, covering both the RVC-based data conversion and the training of the speaker verification systems. To demonstrate the effect of domain adaptation, we generated *converted LS360* in identical fashion by applying RVC to the LibriSpeech train-clean-360 dataset resulting in 104k utterances. The motivation here is to evaluate the performance of our systems on a different domain, and whether retraining on domain-matched data improves performance.

Having noticed that *E-VPC* does not result in proper model convergence when trained according to the VPC pipeline,[2] and to ensure a fair comparison with our other approaches, *E-SPK* and *E-NT* benefit from modest adjustments to the training parameters. The initial learning rate was reduced from 1e-2 to 1e-3 while the number of training epochs was increased from 10 to 20. The ECAPA-TDNN architecture used for the 2024 VPC and in our experiments is a light version with 512-channel layers. During the training of *W-NT* and *W-SPK*, speaker verification task is performed with a module built on top of WavLM that classifies the speaker identities. This module is composed of two linear blocks with rectified linear unit activations and batch normalization (features sizes are 378 for the input, then 512 and 250, and 7323 output classes)

For I and L attack models, the reference *E-VPC* model and the *E-SPK* model are used whereas *E-NT* and *E-NT-360* are trained respectively using the converted Voxceleb1&2 datasets and converted LS360 datasets respectively for 10 epochs. Following [16], *W-SPK*, *W-NT* and *W-NT-360* are fine-tuned from WavLM-Large[3] using original Voxceleb1&2, converted Voxceleb1&2 and converted LS360 respectively for 6 epochs.

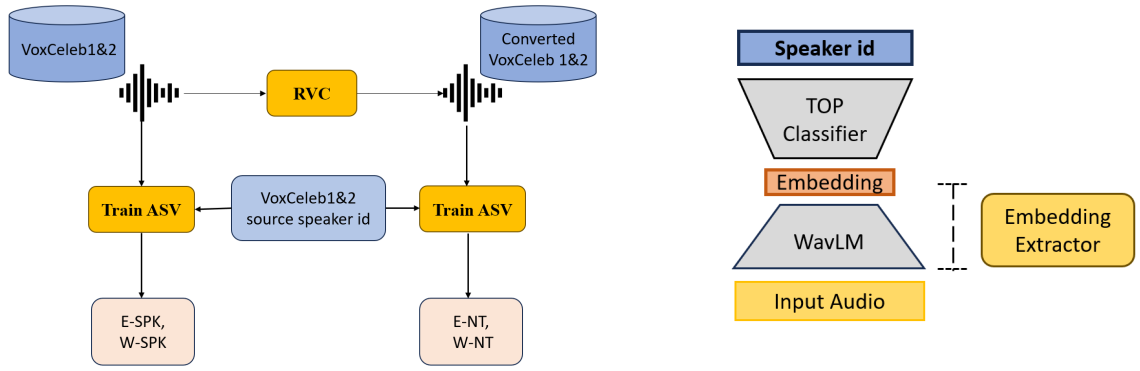
For the semi-informed attack model, and for each anonymisation model, ASV^{anon}_{eval} variants of *W-NT*, *W-SPK*, *E-VPC*, *E-NT* and *E-SPK* are re-trained or fine-tuned for 20 epochs from pre-trained versions using the respective *anonymised LS360* dataset.

5. Results

Results for the VPC baseline systems and the best VPC 2024 submissions described in Section 2 are presented in Table 1. A lower EER indicates greater vulnerability of the anonymisa-

[2] https://github.com/Voice-Privacy-Challenge/Voice-Privacy-Challenge-2024

[3] We used the publicly model available at https://huggingface.co/microsoft/wavlm-large



(a) *Speaker Verification Training Pipeline*

(b) *Embedding extraction used for Speaker Classification*

Figure 4: *Overview of the experimental pipeline. Original VoxCeleb1&2 utterances are converted using Retrieval-based Voice Conversion (RVC), resulting in datasets that preserve, to some extent, non-timbral speaker cues. Speaker verification (ASV) models are trained on both original and converted datasets to assess the impact of non-timbral cues.*

tion system to speaker re-identification, meaning the anonymisation is less effective. This table is used to demonstrate how effectively the new non-timbral representations can compromise voice privacy. Comparisons are made to highlight: the difference in performance for ECAPA and WavLM-based ASV systems; the influence of non-timbral adaptations; the influence of WavLM in capturing non-timbral cues.

