**Introduction**

- Emotional voice conversion (EVC): transform the emotional prosody while preserving the linguistic content and speaker identity;
- Prior studies propose to disentangle the emotional prosody using an encoder-decoder network conditioned on discrete representation, such as one-hot emotion labels, but only learn to remember a fixed set of emotional styles;

**Our contributions:**

In this paper, we propose:
- a one-to-many EVC framework that does not need parallel data;
- to use deep emotional features to describe different emotional styles;
- release a multi-speaker and multi-lingual emotional speech dataset;

**To our best knowledge, it is the first reported study on emotional style transfer for unseen emotions!**

**Analysis of deep emotional features**

- Recent advances of deep learning have led to a shift from human-crafted representations to deep features learnt by neural network;
- As shown in Fig. 1, deep emotional features form clear emotion groups in terms of feature distribution. It suggests the potential to use deep emotional features as the emotion descriptor to encode an emotion class.

**Experiments**

1) ESD: A new multi-speaker and multi-lingual emotional speech dataset
- 350 parallel utterances spoken by Mandarin and English speakers
- For each language, there are 5 male and 5 female speakers in 5 emotions: a) happy, b) sad, c) neutral, d) angry, e) surprise
- During conversion, we use one universal model to conduct emotion conversion from neutral to both seen emotions (happy, sad) and unseen emotion (angry);
- Baseline: VAW-GAN-EVC [2]: one-hot emotion label, one-to-one conversion

2) Objective Evaluation

**3) Subjective Evaluation**

**Conclusions**

- We propose to build a one-to-many emotional style transfer framework that does not require parallel data
- We propose to leverage deep emotional features from SER to describe emotional prosody in a continuous space
- By conditioning the decoder with controllable attributes such as deep emotional features and F0 values, we achieve competitive results for both seen and unseen emotions over the baseline framework
- We release a multi-speaker and multi-lingual emotional speech dataset

**References**
