Emo-StarGAN: A Semi-Supervised Any-to-Many Non-Parallel Emotion-Preserving Voice Conversion

Suhita Ghosh[1*], Arnab Das[1,3*], Yamini Sinha[2,] Ingo Siegert[2], Tim Polzehl[3], Sebastian Stober[1]

1. Artificial Intelligence Lab (AILab), Otto-von-Guericke-University, Magdeburg, Germany
2. Mobile Dialog Systems, Otto-von-Guericke-University, Magdeburg, Germany
3. Speech and Language Technology, German Research Center for Artificial Intelligence (DFKI)

✉ suhita.ghosh@ovgu.de

SPONSORED BY THE Federal Ministry of Education and Research

* equal contribution

Introduction

The increasing use of cloud-based speech devices raises concerns about the confidentiality and protection of shared sensitive data.

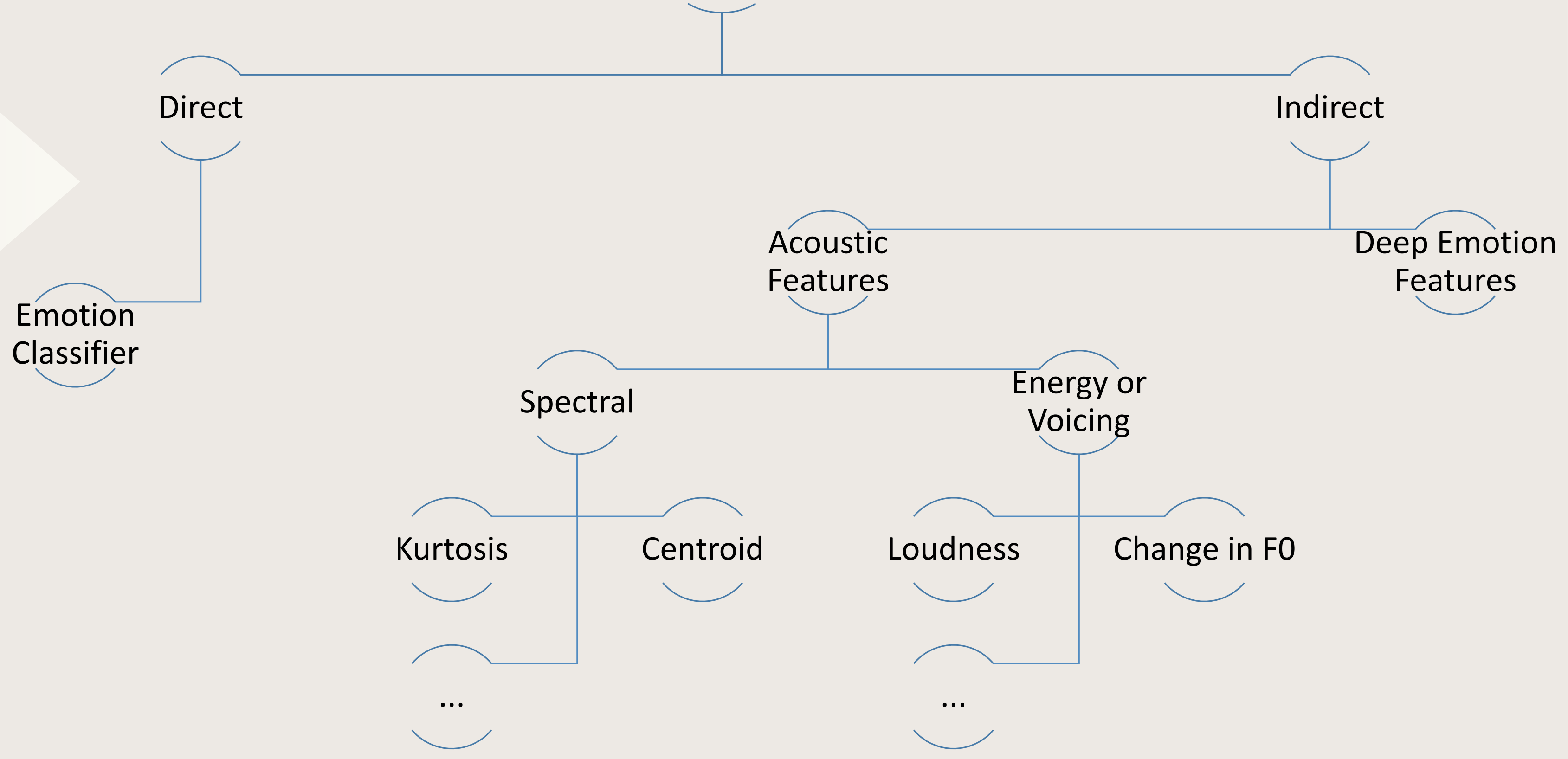
- **Voice conversion** is one of the ways to achieve speech anonymisation, which prevents data misuse.
- **Emotion preservation** is crucial for natural human-computer interaction.
- The state-of-the-art voice conversion methods **fail** to preserve emotions for diverse emotions and acoustic conditions.