5.1. Comparison of ASV models

The *E-SPK* system performs consistently better than *E-VPC* for the S attack model, suggesting that *E-VPC* is suboptimal. *W-SPK* performs better than *E-SPK* for both L and S attacks and hence provides a substantially stronger attack. For the **S** attack, the EER for *E-VPC* and B5 drops from 34% to 23% using *W-SPK*, corresponding to an overestimate of 32% relative. This substantial reduction shows that stronger ASV models offer greater potential to capture residual voice characteristics post-anonymisation, that estimates of performance can be overestimated easily and even the use of a stronger model can reveal weaknesses in anonymisation. Denoising and pre-training using a considerably larger quantity of data supports the extraction of more discriminative representations, leaving WavLM-based methods better equipped to handle the variability introduced by anonymisation and a stronger candidate for evaluation.

5.2. Impact of non-timbral cues on re-identification

Results for *E-SPK* under the **S** attack model are better than those for *E-VPC*, published in [2] on account of the training optimisations described in Section 4. For the **L** attack and for all anonymisation systems, EERs for *E-NT* are lower than those for the *E-SPK* system. For example, the EER of 48% for the B4 baseline (near-to-perfect anonymisation) drops to 41%. These results show that, when timbral information is compromised, the same system trained to use non-timbral information achieves better and more reliable estimates of performance.

Results for the *W-NT* system shown in the penultimate block of Table 1 show that, for the **L** attack, *W-NT* achieves lower EERs than *W-SPK* for all models, echoing the same finding for *E-NT* and *E-SPK* models. For the B4 baseline, for example, the EER of 45% for *W-SPK* drops to 32% for the *W-NT* system.

However, there is less of a difference between the EERs for the **S** attack when comparing *E-SPK* and *E-NT* or W-SPK and W-NT. This can be attributed to the fine-tuning of each ASV model using anonymised data which serves to reduce the differences between cues leveraged by each system. Fine-tuning equips both models with similar capabilities to distinguish between speakers, thereby reducing the advantage of *W-NT* in this context. *E-NT-360* and *W-NT-360* outperform *E-NT* and *W-NT* systems respectively for both **I** and **L** attacks, indicating the benefit of domain adaptation to LibriSpeech data.

5.3. Influence of WavLM in capturing non-timbral cues

While *E-NT* and *W-NT* models are both designed to focus on non-timbral characteristics, their architectures are different. Their use of non-timbral characteristics results in strong performance for both **L** and **S** attacks. By concentrating on non-timbral characteristics, they become less reliant on timbral cues, which enhances performance when these same cues are compromised due to anonymisation. *W-NT* consistently outperforms *E-NT*, supporting our intuition that the transformer-based architecture of *W-NT* is better able to capture and utilise non-timbral cues when comparing original and anonymised data. Its ability to capture subtle characteristic shifts post-anonymisation provides an advantage over the ECAPA-TDNN architecture.

5.4. Comparison with top VoicePrivacy attacker systems

We present the performance of our systems in the same style proposed for the VoicePrivacy Attacker Challenge [26]. Figure 5 illustrates how our systems, especially, *W-SPK* and *W-NT* outperform the top two submissions to the VoicePrivacy Attacker Challenge, namely the systems by Zhang et al. [32] and Lyu et al. [33], for each anonymisation systems. Both of our models achieve low EERs, outperforming both competing systems by a significant margin. Notably, *W-SPK* obtains the lowest EERs for all anonymisation systems except for T10-2. We further discuss this finding in Section 6. These results highlight the importance to anonymisation of obfuscating not just timbral attributes, but also non-timbral attributes.

