Mixed-EVC: Mixed Emotion Synthesis and Control in Voice Conversion

Kun Zhou\textsuperscript{1}, Berrak Sisman\textsuperscript{2}, Carlos Busso\textsuperscript{2}, Bin Ma\textsuperscript{1}, Haizhou Li\textsuperscript{3,4}

\textsuperscript{1} Institute of Intelligent Computing, Alibaba Group, Singapore
\textsuperscript{2} The University of Texas at Dallas, United States of America
\textsuperscript{3} The Chinese University of Hong Kong, Shenzhen, China
\textsuperscript{4} National University of Singapore, Singapore

Abstract

Emotional voice conversion (EVC) traditionally targets the transformation of spoken utterances from one emotional state to another, with previous research mainly focusing on discrete emotion categories. We propose a novel EVC framework, Mixed-EVC, which only leverages discrete emotion training labels. We construct an attribute vector that encodes the relationships among these discrete emotions, which is predicted using a ranking-based support vector machine and then integrated into a sequence-to-sequence (seq2seq) EVC framework. Mixed-EVC not only learns to characterize the input emotional style but also quantifies its relevance to other emotions during training. As a result, users have the ability to assign these attributes to achieve their desired rendering of mixed emotions. Objective and subjective evaluations confirm the effectiveness of our approach in terms of mixed emotion synthesis and control while surpassing traditional baselines in the conversion between discrete emotions.

1. Introduction

Human speech often encompasses a blend of emotions, resulting in the emergence of complex emotional expressions, as evidenced in prior studies \cite{1, 2, 3}. Emotional voice conversion (EVC) aims to manipulate the emotional state of a spoken utterance while keeping speaker identity and linguistic content unchanged \cite{4}. This paper represents a progressive step in the field of EVC, with a unique focus on infusing a quantifiable mixed emotion rendering into a human voice. The primary objective is to enhance the naturalness of human-computer interactions \cite{5}, for example, enriching the emotional responses within a dialogue system \cite{6, 7}.

EVC poses unique challenges due to the complex structure of emotions \cite{8}. People use different words to describe the emotions that they feel, showing that there are nearly 34,000 distinct emotions that a human may experience \cite{9}. To understand how these emotions correlate with each other, scientists analyze them in a valence-arousal space \cite{8}. The evidence for the valence-arousal view comes from statistical analysis of how people report feelings \cite{10}. However, the analysis from a valence-arousal view has not always been able to tell us the real difference between emotions \cite{11}. Plutchik’s emotion wheel \cite{12} provides a more straightforward way to describe emotions. 8 primary emotions: anger, fear, sadness, disgust, surprise, anticipation, trust, and joy, are arranged in an emotion wheel. The diverse amount of emotions could be produced either by changing the intensity or by adding up the primary emotions. Although preliminary studies \cite{13, 14} have explored synthesizing mixed emotions for text-to-speech systems, we observe a lack of study on mixed emotion synthesis in the literature of EVC, with existing studies mostly focusing on the conversion between discrete emotions. In this research, we draw inspiration from the emotion wheel theory and introduce an approach that employs voice conversion techniques to manipulate human emotions into a mixed emotional state.

Speech emotions are inherently supra-segmental and intricate, complex with multiple acoustic cues such as speech quality, pitch, energy, and speaking rate \cite{15}. Addressing these complexities in EVC calls for the modeling of both spectral and prosodic variations at the same time, leading to the study of sequence-to-sequence (seq2seq) architecture for EVC \cite{16, 17, 18, 19}. To learn emotion information, existing EVC frameworks mostly leverage pre-defined discrete emotion labels as the supervision to emotion training, either learning a translation model between emotion pairs \cite{20, 21, 22, 23, 24}, or disentangling emotional elements with auto-encoders \cite{25, 26, 27}. Emotion conversion can be achieved by assigning an emotion label or transferring from an emotional speech. These methods restrict learning richer style descriptions of emotions but rather produce a stereotypical emotional pattern \cite{28}. Consequently, they confine emotions within specific categories, posing challenges when it comes to examining the connections between different emotional states, encompassing the entirety of human emotions, and creating a mixed emotional profile. This paper aims to fill these gaps.

