Disentangling prosody and timbre embeddings via voice conversion

Nicolas Gengembre[1], *Olivier Le Blouch*[1], *Cédric Gendrot*[2]

[1]Orange, France
[2]Laboratoire de Phonétique et de Phonologie, France
nicolas.gengembre@orange.com, olivier.leblouch@orange.com

Abstract

Modern voice conversion and anonymization architectures generally share a design preserving source linguistic content and expressivity while modifying speaker timbre characteristics. This approach leads to a converted signal quite perfectly synchronized with the source signal. In this paper, we hypothesize that this paradigm can help us to quantify the amount of speaker identity preserved in converted voice, refered here as *prosody* (including speech melody and rhythm). Based on this observation, we propose a method to split and disentangle speaker representation into complementary embeddings conveying respectively prosodic and timbre information. Additionally, we propose a method to evaluate prosody preservation in standard voice privacy architectures and we validate the power of prosodic and timbre embeddings to detect related voice attributes.

Index Terms: speaker disentanglement, voice conversion, voice attributes, timbre, prosody

1. Introduction

Speaker embeddings are pillars of Automatic Speaker Verification (ASV) and voice generation applications. They are widely used to clone a reference voice in text-to-speech, voice conversion and voice anonymization, assuming they convey all needed speakers' vocal characteristics. Indeed, with close-to-zero equal error rates (EER) even in hard conditions, ASV tends to be an almost solved problem, while high quality voice cloning is now within everyone's reach. Recent voice generation models such as zero-shot text-to-speech or voice conversion show high performance in terms of speech naturalness and quality, but still do not achieve perfect speaker similarity [1]. This raises the question of the importance of speaker encoders in voice generation pipelines. Mostly coming from ASV field, such as ECAPA [2], x-vectors [3], or GE2E [4], pretrained or trained within an end-to-end pipeline, their reliability to produce converted signals from a reference voice is not often addressed and the extracted embeddings remain non interpretable numerical vectors, disabling fine control on speaker characteristics.

Today, beyond speaker embeddings, a lot of various methods appear to control speaker characteristics at different levels, from a global description of a speaker identity to a fine-grain control. On the one hand, controlling global speaker representations in voice generation are more and more inspired by GPT [5] approaches. By instance, PromptTTS [6] propose to infer a voice from a prompt such as "an old woman whispering", while VALL-E [7], BASE TTS [8] and Tortoise-TTS [9] announce emergent abilities, like stage actors interpreting a didascalia. On the other hand, Text-to-Speech (TTS) systems such as FastSpeech2 [10] propose explicit, frame-based control on three prosodic information (pitch, intensity and phoneme durations), while SpeechSplit2 [11] enables aspect-specific voice conversion by disentangling speech into content, rhythm, pitch, and timbre.

Nowadays, an effort is also made to find a compromise between textual, quite ambiguous, voice descriptions and explicit - but not user-friendly - frame-based control. This effort commonly relies on a dichotomy between timbre and prosody.

Speech timbre can be defined as the specific sound characteristics - such as brightness or nasality - due to the shape of the vocal tract, but also due to the nature of the glottal flow induced by the larynx before going through the vocal tract, two examples being breathy or rough voices[12]. Speech timbre is often considered to be the major factor in an ASV task and is considered rather stable during speech but some aspects can fluctuate according to the situation, including some prosodic phenomena. Prosody is usually intended as the variation of melody and rhythm in speech. Both these parameters are dependent on the syntactic structure of the utterances and on the lexical stress patterns but prosody can be used to convey paralinguistic meanings such as attitudes, emotions or semantic context. It also provides specific information about the speaker - not only by the pitch level which could be considered as a biological factor here - by its specific modulations due to intra-speaker variations in speech. This is confirmed by some studies [13] stating that standard phoneme representations can contain information about speakers. Several attempts at capturing the prosodic specificities of speakers for ASV [14] have been published until speaker embeddings were used more recently with the aforementioned efficiency as they thoroughly encode both prosody and timbre.