Table 1: *EERs (%) for ignorant (I), lazy-informed (L) and semi-informed (S) attack models for the LibriSpeech test subset and for each anonymisation baseline and each ASV system. Un. denotes EERs for unprotected data. Best results for S and I&L are in bold.*

	E-VPC			**E-SPK**			**E-NT**			**E-NT-360**		**W-SPK**			**W-NT**			**W-NT-360**	
Un.	4.6			0.3			15.0			9.1		0.4			7.8			6.8	
	I	L	S	I	L	S	I	L	S	I	L	I	L	S	I	L	S	I	L
B3	47.4	45.7	27.3	48.3	47.5	26.2	42.9	42.7	26.4	40.4	36.5	47.6	44.8	17.5	38.2	34.7	**17.4**	35.0	**28.1**
B4	47.8	49.5	30.3	47.8	47.9	27.1	40.7	40.9	28.0	38.2	38.0	44.6	44.5	**14.5**	34.2	32.0	14.8	**28.1**	30.7
B5	49.1	48.7	34.3	49.8	49.7	31.6	47.2	46.7	31.2	45.2	43.5	48.8	48.7	**22.5**	42.5	42.0	23.2	38.3	**38.3**
T8-5	45.5	48.2	40.9	42.4	47.1	41.1	39.1	43.2	41.9	38.1	43.9	41.7	45.2	**22.8**	32.8	36.3	23	**31.9**	34.7
T10-2	36.2	35.9	40.8	36.8	38.4	35.7	31.3	31.6	38.2	23.2	25.3	32.8	34.7	32.2	23.6	22.1	**29.2**	14.5	**14.4**
T12-5	49.1	51.1	33.2	49.5	50.2	33.8	48.3	48.5	32.5	45.8	44.7	47.1	48.6	**23.9**	44.4	43.2	**23.9**	40.1	**37.2**
T25-1	48.8	49.5	39.8	49.9	49.5	35.8	47.1	48.2	36.8	45.4	43.6	48.2	48.9	**24.1**	44.7	44.1	27.9	40.4	**37.6**

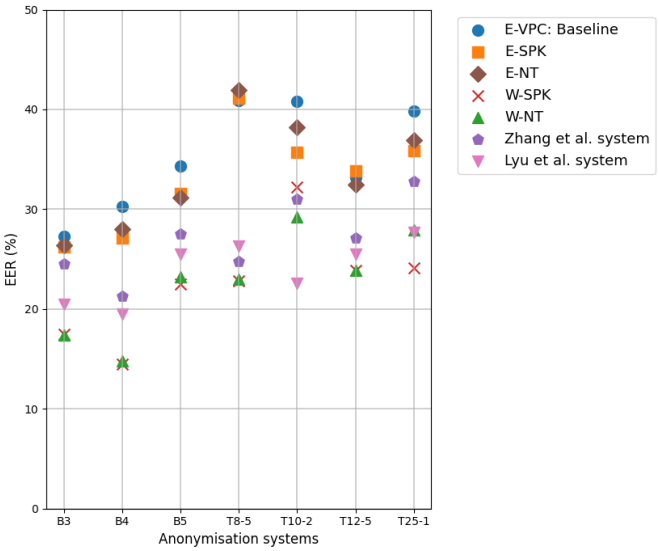


Figure 5: *EERs for the semi-informed attack model (S) for the LibriSpeech test. Results shown for our systems and the two best systems from the VoicePrivacy Attacker Challenge [27].*

6. Discussion

This study sheds new light on the role of timbral and non-timbral cues in ASV and their influence in voice anonymisation and evaluation. Our findings indicate that non-timbral cues can survive anonymisation processes, especially when using certain ASV models. This suggests that current anonymisation methods may need to incorporate strategies specifically targeting these cues. For all ASV models, the focus on non-timbral cues, even under an ignorant attack model, leads to lower EERs than global cues under a stronger lazy-informed attack model. When a stronger ASV system is fine-tuned using anonymised data under the strongest, semi-informed attack model, the advantage of systems trained to focus on non-timbral characteristics is diminished suggesting that fine-tuning using anonymised data is somewhat effective in adapting to use of the same non-timbral cues.

These findings carry important implications for the development and evaluation of voice anonymisation systems. Most anonymisation systems do not fully protect the voice identity since residual, non-timbral characteristics such as prosody, speaking style and accent, remain critical identifiers, even after anonymisation. Many previously reported estimates of anonymisation performance might not be fully trustworthy, adding further weight to the arguments within the community to consider stronger approaches to evaluation. Similar concerns were raised by Williams et al. [43], who showed that anonymised voices, while harder for ASV systems to recognise, remain perceptually easy to imitate, further highlighting unresolved privacy risks and the need for stronger anonymisation strategies.