This paper presents the first investigation of mixed emo-
The rest of the paper is organized as follows: We motivate our work by introducing the background in Section 2. We then introduce our proposed method in Section 3. In Section 4, we report our experimental results, and introduce our interactive demo in Section 5. Section 6 concludes the paper.

The Emotion Wheel [12], proposed by Robert Plutchik in 1980, illustrates the vast range of human emotions and their interrelationships. It suggests that combinations of eight primary emotions—joy, trust, fear, surprise, sadness, anticipation, anger, and disgust—can give rise to nuanced emotions. For example, emotions such as excitement, outrage, and disappointment can emerge from combinations of surprise and happiness, anger, and sadness, respectively, as illustrated in Figure 1. The Emotion Wheel acknowledges the complexity of emotional blends, bridging the gap between psychological emotion studies and computational speech emotion modeling. It highlights the challenges in speech emotion recognition and synthesis, emphasizing the need to consider a wider spectrum of emotions than traditionally accounted for. In our research, we integrate insights from the Emotion Wheel into voice conversion models, proposing a method to quantifiably render mixed emotions in human voice, which we will discuss next.

3. Mixed-EVC

In this section, we present the details of our proposed emotional voice conversion framework, Mixed-EVC. We begin by introducing a novel approach that facilitates the synthesis of mixed emotions (Sec. 3.1). Following that, we introduce the core components and the training diagram of Mixed-EVC (Sec. 3.2). Lastly, we describe how to render and control the desired mixed emotions (Sec. 3.3).

3.1. Modeling Emotion Attribute with Pairwise Ranking

Most emotional speech databases group utterances into several discrete categories thus restricting to learning of richer descriptions of the emotional style [4]. Instead of predicting the presence of a specific emotion, we propose to model the relative difference between a pair of emotions, which we define as “Emotion Attribute”. During the training, we characterize the input emotional style by measuring its relevance to other emotions by explicitly modeling emotion attributes. For example, while it is hard to give a consensus emotion label to an utterance, espe-
We employ the idea of relative attributes \[30\] to encode the relationship between discrete emotions into a quantifiable vector. Relative attributes were first proposed in computer vision \[30\] and later applied in speech processing \[31, 32, 33\]. Prior studies show that relative attributes provide a more detailed description of an image, thus surpassing traditional classifiers on image recognition tasks \[30\]. Inspired by that, we propose to learn a ranking function for each emotion attribute given relative similarity constraints on paired emotional samples. More specifically, we are given a training set \( T = \{x_n\} \), where \( x_n \) is the acoustic feature vector of the \( n^{th} \) training sample with \( M \) emotion labels. Our goal is to learn a ranking function \( f_m \) given below:

\[
f_m(x_n) = W_m x_n,
\]

where \( m = 1, ..., M \) and \( W \) is a weighting matrix. We construct an ordered set \( O_m \) and an unordered set \( U_m \) as the supervision sets, which satisfy the following constraints:

\[
\forall (x_i, x_j) \in O_m : W_m x_i > W_m x_j \tag{2}
\]

\[
\forall (x_i, x_j) \in U_m : W_m x_i = W_m x_j \tag{3}
\]

The sample \( x_i \) should have a stronger presence of emotion attribute \( a_m \) than \( x_j \) in the ordered set \( O_m \), while their presence should be similar in the unordered set \( U_m \). The weighting matrix \( W \) is estimated by solving a support vector machine (SVM) problem \[35\]. The emotion attribute thus corresponds to the distance from the decision boundary. We generate pairs for all emotions and iteratively conduct the aforementioned ranking training in a pairwise manner. During inference, the trained ranking function can automatically predict the emotion attribute, which signifies the relevance or similarity between two emotional styles. Each emotion attribute is standardized within the \([0, 1]\) range to be easily quantifiable, where a lower value indicates a greater similarity in emotional style. All these emotion attributes together compose an emotion attribute vector, quantifying the degree of relevance between an input emotional style and all other emotions. Next, we explain how those attributes contribute to the EVC training.