Prosody vs timbre dichotomy is observed in Tortoise-TTS, which enriches the speaker's modelization by extracting two representations from the audio reference: a standard *speaker embedding* injected in the vocoder and GPT conditioning latents used to condition hidden representations of linguistic content. This raises the question of the adequate place in the generation pipeline to link specific speaker information (i.e. prosodic information in early "content" layers, timbral information in last "vocoder" layers). In a same spirit, authors of [15] learn an extractor of prosody representations to help emotion detection in speech signal, and a hierarchical decomposition of timbre and cadence is studied in [16] for zero-shot speech synthesis.

Standard voice conversion [17] and anonymization [18] systems also share architectures dealing with this dichotomy. Indeed, their objectives are basically to modify speaker characteristics while preserving two modalities, content and expressiveness, which are very closely related to prosody. That is where our study takes roots: what is the amount of prosodic and timbral information really converted by voice conversion?

The main contribution of this work lies in splitting the standard speaker embedding in two complementary prosody and timbre embeddings via their training on a converted version of the VoxCeleb dataset. Then a focus is made on the importance of prosodic information in speaker identification, addressing the bias of persistent speaker information in voice privacy architectures [19]. Finally ASV-based experiments and vocal attributes classifications are conducted to confirm the expected information preservation.

2. Methodology

Decomposing a voice into prosody and timbre is formally addressed by defining two complementary embeddings that describe each component independently (without any information shared by both embeddings) and that can represent, when combined together, the whole speaker characteristics, as a speaker embedding would do. The present study relies on the assumption that a number of voice conversion systems, including high performance ones, actually convert only the timbre of the voice, while keeping the prosody unchanged. Indeed, such systems rely on the decomposition of the voice signal into two parts : the global (constant over time) speaker characteristics and the verbal (transient) content. This latter part includes the phonemes dynamics, accentuation, and in some cases pitch variations and intensity contours, that are closely connected to the definition of prosody. The tool used in this study for conversion is RVC[1], or Retrieval-based Voice Conversion, based on VITS [20] and trained on the VCTK [21] dataset, which respects the previous assumptions thanks to its f0-free configuration. Starting from this observation, a prosody embeddings extractor is derived from a speaker embeddings model, but trained on converted voice data, as described in 2.1. A timbre embedding is defined as the complementary embedding to recover the whole speaker information, and the related model is described in 2.2.

2.1. Prosody extraction

The first step of this work consists in defining a prosody embeddings model as for speaker verification, except that it is trained on *converted data* to recover the source speakers, from which, according to our assumption, only the prosody remains. In that aim, one converts the voxceleb 2 dataset [22] (5965 speakers) into the first 99 speakers of the VCTK dataset (the remaining 10 speakers have been kept for tests on unseen voices). In order to benefit from the advantages of the self supervised learning representations (specially their ability to address a large range of tasks), the model is derived from a pretrained wavLM architecture[2] [23], followed by a stats-pooling (mean and variance over time of the last hidden layer output of the wavLM), two blocks of one linear layer followed by a rectified linear unit activation and a batch normalization (with output sizes equalling 512 and 250 respectively) and a final classification layer trained to classify the source speakers identities, using the additive angular margin paradigm [24] (margin 0.25). The so-called prosody embedding, denoted as $\mathbf{e_{pro}}$, is defined as the output of the penultimate layer and has 250 dimensions. The model trained in this way, designed to capture the prosodic part of the voice, is hereafter called W-PRO. Similarly, a system with the same architecture but trained on the original unconverted VoxCeleb1&2 dataset (7323 speakers), used as a baseline, captures the whole speakers characteristics and is denoted as W-SPK.