We present an emotion-preserving non-parallel voice conversion technique, trained on novel affect-aware losses using acoustic and deep emotion features.

Methods

1

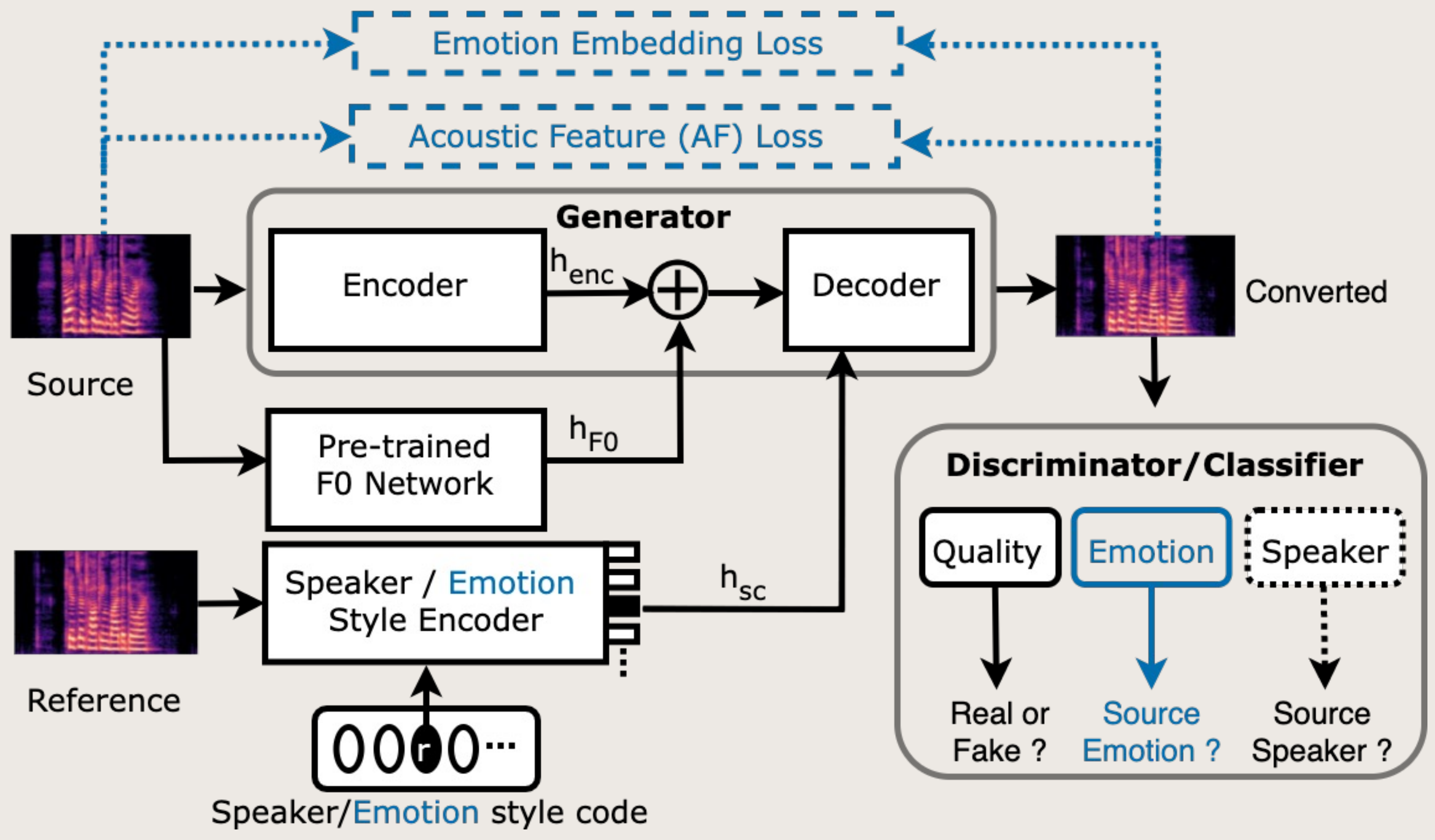
EMOTION SUPERVISION TECHNIQUES



2

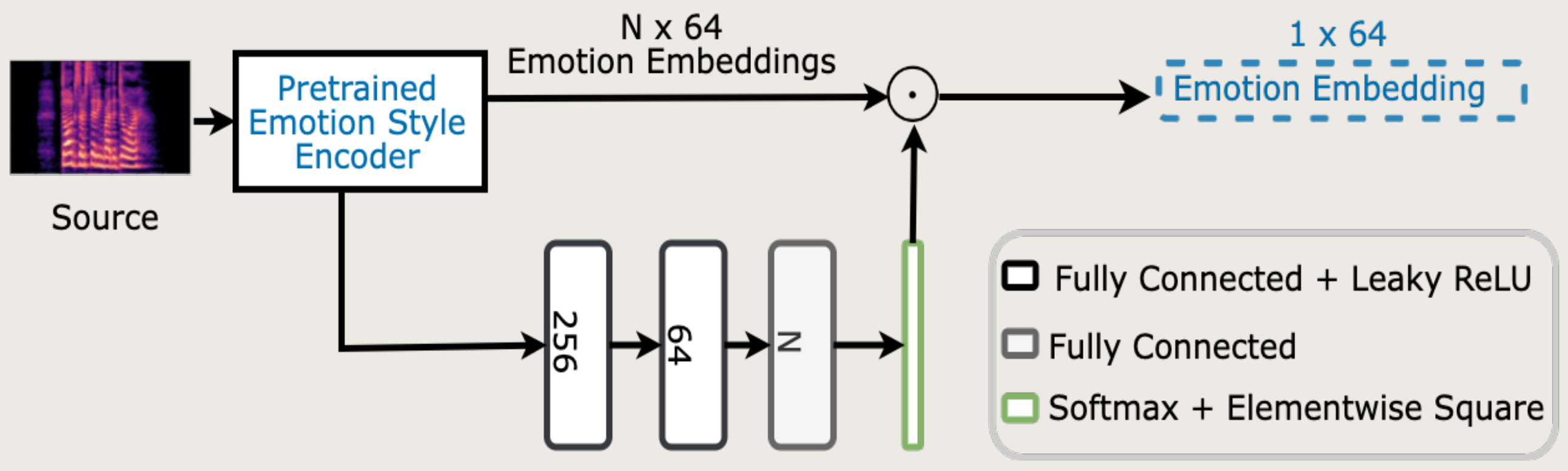
ARCHITECTURE

- ➢ The proposed framework is adapted from StarGANv2-VC [1].
- ➢ For **voice conversion,** the style encoder captures speaker embeddings.
- ➢ The emotion embeddings are derived through the same framework, but trained for **emotion conversion**.