### 3.2. Seq2Seq Emotional Voice Conversion Training

We integrate our emotion attribute vector into a seq2seq-based emotional voice conversion framework, as illustrated in Figure \[2\]. Our approach, distinct from traditional frame-wise modeling \[19\], jointly models feature mapping and alignment and automatically predicts speech duration during run-time.

Given an input speech, the linguistic encoder is trained to predict a sequence of linguistic embeddings (“Linguistic Modeling”), while the emotion encoder focuses on encoding the emotional style into an utterance-level emotion embedding (“Style Modeling”). The pre-trained relative functions further predict an emotion attribute vector that signifies the relevance of the input emotional style to other emotions (“Relevance Modeling”). This emotion attribute vector is then transformed using a fully connected (FC) layer, resulting in a relative embedding. Ultimately, the decoder learns to reconstruct the input emotional style by combining information from both the emotion and relative embeddings. Our framework not only explicitly characterizes the input emotional style but also establishes a relationship with other emotions.

To overcome the instability issues in seq2seq models, we conduct emotional training with the following strategies: (1) introducing text supervision \[35\], and (2) pre-training with a large neutral-speaking corpus \[16\]. We use text transcriptions.
Table 1: MOS with 95% confidence interval to evaluate the speech quality of synthesized mixed emotions.

<table>
<thead>
<tr>
<th>Configuration</th>
<th>MOS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mixing Angry with Surprise</td>
<td>4.78 ± 0.17</td>
</tr>
<tr>
<td>+ 0% Surprise</td>
<td>3.50 ± 0.24</td>
</tr>
<tr>
<td>+ 30% Surprise</td>
<td>3.34 ± 0.26</td>
</tr>
<tr>
<td>+ 60% Surprise</td>
<td>3.15 ± 0.34</td>
</tr>
<tr>
<td>+ 90% Surprise</td>
<td>3.20 ± 0.31</td>
</tr>
<tr>
<td>Mixing Happy with Surprise</td>
<td>4.85 ± 0.12</td>
</tr>
<tr>
<td>+ 0% Surprise</td>
<td>3.63 ± 0.27</td>
</tr>
<tr>
<td>+ 30% Surprise</td>
<td>3.53 ± 0.19</td>
</tr>
<tr>
<td>+ 60% Surprise</td>
<td>3.17 ± 0.30</td>
</tr>
<tr>
<td>+ 90% Surprise</td>
<td>3.05 ± 0.34</td>
</tr>
<tr>
<td>Mixing Sad with Surprise</td>
<td>4.74 ± 0.12</td>
</tr>
<tr>
<td>+ 0% Sad</td>
<td>3.17 ± 0.35</td>
</tr>
<tr>
<td>+ 30% Sad</td>
<td>3.48 ± 0.30</td>
</tr>
<tr>
<td>+ 60% Sad</td>
<td>3.42 ± 0.32</td>
</tr>
<tr>
<td>+ 90% Sad</td>
<td>3.08 ± 0.36</td>
</tr>
</tbody>
</table>

3.3. Mixed Emotion Rendering and Control

During run-time, Mixed-EVC carries out the conversion of the source emotion into an emotional mixture, illustrated in Figure 3. This process involves several steps: first, the linguistic encoder transforms the source linguistic content into internal representations. Next, the emotion encoder encapsulates a set of reference speech into an utterance-level emotion embedding. Additionally, attributes of other emotion types can be introduced by manually specifying an attribute vector. This flexibility enables us to adjust the proportions of each emotion type, thereby enabling the creation of diverse emotional mixtures.