[1] https://github.com/RVC-Project

[2] https://huggingface.co/microsoft/wavlm-large

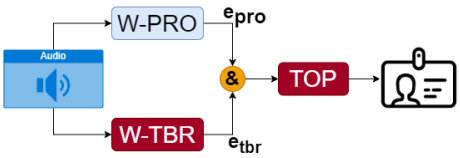

Figure 1: *Timbre model training. W-PRO is frozen while the red parts TOP and W-TBR are trained. The symbol "&" denotes concatenation. The TOP module classifies speakers identities.*

2.2. Timbre extraction

With W-PRO in hand, one can design another system called W-TBR that extracts the complementary information required to fully recognize speakers, assumed to be the voice timbre embedding, which is denoted as $\mathbf{e_{tbr}}$. The previously trained W-PRO system is used to extract $\mathbf{e_{pro}}$ from the data. In parallel, W-TBR produces $\mathbf{e_{tbr}}$, and both embeddings are concatenated to feed another module called TOP that classifies the speakers identities. TOP 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). TOP and W-TBR are trained jointly on the original unconverted data of VoxCeleb1&2, while W-PRO is frozen (see Figure 1). By these means, W-TBR is guided to convey the timbre information, complementary to the prosodic information. It is designed with a similar architecture as W-PRO except that its embeddings size is reduced to 128 dimensions to avoid the capture of redundant information (that might otherwise include prosodic components), following the bottleneck paradigm described for instance in [25].[3]

2.3. Training details

W-SPK, W-PRO and W-TBR have been trained on 6 epochs and a batch size of 16, filled with randomly picked excerpts of random durations in the range 3-5 seconds, on one NVIDIA A100 GPU board. The learning rates depend on the layers : 10^{-5} for the encoder layers of the wavLM (transformers), 0 for the feature extraction convolutional layers, and 10^{-4} for the linear classification layers after the wavLM ones and the TOP module.

3. Experiments

This section exposes the experiments led to judge the quality of proposed prosody and timbre embeddings. First, standard ASV experiments are described in subsection 3.1, then the case of Voice Privacy is addressed through the bias of residual speaker information in baseline architectures. Finally, usual voice community datasets and associated vocal attributes are browsed in order to confirm the expected links between embeddings type and related attributes.

3.1. Speaker Verification

The ECAPA-TDNN[4] public speaker embedding extractor pretrained by speechbrain [26] teams and refered in this article as ECAPA is used as an anchor to compare our results. On the commonly used Voxceleb test clean evaluation benchmark, all

[3] Apart from this empirical parameter setting, a more comprehensive study on the optimal dimensions of the prosodic and timbral embeddings is kept for future work.

[4] https://huggingface.co/speechbrain/spkrec-ecapa-voxceleb

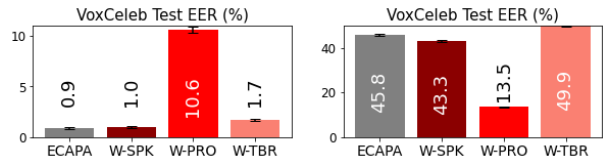

Figure 2: *Comparative performance of speaker embeddings on the Voxceleb Test Clean dataset in terms of EER (lower value means better identification).* left *: original dataset,* right *: dataset converted to the first 99 VCTK speakers.*

speaker embeddings reach EERs lower than 2% as shown on Figure 2-*left* except W-PRO that achieves 10.6%. That confirms that W-PRO is not made to be used as an ASV by itself, but also shows that it still captures speakers' identity to some extent.

As a first clue to confirm the intuitions behind this work, the EERs computed on the converted version of Voxceleb test clean are presented in Figure 2-*right*. In this evaluation, each utterance is converted to a speaker which is randomly chosen among the 99 seen during the training phase of our systems, and trials are labeled as targets when utterances share the same source speaker. All systems except W-PRO achieve EERs higher than 40%, confirming that they are unable to identify the source speakers. But W-PRO achieves 13.5% EER, which means that this model succeeds in pairing a same source speaker in a large number of trials, thus proving that the link between speaker identities and converted signals is not broken. In other words, it is somehow able to identify source speakers by means of its prosodic characteristics.