We acknowledge that our findings are entangled implicitly in results for the VPC semi-informed attack model, yet we make them more explicit. The retraining of $ASV_{\text{eval}}^{\text{anon}}$ systems for the semi-informed attack model is not dissimilar to the protocol used in our work — the obfuscation of timbral cues via anonymisation and then the refocusing on what remains. Our use of the RVC system is akin to anonymisation. By using RVC-generated data to train an $ASV_{\text{eval}}^{\text{anon}}$ system, we better isolate and can better observe the difference between what is *domain adaptation* and what is *instead* the use of *non-timbral information*. The E-NT-360 system is equivalent to an $ASV_{\text{eval}}^{\text{anon}}$ system which is domain-adapted to data generated using the RVC-based anonymisation system. Its use for the evaluation of other anonymisation systems for the ignorant and lazy-informed attack models then reveals the real influence of non-timbral cues in contrast to what is purely domain adaptation (results for semi-informed attacks).

Our *W-NT* and *W-SPK* systems outperform Lyu et al.'s approach [33] for all anonymisation systems except T10-2. T10-2 employs a serial disentanglement strategy to gradually disentangle the global speaker identity and time-variant linguistic content and paralinguistic information. We hypothesise that this system is particularly effective at obfuscating non-timbral attributes, the same attributes upon which our systems are trained to focus. Interestingly, Lyu et al.'s attack system also uses WavLM as a feature extractor, integrating LoRA modules into a pretrained ResNet34 ASV model, allowing fine-grained adaptation to anonymised data distributions without affecting ASV performance on unprotected audio. The architecture and training protocol likely contribute to its superior performance for T10-2, despite the stronger generalisation of our models across other anonymisation systems.

We also acknowledge some limitations of the study reported in this paper. While the *E-NT* and W-NT systems are trained to focus on residual, non-timbral cues, we cannot be certain that this is *all* they use. Some timbral information corresponding to the original speaker may remain in RVC-converted data used for the training of *E-NT* and *W-NT* systems. The similarity in results for the semi-informed attack model does

nonetheless suggest that NT-variant systems learn the same, or similar cues when they are trained using anonymised data. While it is likely that these cues are of non-timbral origins, confirmation demands further investigation to better understand the distinction and interplay between timbral and non-timbral characteristics and their contrasting prevalence in both speaker and auxiliary representations used for anonymisation.

While we use the term "non-timbral" to describe residual speaker characteristics such as prosody, rhythm, accent, and speaking style, we acknowledge that the boundary between timbral and non-timbral attributes is not clean-cut. In practice, speaker identity information is distributed across multiple entangled acoustic dimensions. Our methodology relies on the assumption that RVC-based voice conversion suppresses most timbral cues while preserving non-timbral traits. However, we are currently unable to quantify the residual speaker information carried by each specific non-timbral attribute such as rhythm, prosody, accent, or speaking style. The extent to which each of these characteristics contributes to re-identification remains an open question, and is the subject of ongoing investigation. Early work by Tomashenko et al. [40] has started to investigate the role of phoneme duration, a non-timbral attribute in speaker re-identification. These findings reinforce the relevance of non-timbral cues, but further analysis is needed to disentangle their individual contributions and interactions, especially under different anonymisation strategies.

What is clear from the results reported in this paper and elsewhere is that the key to improving anonymisation performance lies in the sanitisation of speaker-dependent cues contained within the auxiliary representations, *whatever* their nature. This will be difficult given that F0, energy, phone duration and even phonetic transcripts, etc., may preserve residual speaker information, but are often essential to the downstream application and cannot be removed without sacrificing utility. Still, we predict some remaining potential to obfuscate residual speaker-dependent, but application-independent cues. Future work should target the anonymisation of these residual cues to improve performance. In parallel, and as a matter of course, we should explore the use for evaluation of stronger ASV systems which use both timbral and non-timbral cues.