4. Experiments

4.1. Experimental Setup

We conduct experiments with the ESD dataset [4]. There are many speakers, where we randomly choose one female speaker ("0019") with five emotions ("Neutral", "Happy", "Sad", "Angry" and "Surprise"). We follow the data partition in the ESD dataset, and for each emotion, we use 300, 30, and 20 utterances for training, testing, and evaluation, respectively. We train a universal EVC model for all emotions. At run-time, we convert Neutral to 4 different mixtures of emotions that are:

- **Outrage, Excitement, Disappointment:** where we convert Neutral to Angry, Happy and Sad (Neutral-Angry, Neutral-Happy, Neutral-Sad) respectively while introducing different percent-ages of Surprise into the mixture. These 3 mixtures have been studied in emotion theory [12].
- **Bittersweet:** where we convert Neutral to Happy while introducing different percentages of Sad. Even though happiness and sadness are two oppositely valenced emotions in Russell’s model [36], there are some debates that agree with the co-existence of happiness and sadness [2].

The training pipeline is described below: We first pre-train the relative ranking function between each emotion pair. Each utterance is represented by a 384-dimensional feature vector defined by the INTERSPEECH Emotion Challenge [37], which is used to train the ranking functions. The ranking functions achieved 98% accuracy in classifying the emotion categories on the test set of the ESD dataset. We then follow a two-stage training strategy [16] to train our proposed framework: (1) Style Pre-training with the VCTK Corpus [38], and (2) Seq2Seq EVC Training with the ESD dataset. The inputs to the EVC model are acoustic features represented by an 80-dimensional Mel-spectrogram extracted every 12.5 ms with a frame size of 50 ms for the short-time Fourier transform (STFT) and phoneme sequences converted by the Festival [39] G2P tool. It should be noted that we only use acoustic features as inputs during the conversion.

Our proposed framework has a similar structure to the model presented in [35]. The linguistic encoder comprises an encoder, a 2-layer 256-cell BLSTM, and a decoder employing a 1-layer 512-cell BLSTM with an attention layer followed by a fully connected (FC) layer with an output channel of 512. The decoder has the same model architecture as that of Tacotron [40]. The text encoder is a 3-layer 1D CNN with a kernel size of 5 and a channel of 512. The style encoder is a 2-layer, 128-cell
Mixed-EVC (proposed) | No Difference | Seq2Seq-EVC
--- | --- | ---
Neu-Sur | 66.70% | 14.30% | 19.00%
Neu-Sad | 61.90% | 9.50% | 28.60%
Neu-Hap | 52.40% | 23.80% | 23.80%
Neu-Ang | 57.10% | 14.30% | 28.60%

Figure 5: Prereference test results of emotion similarity between the proposed Mixed-EVC and the baseline Seq2Seq-EVC over one-to-one emotion conversion.

BLSTM followed by an FC layer with an output channel of 64. The classifier is a 4-layer FC with 512, 512, 512, and 5 channels. For training, we specify a batch size of 64 for style pre-training and 4 for emotion training. During style pre-training, the learning rate is set at 0.001, and a weight decay of 0.0001 is applied. For emotion pre-training, we adopt a learning rate halving strategy every 7 epochs.

4.2. Objective Evaluation

We assess the quality of our synthesized mixed emotions by employing a speech emotion recognition (SER) model pre-trained on the ESD dataset. This SER model shares a similar structure to the one presented in [41], encompassing components such as a 3-D CNN layer, a BLSTM layer, an attention layer, and an FC layer. Our analysis leverages the classification probabilities obtained from the final softmax layer of the SER. We contend that these classification probabilities serve as a concise summary of emotional cues aggregated from preceding layers, aiding in decision-making and providing us with a valuable tool for investigating emotional blends. We report the classification probabilities in Figure 4. We first evaluate three different mixed effects that are Outrage, Excitement and Disappointment, where we convert Neutral to Angry, Neutral to Happy and Neutral to Sad respectively. These transformations are done while gradually increasing the percentage (0%, 30%, 60% and 90%) of Surprise. As shown in Figure 4(a), (b), and (c), we observe that the probability of Surprise consistently increases when we increase the percentage of Surprise from 0% to 90%. The likelihoods of Angry, Happy, and Sad consistently maintain the highest in three different emotional mixtures. This pattern arises from their distinct characterization by the emotion encoder, establishing them as the predominant emotional components within the mixture. We then evaluate the mixed effect of Bittersweet as shown in Figure 4(d), where Sad is further introduced when we convert Neutral to Happy. As shown in Figure 4(d), we find a similar observation as in Figure 4(a), (b) and (c). These observations indicate that mixed emotions can be objectively recognized by a pre-trained SER.