Two complementary experiments have been conducted in order to check if such an ability of W-PRO is biased by the target speakers identities that were also used during the training of W-PRO, or by the conversion system used. In the first one we convert the test set to a random speaker among the remaining 10 VCTK unseen speakers, while in the second one we use a pretrained FreeVC model [27] instead of RVC. Both experiments lead to results close to the one presented here-above (EER = 14.1% and 13.4% resp., to be compared with 13.5% and with confidence intervals of 0.4%). We then conclude that the ability of W-PRO to recognize a source speaker from its prosody still holds for more generic voice conversions.

3.2. Voice Privacy

The success in pairing source speakers in converted speech naturally leads to an interest in voice privacy. Indeed, as anonymization and voice conversion systems tend to share their underlying architectures, it is interesting to know if the same effect is observed. In 2020, the Voice Privacy Challenge[28] proposed to evaluate anonymization systems under four conditions : *Unprotected* (raw enrollments vs raw trials), *Ignorant attacker* (raw enrollments vs anonymized trials), *Lazy-informed* (anonymized enrollments vs anonymized trials) and *Semi-informed* (anonymized enrollments vs anonymized trials with the ASV model retrained on anonymized data). Ignorant attacker and Lazy-informed conditions were dropped from the next evaluation in 2022 [18]. This section arguments that, despite the good results (*i.e.* EER close to 50%) obtained by most anonymization systems under the ignorant attacker scenario, their performance might drop when facing speaker identification systems focused on prosodic characteristics only.

The scores presented in Figure 3-*top* describe the performance of ASV systems under the *unprotected scenario*, a conventional ASV evaluation where only raw, non anonymized, enrollments and trials are used. All systems except VPC-A (*a.k.a* ASV_{eval}^{anon} in [18]) and W-PRO achieve EERs lower than 10%. As expected, VPC-A and W-PRO, both trained to extract source speaker, are poor candidates for this task. Nevertheless, as seen on Figure 3-*bottom*, W-PRO reduces the EER in *ignorant attacker* condition by 20 to 30 points relatively to other ASV systems, without any adaptation or information coming from anonymization systems. As a conclusion, W-PRO, simply born from theoretical knowledge of standard voice conversion architectures gives a way to measure the residual speaker information bias in similar anonymization baselines (ie. different from "ASR+TTS"-based systems). As a matter of fact, it could enrich the pool of voice privacy metrics as a measurement of prosody preservation, partially related to speaker identity, in anonymized speech. To illustrate: applying this method to the recent codec-based approach proposed in [19] shows a significant improvement of approximately 10 points on these datasets compared to B1.b baseline, confirming a vanishing of speaker information in anonymized content.

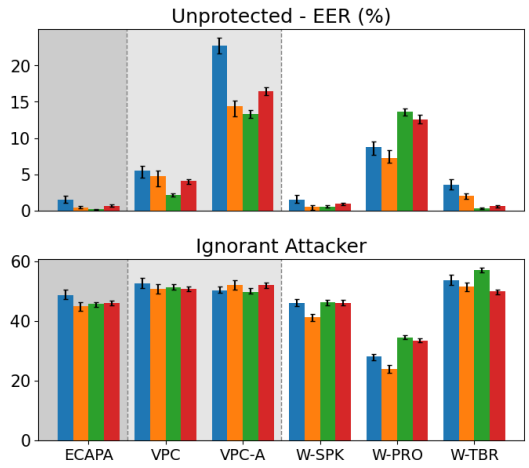

Figure 3: *Voice Privacy experiments. EERs under* unprotected *and* ignorant attacker *conditions on LibriSpeech [29] and VCTK [21] as defined in [18] on Librispeech dev (blue) and test (yellow), VCTK dev (green) and test (red). Male and female splits are merged for readability. The anonymized signals are generated via the B1.b baseline system described in [18]. VPC & VPC-A are original ASV systems supplied by the VPC2022 framework* [5]*, trained on LibriSpeech-train-clean-360.*

3.3. Prosodic or timbral attributes

In this section we analyze how prosody and timbre embeddings behave when used to classify lower level voice attributes such as emotion, style, age, gender, accent and language. In that purpose, we selected datasets where annotations of these attributes were available. The datasets used are listed in Table 1. The table also describes the evaluation protocol used in each case (cross validation, subset splits).

