Contribution ID: **48** Type: **Poster**

EMOCONV-DIFF: Diffusion-based Speech Emotion Conversion for Non-parallel and In-the-wild Data

Speech emotion conversion is the task of converting the expressed emotion of a spoken utterance to a target emotion while preserving the lexical content and speaker identity. While most existing works in speech emotion conversion rely on acted-out datasets and parallel data samples, in this work we specifically focus on more challenging in-the-wild scenarios and do not rely on parallel data. To this end, we propose a diffusion-based generative model for speech emotion conversion, the EmoConv-Diff, that is trained to reconstruct an input utterance while also conditioning on its emotion. Subsequently, at inference, a target emotion embedding is employed to convert the emotion of the input utterance to the given target emotion. As opposed to performing emotion conversion on categorical representations, we use a continuous arousal dimension to represent emotions while also achieving intensity control. We validate the proposed methodology on a large in-the-wild dataset, the MSP-Podcast v1.10. Our results show that the proposed diffusion model is indeed capable of synthesizing speech with a controllable target emotion. Crucially, the proposed approach shows improved performance along the extreme values of arousal and thereby addresses a common challenge in the speech emotion conversion literature.

I want to give a Lightning Talk

no

Author: RAJ PRABHU, Navin (Signal Processing)

Co-authors: Mr LAY, Bunlong (Universität Hamburg); Prof. LEHMANN-WILLENBROCK, Nale (Universität Hamburg); Mr WELKER, Simon (Universität Hamburg); Prof. GERKMANN, Timo

Presenter: RAJ PRABHU, Navin (Signal Processing)

Session Classification: Poster Session