4.3. Subjective Evaluation

We then conduct listening experiments to evaluate our synthesized results in terms of speech quality and emotion perception. 15 subjects participated in all the experiments and each of them listened to 112 synthesized utterances in total.

We first report mean opinion scores (MOS) results for speech quality, where all participants are asked to listen to the reference speech (“Ground truth”) and the synthesized speech with mixed emotions and score the “quality” of each speech sample on a 5-point scale (‘5’ for excellent, ‘4’ for good, ‘3’ for fair, ‘2’ for poor, and ‘1’ for bad). As shown in Table 1, the audio quality of synthesized speech slightly decreases as we increase the percentage of emotional mixture. Nonetheless, the quality remains between fair and good across the board.

We then conduct best-worst scaling (BWS) tests to evaluate the emotion perception of the synthesized mixed emotions. All participants are asked to choose the best and the worst emotion according to their perception of the mixed emotion (Outrage, Excitement, Disappointment and Bittersweet). As shown in Table 2(a), (b) and (c), we observe that most participants can perceive the mixed feelings and choose those with 90% of Surprise as the “Best”, and those of 0% of Surprise as the “Worst”. We take one step further to evaluate the perception of Bittersweet as shown in Table 2(d). Most participants chose those with 90% Sad as the “Best” confirming the strength of our approach. The results of Bittersweet are not as distinguishable as those of the other three mixed emotions, which suggests that it could be more challenging for listeners to perceive a Bittersweet feeling. These results show that we are able to create new emotional feelings that are subtle but do not exist in the database. We also show the effectiveness of controllability by varying the percentages of primary emotions to synthesize different emotional mixtures.

We further conduct a preference test to evaluate the performance of Mixed-EVC on one-to-one emotion conversion. We choose Seq2Seq-EVC [16] as the baseline. Participants are asked to choose the best one in terms of the emotion similarity with the reference emotion. As shown in Figure 5, our proposed Mixed-EVC outperforms the baseline Seq2Seq-EVC in all conversion pairs. These results show that Mixed-EVC could not only synthesize emotional mixtures but also have superior performance on emotion conversion.

5. An Interactive Demo - Emotion Triangle

Mixed emotion synthesis is a key method to model transitions between discrete emotional states. In our application study, we developed an “Emotion Triangle” as an interactive demo for users to explore these transitions. This model challenges us to accurately represent the internal states that occur between distinct emotions. Our approach, the Mixed-EVC, enables the synthesis of these intermediary states by blending various percentages of basic emotions—happiness, anger, sadness, and surprise. This tool not only aids in understanding how different feelings can combine to form complex, mixed emotional states.
but also demonstrates the effectiveness of Mixed-EVC in controlling these blends. As shown in Figure 6, each corner of the triangle illustrating a specific emotional mixture, representing excitement, outrage, and disappointment, thereby further validating the controllability of Mixed-EVC. For a practical exploration of these concepts, we encourage readers to visit the online demo.

6. Conclusion

We introduce Mixed-EVC, a seq2seq emotional voice conversion (EVC) framework, to address the existing research gap in mixed emotion synthesis within the context of EVC. We formulate emotional styles as an attribute and explicitly model the degree of relevance between different emotions through a ranking-based SVM. By manually adjusting relevance at runtime, Mixed-EVC could produce different emotional mixtures. Both objective and subjective evaluations show the effectiveness of synthesizing different mixed emotions. The speech samples are publicly available.

7. References