To evaluate the amount of attribute-related information hidden in the different embeddings, Support Vector Classifiers are trained to predict these attributes directly from the embeddings provided by frozen models, as the aim of the experiment is not to obtain optimal performance for these tasks (as opposed to what is done in [15] where the prosody model is fine-tuned, so it no more produces an embedding describing only the prosody).

Table 1: *Prosody and timbre attributes datasets. The metric used for all the experiments is the Weighted Accuracy WA, defined as in [15]. For the ArtieBias dataset, 80%/20% splits are defined according to each attribute. For PTSVOX and IEMOCAP, the "script" attribute refers to improvised/spontaneous versus scripted/read speaking style.*

Dataset	#samples	#speakers	Attributes(#classes)	Evaluation protocol
IEMOCAP[30]	5531	10	emotion(4), gender(2), script(2)	Leave-one-session-out
RAVDESS[31]	1440	24	emotion(8)	Leave-one-speaker-out, 24 folds
ESD[32]	35000	20	emotion(5), language(2)	Leave-one-speaker-out, 20 folds
ArtieBias[33]	1712	970	accent(16), age(7), gender(2)	80% training, 20% test
PTSVOX[34]	7506	24	script(2)	Leave-one-speaker-out, 24 folds

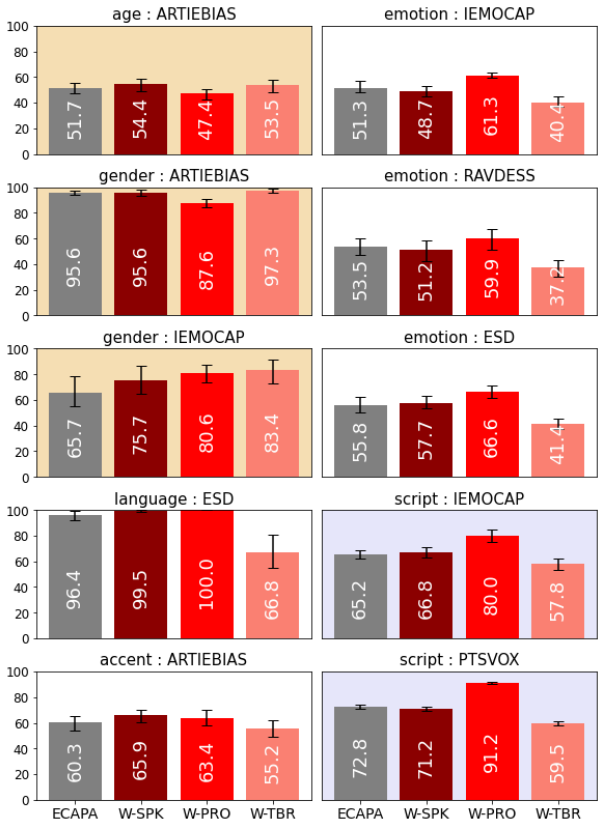

Figure 4: *Performance of embeddings on prosodic and timbral attributes from datasets in Table 1. Timbre-related speakers characteristics (age, gender) have a yellow background.*

The classification accuracies given in Figure 4 give clues about which attributes are best captured by which embeddings. Unsurprisingly, attributes related to emotions (RAVDESS, IEMOCAP and ESD) and speaking style (PTSVOX and IEMOCAP) are much better captured by W-PRO while W-SPK and W-TBR leads to very low results, with more than 20% difference in every cases. This is also the case for language and accent attributes, but the results are less salient and with greater confidence intervals due probably to less diverse test sets (few accents, few speakers...).