7. Conclusions and Perspectives

In this paper we report experimental evidence of the critical need to address both timbral and non-timbral characteristics in voice anonymisation as well as in the use of ASV systems used for evaluation. Our analysis reveals a significant gap of up to 32% relative between estimates of anonymisation performance reported in the literature and comparable results achieved using stronger ASV systems and those specially designed to focus upon the use of residual, non-timbral cues. Our systems outperform the best systems submitted to the VoicePrivacy Attacker challenge for all but one anonymisation system, showing the influence of non-timbral cues in the re-identification of anonymised speakers. Stronger verification systems like WavLM are able to learn these cues upon fine-tuning with anonymised data and offer some potential to identify the source, and to study the influence of such residual cues. Based upon the results of this work, we are now studying the relative influence of specific characteristics like prosody, rhythm, style and accent, all non-timbral cues which carry speaker-dependent information and which can be used by an adversary to re-identify the original speaker post-anonymisation. Obfuscation of these characteristics will likely be challenging and will demand the design of new, specific anonymisation techniques. Arguably much more important in the nearer term is the adoption of similarly strong verification models for the benchmarking of competing anonymisation solutions to improve evaluation reliability.

8. References

[1] A. Nautsch, A. Jiménez, A. Treiber, J. Kolberg, C. Jasserand, E. Kindt, H. Delgado, M. Todisco, M. A. Hmani, A. Mtibaa, M. A. Abdelraheem, A. Abad, F. Teixeira, D. Matrouf, M. Gomez-Barrero, D. Petrovska-Delacrétaz, G. Chollet, N. Evans, T. Schneider, J.-F. Bonastre, B. Raj, I. Trancoso, and C. Busch, "Preserving privacy in speaker and speech characterisation," *Computer Speech & Language*, vol. 58, pp. 441–480, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0885230818303875

[2] N. Tomashenko, X. Miao, P. Champion, S. Meyer, X. Wang, E. Vincent, M. Panariello, N. Evans, J. Yamagishi, and M. Todisco, "The voiceprivacy 2024 challenge evaluation plan," 2024.

[3] G. Maimon and Y. Adi, "Speaking style conversion in the waveform domain using discrete self-supervised units," in *Findings of the Association for Computational Linguistics: EMNLP 2023*, H. Bouamor, J. Pino, and K. Bali, Eds. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 8048–8061. [Online]. Available: https://aclanthology.org/2023.findings-emnlp.541

[4] C. R. Pernet and P. Belin, "The role of pitch and timbre in voice gender categorization," *Frontiers in psychology*, vol. 3, p. 23, 2012.

[5] H. Ming, D. Huang, L. Xie, S. Zhang, M. Dong, and H. Li, "Exemplar-based sparse representation of timbre and prosody for voice conversion," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 5175–5179.

[6] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: emphasized channel attention, propagation and aggregation in TDNN based speaker verification," in *Interspeech 2020*, H. Meng, B. Xu, and T. F. Zheng, Eds. ISCA, 2020, pp. 3830–3834.

[7] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao *et al.*, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[8] 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*, 2020, pp. 1693–1697.

[9] N. Tomashenko, X. Wang, E. Vincent, J. Patino, B. M. L. Srivastava, P.-G. Noé, A. Nautsch, N. Evans, J. Yamagishi, B. O'Brien *et al.*, "The voiceprivacy 2020 challenge: Results and findings," *Computer Speech & Language*, vol. 74, p. 101362, 2022.

[10] 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 *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

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

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

[13] T. Bäckström, "Privacy in speech technology," in *Survey Talk - Interspeech 2022*, 2022. [Online]. Available: https://acris.aalto.fi/ws/portalfiles/portal/88447884/2022surveytalk.pdf

[14] B. M. L. Srivastava, N. Vauquier, M. Sahidullah, A. Bellet, M. Tommasi, and E. Vincent, "Evaluating voice conversion-based privacy protection against informed attackers," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 2802–2806.

[15] 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.

[16] N. Gengembre, O. Le Blouch, and C. Gendrot, "Disentangling prosody and timbre embeddings via voice conversion," in *Interspeech 2024*. International Speech Communication Association, 2024.

[17] J. Williams and S. King, "Disentangling style factors from speaker representations," in *Interspeech 2019*, 2019, pp. 3945–3949.

[18] J. Williams, Y. Zhao, E. Cooper, and J. Yamagishi, "Learning disentangled phone and speaker representations in a semi-supervised vq-vae paradigm," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 7053–7057.

[19] O. Zhang, O. Le Blouch, N. Gengembre, and D. Lolive, "An extension of disentanglement metrics and its application to voice," in *Interspeech 2023*, 2023, pp. 2878–2882.