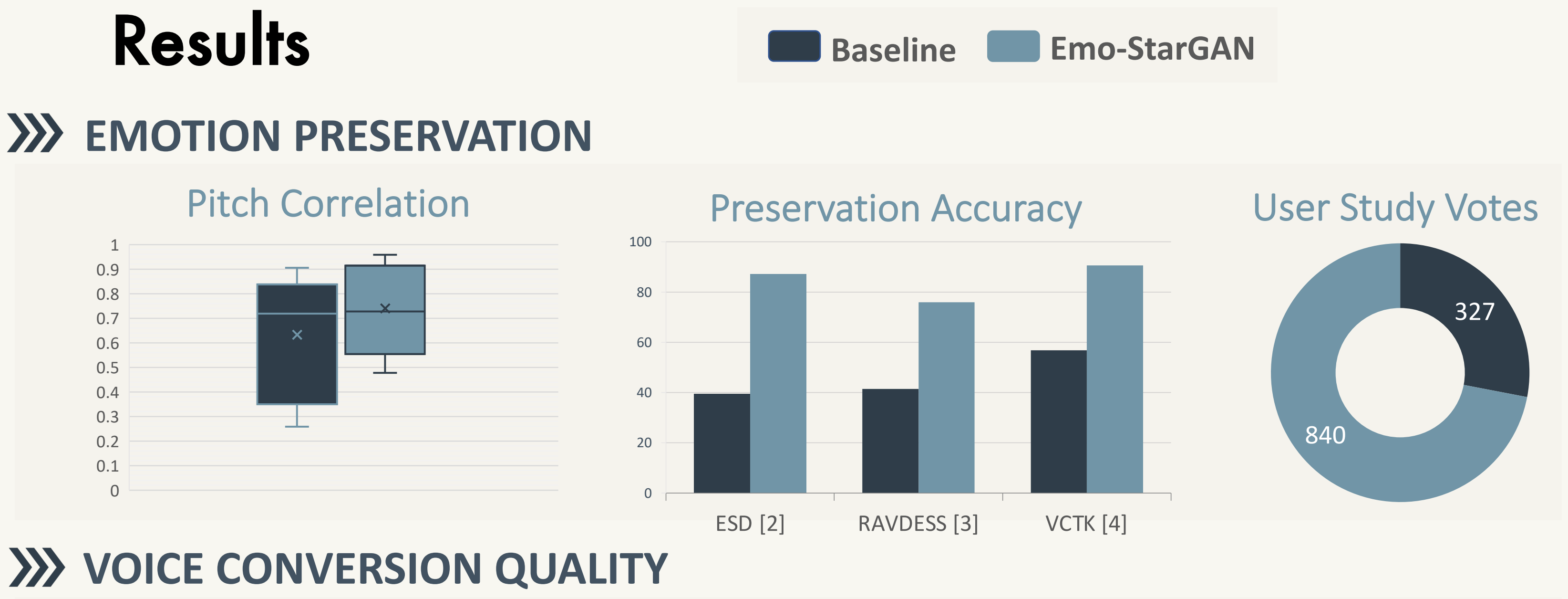
3

AUTOMATIC EMOTION EMBEDDING EXTRACTION

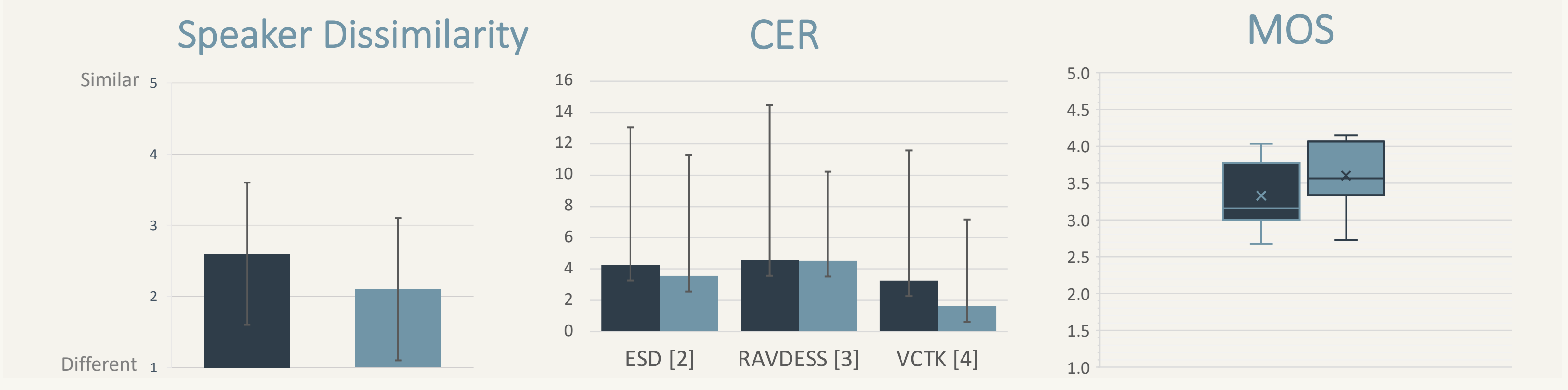

Source
Pretrained Emotion Style Encoder
N x 64 Emotion Embeddings
1 x 64 Emotion Embedding
256
64
N
Fully Connected + Leaky ReLU
Fully Connected
Softmax + Elementwise Square