Oppositely, speakers characteristics such as age or gender are better conveyed by W-TBR and W-SPK (charts with a yellow background in Figure 4). These results confirm that absolute speaker f0, closely related to gender, is not well captured by W-PRO (which was expected as it is not preserved by the conversion system used to train it), neither are the duration and f0 variations due to age.

4. Discussion and Conclusion

This study details a decomposition of speaker representations in two complementary timbre and prosody embeddings, mining the hypothesis of a *dichotomy by design* shared by standard voice conversion architectures and showing good abilities for different tasks. The prosodic model is, in some situations, able to identify speakers after voice conversion or anonymization. It notably confirms that anonymized speech signals still contain speaker information while providing a measurement to assess the extent to which the information persists. The produced prosodic embeddings can also lead to accurate classification of voice attributes known to be conveyed by prosody, such as emotions or speaking style, while the timbre embeddings are relevant for age and gender, as expected. In the training phase, W-PRO appears to capture the prosodic level broadly enough to learn both the characteristics that hold allover each speaker's utterances and attributes such as emotions or expressivity, different from one utterance to another. Moreover, the two embeddings seem to be complementary since when one embedding captures correctly one attribute, the performance of the other is usually low, and conversely. This tends to demonstrate that these two components have been correctly disentangled by our method, although complementary experiments would be necessary to confirm and, if possible, to improve, this statement - for instance by adding a regularization term in the loss function that minimizes the mutual information between the embeddings.

Further analyses are needed so as to improve the proposed encoders and be able to include them in a fine-grained voice conversion. This kind of decomposition shall be useful to inject the right piece of speaker information at the right place in a generative pipeline. This would open new ways of partial cloning, enabling to choose the timbre of a speaker and the speaking style of another. In the same idea, one could imagine to re-split these representations into finer grain embeddings in order to extract speaker-dependant style and emotions from prosody or breathiness and roughness from timbre, requiring new dedicated dataset annotations or smart automatic disentanglement.

The phonetics and sociophonetics domains could also benefit from these results: aging is known for slowing speech rhythm and the attributes of gender are specifically analyzed in transgender studies.

Finally, this dichotomy approach obviously remains a simplified scheme of reality and future works will address new methods to refine and capture the vocal attributes located on the frontier between timbre and prosody.

5. Acknowledgements

The research reported here was supported by the ANR-23-CE23-0018 EVA project.

6. References

[1] K. Milewski, S. Zaporowski, and A. Czyżewski, "Comparison of the ability of neural network model and humans to detect a cloned voice," *Electronics*, vol. 12, no. 21, p. 4458, 2023.

[2] 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.

[3] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *ICASSP*. IEEE, 2018, pp. 5329–5333.

[4] L. Wan, Q. Wang, A. Papir, and I. L. Moreno, "Generalized end-to-end loss for speaker verification," in *ICASSP*. IEEE, 2018, pp. 4879–4883.

[5] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[6] Z. Guo, Y. Leng, Y. Wu, S. Zhao, and X. Tan, "Prompttts: Controllable text-to-speech with text descriptions," in *ICASSP*. IEEE, 2023, pp. 1–5.

[7] C. Wang, S. Chen, Y. Wu, Z. Zhang, L. Zhou, S. Liu, Z. Chen, Y. Liu, H. Wang, J. Li *et al.*, "Neural codec language models are zero-shot text to speech synthesizers," *arXiv preprint arXiv:2301.02111*, 2023.

[8] M. Łajszczak, G. Cámbara, Y. Li, F. Beyhan, A. van Korlaar, F. Yang, A. Joly, Á. Martín-Cortinas, A. Abbas, A. Michalski *et al.*, "Base tts: Lessons from building a billion-parameter text-to-speech model on 100k hours of data," *arXiv preprint arXiv:2402.08093*, 2024.

