A Curriculum Learning Paradigm for Speech Emotion Transfer: Overcoming Real-Data Training Challenges

Abstract

Speech-to-speech emotion transfer modifies the emotional tone of a speech signal while preserving its linguistic content and speaker identity. This task is vital for expressive speech synthesis, emotionally adaptive voice conversion, and naturalistic human-computer interaction. However, models trained directly on real speech often struggle to generalize due to high variability in emotional expression, caused by factors such as speaker individuality, cultural norms, and contextual nuances. These challenges make it difficult for models to disentangle and manipulate emotional signals without compromising speech naturalness or identity.

In this work, we introduce *EmoTransfer*, an end-to-end audio-conditioned emotion transfer model that bypasses the need for text input and directly operates on acoustic features. To overcome the limitations of real-world data, we propose a curriculum learning paradigm in which the model is first trained on synthetic speech generated by a controllable text-to-speech (TTS) system with predefined emotional attributes. This phase provides a low-noise, highly structured environment for learning foundational emotion representations. The model is then fine-tuned on a progressively mixed dataset of real and synthetic speech to adapt to the complexity of real-world emotional variation.

Our experiments show that curriculum learning significantly enhances the model's generalization ability. EmoTransfer achieves a top-1 emotion similarity of 98% in the same-speaker/text setting and 97% in cross-speaker/text conditions, outperforming state-of-the-art baselines. Moreover, it receives the highest Mean Opinion Score (MOS) across emotion naturalness and speaker similarity in subjective evaluations.

To analyze how curriculum learning shapes emotion representation, we perform a t-SNE analysis on the latent emotion embeddings across four training settings (Figure 1). The model trained solely on synthetic data transfers emotion well within synthetic speech but exhibits poor clustering and generalization to real speech. Conversely, the model trained only on real speech displays entangled and poorly separated emotional clusters, indicating difficulty in learning distinct emotion patterns. When trained on a mixed dataset with 60% real and 40% synthetic data, the emotional embeddings begin to separate more clearly. Finally, fine-tuning on a dataset with 80% real and 20% synthetic speech results in well-structured and clustered emotion representations in t-SNE space, corresponding to improved emotion transfer performance. These findings underscore the effectiveness of curriculum learning in bridging the gap between synthetic and real speech and establishing robust emotional representations.

In addition to the proposed methodology, we contribute: (1) *EmoTransfer*, a novel audio-conditioned speech emotion transfer model; (2) a curriculum learning framework that strategically combines synthetic and real data; (3) an in-depth analysis of emotion representation learning using t-SNE visualizations; and (4) a publicly available dataset of 27 speakers, 9 emotions, and 27,000 audio samples, released at https://huggingface.co/datasets/anonymousforemotion/ET. Demos and source code are available at https://demopagea.github.io/EmoTransfer-demo/ and https://anonymous.4open.science/r/EmoTransfer-F882/.

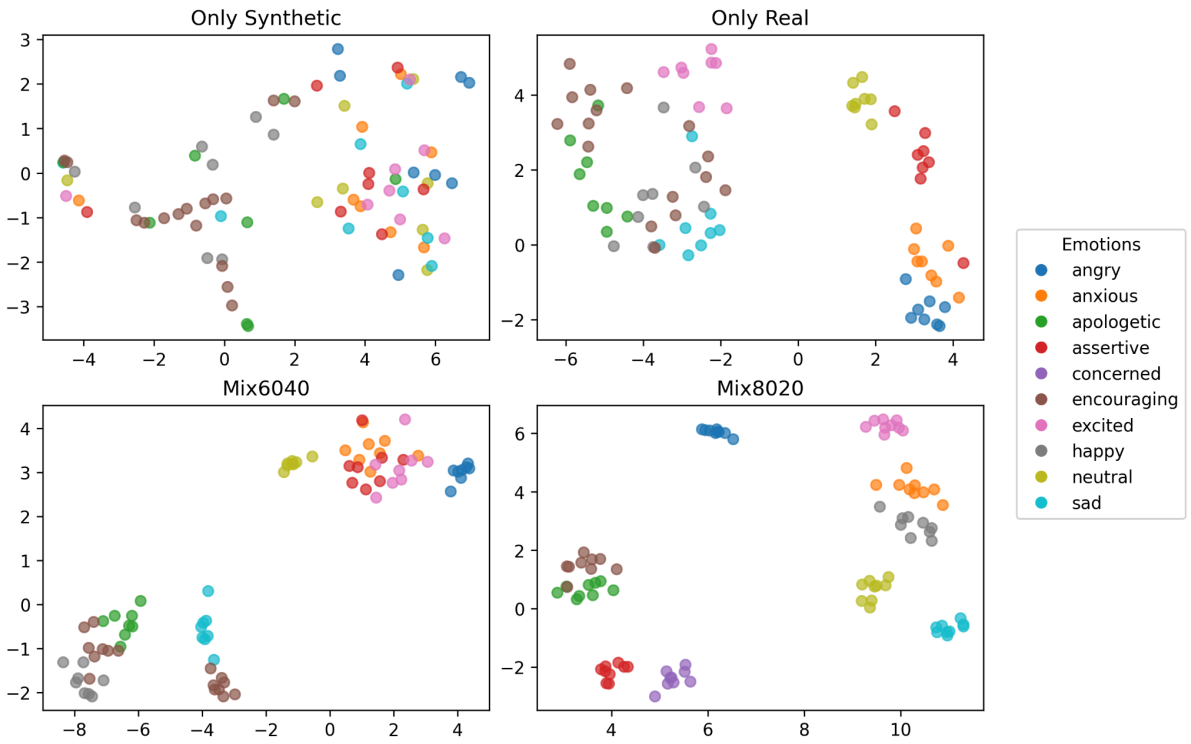

Figure 1. t-SNE analysis of emotion embeddings of real speech in curriculum learning models