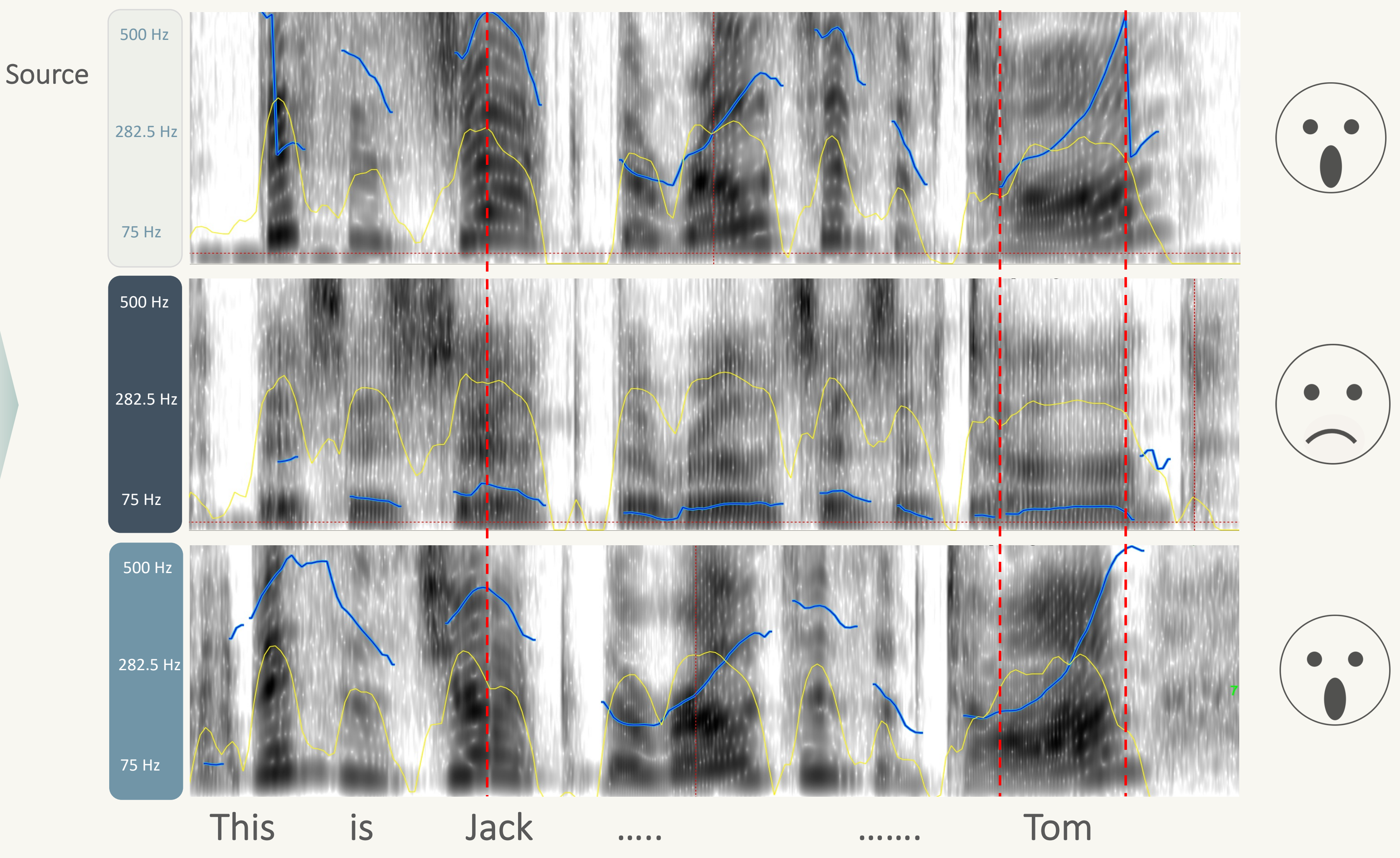
Results

EMOTION PRESERVATION



VOICE CONVERSION QUALITY





Discussion

Anonymise
Preserve Emotion
Preserve Intelligibility
Preserve Quality

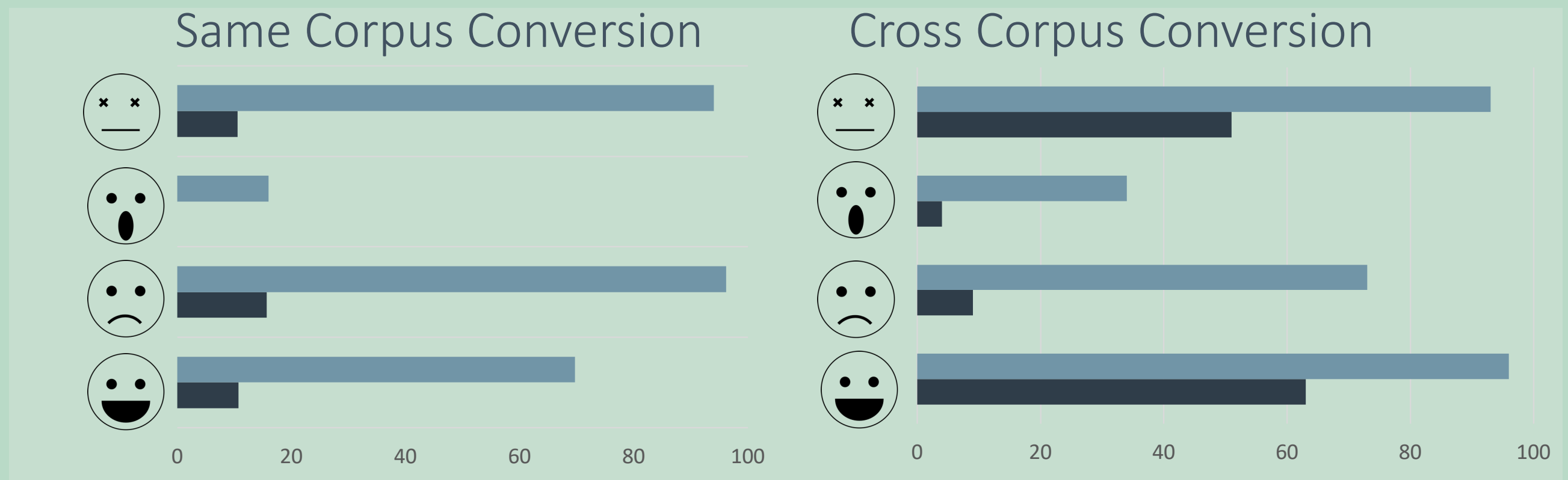
WHICH TECHNIQUE PRESERVES EMOTION THE MOST?

- ➢ The indirect method using **deep emotion embeddings** contributes more than direct supervision by the emotion classifier.
- ➢ **Spectral features** are more helpful than energy-based ones, as they add additional information about higher-level harmonics [5].

WHICH EMOTIONS ARE DIFFICULT TO PRESERVE?

Surprise is most difficult to preserve, reported similarly for emotion recognition [6].

EMOTION PRESERVATION ACCURACY



Conclusion

- ➢ In this work, we propose a semi-supervised emotion-preserving voice conversion framework trained on non-parallel data.
- ➢ Further, we introduce method-agnostic affect-aware losses, which can be used even in the absence of emotion labels.
- ➢ The proposed method significantly improves emotion preservation over vanilla StarGANv2-VC without compromising intelligibility and anonymisation.

As future work, we plan to improve emotion preservation for complex emotions by incorporating losses beneficial to a specific emotion. Further, we plan to extend the method with emotion embeddings learned from multi-label and arousal-valence labelled datasets.

References

1. Li, Y. A. Li, A. Zare, and N. Mesgarani, “StarGANv2-VC: A Diverse, Unsupervised, Non-Parallel Framework for Natural-Sounding Voice Conversion,” in Proc. Interspeech 2021, 2021, pp. 1349–1353.
2. Zhou, Kun, et al. "Emotional voice conversion: Theory, databases and ESD." Speech Communication 137 (2022): 1-18.
3. Livingstone, Steven R., and Frank 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 13.5 (2018): e0196391.
4. Yamagishi, Junichi et al. “CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit (version 0.92).” (2019).
5. F. Weninger, F. Eyben, B. W. Schuller, M. Mortillaro, and K. R. Scherer, “On the acoustics of emotion in audio: what speech, music, and sound have in common,” Frontiers in psychology, vol. 4, p. 292, 2013.
6. L. Chen, X. Mao, Y. Xue, and L. L. Cheng, “Speech emotion recognition: Features and classification models,” Digital signal processing, vol. 22, no. 6, pp. 1154–1160, 2012.