[9] J. Betker, "Better speech synthesis through scaling," *arXiv preprint arXiv:2305.07243*, 2023.

[10] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "Fastspeech 2: Fast and high-quality end-to-end text to speech," in *ICLR*, 2020.

[11] 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*. IEEE, 2022, pp. 6332–6336.

[12] F. Nolan, "Forensic Speaker Identification and the Phonetic," *A Figure of Speech: A Festschrift for John Laver*, p. 385, 2014, publisher: Routledge.

[13] P. Champion, D. Jouvet, and A. Larcher, "Are disentangled representations all you need to build speaker anonymization systems?" in *INTERSPEECH 2022-Human and Humanizing Speech Technology*, 2022.

[14] E. Shriberg, L. Ferrer, S. Kajarekar, A. Venkataraman, and A. Stolcke, "Modeling prosodic feature sequences for speaker recognition," *Speech Communication*, vol. 46, no. 3, pp. 455–472, 2005, quantitative Prosody Modelling for Natural Speech Description and Generation.

[15] L. Qu, T. Li, C. Weber, T. Pekarek-Rosin, F. Ren, and S. Wermter, "Disentangling prosody representations with unsupervised speech reconstruction," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023.

[16] J. Y. Lee, J.-S. Bae, S. Mun, J. Lee, J.-H. Lee, H.-Y. Cho, and C. Kim, "Hierarchical timbre-cadence speaker encoder for zero-shot speech synthesis," in *INTERSPEECH*, 2023.

[17] T. Walczyna and Z. Piotrowski, "Overview of voice conversion methods based on deep learning," *Applied Sciences*, vol. 13, no. 5, p. 3100, 2023.

[18] N. Tomashenko, X. Wang, X. Miao, H. Nourtel, P. Champion, M. Todisco, E. Vincent, N. Evans, J. Yamagishi, and J.-F. Bonastre, "The voiceprivacy 2022 challenge evaluation plan," *arXiv preprint arXiv:2203.12468*, 2022.

[19] M. Panariello, F. Nespoli, M. Todisco, and N. Evans, "Speaker anonymization using neural audio codec language models," 2024.

[20] J. Kim, J. Kong, and J. Son, "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech," in *International Conference on Machine Learning*. PMLR, 2021, pp. 5530–5540.

[21] 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.

[22] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," in *INTERSPEECH*, 2018.

[23] 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.

[24] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4690–4699.

[25] 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.

[26] M. Ravanelli, T. Parcollet, P. Plantinga, A. Rouhe, S. Cornell, L. Lugosch, C. Subakan, N. Dawalatabad, A. Heba, J. Zhong *et al.*, "Speechbrain: A general-purpose speech toolkit," *arXiv preprint arXiv:2106.04624*, 2021.

[27] J. Li, W. Tu, and L. Xiao, "Freevc: Towards high-quality text-free one-shot voice conversion," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[28] 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.

[29] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *ICASSP*. IEEE, 2015, pp. 5206–5210.

[30] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, pp. 335–359, 2008.

[31] S. R. Livingstone and F. A. Russo, "The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english," *PloS one*, vol. 13, no. 5, p. e0196391, 2018.

[32] K. Zhou, B. Sisman, R. Liu, and H. Li, "Seen and unseen emotional style transfer for voice conversion with a new emotional speech dataset," in *ICASSP*. IEEE, 2021, pp. 920–924.

[33] J. Meyer, L. Rauchenstein, J. D. Eisenberg, and N. Howell, "Artie bias corpus: An open dataset for detecting demographic bias in speech applications," in *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 2020, pp. 6462–6468.

[34] A. Chanclu, L. Georgeton, C. Fredouille, and J.-F. Bonastre, "Ptsvox: une base de données pour la comparaison de voix dans le cadre judiciaire (ptsvox: a speech database for forensic voice comparison)," in *JEP, TALN, RÉCITAL*, 2020, pp. 73–81.