[20] I. B. Amor and J.-F. Bonastre, "Ba-lr: Binary-attribute-based likelihood ratio estimation for forensic voice comparison," in *2022 International Workshop on Biometrics and Forensics (IWBF)*, 2022, pp. 1–6.

[21] C. H. Chan, K. Qian, Y. Zhang, and M. Hasegawa-Johnson, "Speechsplit2. 0: Unsupervised speech disentanglement for voice conversion without tuning autoencoder bottlenecks," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6332–6336.

[22] K. Qian, Y. Zhang, S. Chang, X. Yang, and M. Hasegawa-Johnson, "Autovc: Zero-shot voice style transfer with only autoencoder loss," in *International Conference on Machine Learning*. PMLR, 2019, pp. 5210–5219.

[23] P. Champion, T. Thebaud, G. Le Lan, A. Larcher, and D. Jouvet, "On the invertibility of a voice privacy system using embedding alignment," in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2021, pp. 191–197.

[24] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 5329–5333.

[25] M. Panariello, M. Todisco, and N. Evans, "Vocoder drift in x-vector–based speaker anonymization," in *Interspeech 2023*, 2023, pp. 2863–2867.

[26] N. Tomashenko, X. Miao, E. Vincent, and J. Yamagishi, "The first voiceprivacy attacker challenge evaluation plan," *arXiv preprint arXiv:2410.07428*, 2024.

[27] ——, "The first voiceprivacy attacker challenge," in *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025, pp. 1–2.

[28] 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," in *4th Symposium on Security and Privacy in Speech Communication*, 2024, pp. 61–66.

[29] 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," in *4th Symposium on Security and Privacy in Speech Communication*, 2024, pp. 67–71.

[30] N. Kuzmin, H.-T. Luong, J. Yao, L. Xie, K. A. Lee, and E.-S. Chng, "Ntu-npu system for voice privacy 2024 challenge," in *4th Symposium on Security and Privacy in Speech Communication*, 2024, pp. 72–79.

[31] W. Gu, Z. Liu, L. Chen, R. Wang, C. Guo, W. Guo, K. A. Lee, and Z.-H. Ling, "Ustc-polyu system for the voiceprivacy 2024 challenge," 2024.

[32] Y. Zhang, Z. Bi, F. Xiao, X. Yang, Q. Zhu, and J. Guan, "Attacking voice anonymization systems with augmented feature and speaker identity difference," in *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025, pp. 1–2.

[33] X. Lyu, Y. Wang, T. Zhao, and H. Liu, "Fast adaptation of pre-trained speaker verification system for source speaker tracking," in *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025, pp. 1–2.

[34] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[35] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," 2021. [Online]. Available: https://arxiv.org/abs/2106.09685

[36] Y. Li, Y. Zheng, Z. Guo, Y. Wang, J. Yin, and H. Fei, "Specwav-attack: Leveraging spectrogram resizing and wav2vec 2.0 for attacking anonymized speech," in *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025, pp. 1–2.

[37] C. O. Mawalim, A. Adila, and M. Unoki, "Fine-tuning titanet-large model for speaker anonymization attacker systems," in *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025, pp. 1–2.

[38] N. R. Koluguri, T. Park, and B. Ginsburg, "Titanet: Neural model for speaker representation with 1d depth-wise separable convolutions and global context," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 8102–8106.

[39] H. L. Xinyuan, A. Garg, Z. Cai, K. Duh, L. P. García-Perera, S. Khudanpur, N. Andrews, and M. Wiesner, "Hltcoe submission to the voiceprivacy attacker challenge," in *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025, pp. 1–2.

[40] 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)*, 2025, pp. 1–5.

[41] M. Panariello, N. Tomashenko, X. Wang, X. Miao, P. Champion, H. Nourtel, M. Todisco, N. Evans, E. Vincent, and J. Yamagishi, "The voiceprivacy 2022 challenge: Progress and perspectives in voice anonymisation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 3477–3491, 2024.

[42] C. Veaux, J. Yamagishi, K. MacDonald *et al.*, "Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit," *University of Edinburgh. The Centre for Speech Technology Research (CSTR)*, vol. 6, p. 15, 2017.

[43] J. Williams, K. Pizzi, N. Tomashenko, and S. Das, "Anonymizing speaker voices: Easy to imitate, difficult to recognize?" in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 12 491–12 495.