Transparent Exchange of Speaker Attributes

Jiusi Zheng[1,2]*, Martha Larson*[1]*, Tom Bäckström*[2]

[1]Institute for Computing and Information Sciences, Radboud University, Netherlands
[2]Department of Information and Communications Engineering, Aalto University, Finland

jiusi.zheng@ru.nl, martha.larson@ru.nl, tom.backstrom@aalto.fi

1. Research Objective

Current voice anonymization methods either rely on holistic speaker embeddings or enforce strict disentanglement between latent factors. Holistic embeddings tend to entangle multiple attributes, making it difficult to modify a single attribute without inadvertently affecting others. In contrast, disentangled representations enforce rigid one-to-one mappings between latent factors and voice attributes, often lacking flexibility and failing to capture fine-grained variations.

Our research aims to achieve attribute-preserving voice anonymization in an interpretable and controllable manner. Here, interpretability refers to the ability to manipulate the learned speaker representations using structurely grounded components (e.g., frequency bands or temporal scales), thereby enabling a clearer understanding of how modifications in the speaker representation affect the output speech; controllability denotes the ability to maintain specific voice attributes (e.g., emotion, accent, gender, pitch, and intensity) during voice anonymization.

To accomplish this, we will first decompose the speech signal into structured components, which encourage the encoded representations to follow a more understandable structure. Working on the decomposed components, we then blindly disentangle the target attribute from speaker identity, enabling the generation of anonymized speech that preserves the target attribute. This approach is expected to offer two main advantages. First, we can know how and where we modify in the acoustic representation to achieve attribute-preserving voice anonymization. Second, by introducing an intermediate structured representation, the method facilitates more effective disentanglement between speaker identity and target attributes, reducing the reliance on large amounts of labeled data while remaining broadly applicable. Currently, we have explored frequency components as one form of inductive bias, and we are investigating additional biases to further enhance the framework.

2. Related Work

Recent work on voice anonymization and controllable speech synthesis can be broadly categorized into holistic speaker representation approaches and disentanglement-based methods. Holistic approaches [1, 2, 3, 4] typically rely on global speaker embeddings and achieve anonymization by transforming them. Although effective in anonymizing speaker identity, these methods inherently anonymize multiple voice attributes simultaneously, including gender, emotion, and speaking style, making it difficult to modify a specific attribute without inadvertently affecting others. As a result, they offer limited controllability for applications that require selective attribute preservation.

To address this limitation, disentanglement-based approaches [5, 6, 7] aim to decompose speech into independent latent factors. However, they typically rely on the independence between latent factors and explicit one-to-one mappings between factors and voice attributes. In practice, high-level voice attributes, such as emotion, health condition, gender, and accent, are inherently correlated and distributed across the speech signal. Moreover, most existing approaches attempt to directly disentangle semantic attributes in a single stage, without explicitly considering the underlying acoustic structure of speech. In contrast, we adopt a two-stage perspective. We first learn a structured representation that organizes speaker embeddings according to the time scale structure of the speech signal. Building on these decomposed factors, we disentangle voice attributes without requiring strict independence of the latent components. This design allows attributes to be controlled in a more flexible and fine-grained manner, while also providing better alignment with speech science, as it facilitates analysis of how different attributes are distributed across different acoustic dimensions.

From an information-theoretic standpoint, although recent works [5, 6, 7] have attempted to disentangle voice attributes within speaker embeddings using supervised multi-task and adversarial objectives, such disentanglement remains fundamentally challenging: Attribute annotations are often scarce or subjective, and many attributes share underlying acoustic cues, resulting in inherently correlated representations. This motivates an alternative perspective that focuses on acoustic structure. In particular, time scales provide an interpretable axis along which voice attributes can be analyzed, enabling more transparent control without requiring attribute labels.

3. Open Questions

Despite recent progress in attribute-preserving voice anonymization, several key challenges remain in bridging existing approaches toward transparent, controllable voice anonymization. First, we need to experimentally verify if our proposed method is suitable for controllable voice anonymization. Second, it remains an open question whether incorporating signal-level supervision during training can provide a suitable foundation for controllable manipulation of voice attributes. We plan to address these challenges by exploring structured and decomposed representations that bridge signal-level manipulation and attribute-level control, enabling more flexible voice attribute manipulation for privacy-preserving applications.

4. References

[1] 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 *10th ISCA Workshop on Speech Synthesis (SSW 10)*, 2019, pp. 155–160.

[2] 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)*. IEEE, 2023, pp. 1–5.

[3] 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.

[4] P. Champion, "Anonymizing speech: Evaluating and designing speaker anonymization techniques," Ph.D. dissertation, University of Lorraine, 2023.

[5] F. Teixeira, A. Abad, B. Raj, and I. Trancoso, "Privacy-oriented manipulation of speaker representations," *IEEE Access*, vol. 12, pp. 82 949–82 971, 2024.

[6] C. Luu, S. Renals, and P. Bell, "Investigating the contribution of speaker attributes to speaker separability using disentangled speaker representations," in *Interspeech 2022*, 2022, pp. 610–614.

[7] J.-H. Huang, W.-T. Lee, and C.-H. Wu, "USD-AC: Unsupervised speech disentanglement for accent conversion," in *Interspeech 2024*, 2024, pp. 4388–4392.