



## Emotional voice conversion: Theory, databases and ESD

Kun Zhou<sup>a,\*</sup>, Berrak Sisman<sup>b</sup>, Rui Liu<sup>b</sup>, Haizhou Li<sup>a,c</sup>

<sup>a</sup> Department of Electrical and Computer Engineering, National University of Singapore, Singapore

<sup>b</sup> Information Systems Technology and Design Pillar, Singapore University of Technology and Design, Singapore

<sup>c</sup> Machine Listening Lab, University of Bremen, Germany

### ARTICLE INFO

#### Keywords:

Emotional voice conversion  
Emotional speech database (ESD)  
Overview  
Voice conversion

### ABSTRACT

In this paper, we first provide a review of the state-of-the-art emotional voice conversion research, and the existing emotional speech databases. We then motivate the development of a novel emotional speech database (ESD) that addresses the increasing research need. With this paper, the ESD database<sup>1</sup> is now made available to the research community. The ESD database consists of 350 parallel utterances spoken by 10 native English and 10 native Chinese speakers and covers 5 emotion categories (neutral, happy, angry, sad and surprise). More than 29 h of speech data were recorded in a controlled acoustic environment. The database is suitable for multi-speaker and cross-lingual emotional voice conversion studies. As case studies, we implement several state-of-the-art emotional voice conversion systems on the ESD database. This paper provides a reference study on ESD in conjunction with its release.

### 1. Introduction

Speech is more than lexical words and carries the emotion of the speakers. Previous research (Mehrabian and Wiener, 1967) indicates that the spoken words, which are referred to as verbal attitude, convey only 7% of the message when communicating feelings and attitudes. Non-verbal vocal attributes (38%) and facial expression (55%) have a major impact on the expression of social attitudes. Non-verbal vocal attributes reflect the emotional state of a speaker and play an important role in daily communication (Arnold, 1960). Emotional voice conversion (EVC) is a technique that aims to convert the emotional state of the utterance from one to another while preserving the linguistic information and speaker identity, as shown in Fig. 1(a). It allows us to project the desired emotion into a human voice, for example, to act or to disguise one's emotions. Speech synthesis technology has been progressing by leaps and bounds towards human-like naturalness. However, it has not adequately equipped with human-like emotions (Schuller and Schuller, 2018). Therefore, emotional voice conversion bears huge application potential in human-machine interaction (Tits et al., 2020; Pittermann et al., 2010; Crumpton and Bethel, 2016; Li et al., 2019), for example, to enable robots to respond to humans with emotional intelligence.

In general, voice conversion aims to convert the speaker identity in human speech while preserving its linguistic content, which is also referred to as speaker voice conversion (Childers et al., 1989), as shown in Fig. 1(b). Earlier studies of voice conversion are focused on modeling

the mapping between source and target features with some statistical methods, which include Gaussian mixture model (GMM) (Toda et al., 2007), partial least square regression (Helander et al., 2010), frequency warping (Erro et al., 2009a) and sparse representation (Çişman et al., 2017; Sisman et al., 2018b, 2019b). Deep learning approaches, such as deep neural network (DNN) (Chen et al., 2014; Lorenzo-Trueba et al., 2018a), recurrent neural network (RNN) (Nakashika et al., 2014), generative adversarial network (GAN) (Sisman et al., 2018c) and sequence-to-sequence model with attention mechanism (Zhang et al., 2019b; Huang et al., 2019) have advanced the state-of-the-art. For effective modeling, parallel training data are required in general. Recently, voice conversion techniques with non-parallel training data have been studied, for example, domain translation (Kaneko and Kameoka, 2017; Kameoka et al., 2018), multitask learning (Zhang et al., 2019a) and speaker disentanglement (Hsu et al., 2016, 2017; Tanaka et al., 2019) are among the successful attempts. The recent progress in voice conversion becomes the source of inspiration for emotional voice conversion studies.

Both speaker voice conversion and emotional voice conversion aim to preserve the speech content and to convert para-linguistic information. In speaker voice conversion, the speaker identity is thought to be characterized by the speaker's physical attributes, which are determined by the voice quality of the individual. Thus, spectrum conversion has been the focus (Kain and Macon, 1998; Desai et al.,

\* Corresponding author.

E-mail addresses: [zhoukun@u.nus.edu](mailto:zhoukun@u.nus.edu) (K. Zhou), [berraksisman@u.nus.edu](mailto:berraksisman@u.nus.edu) (B. Sisman), [liu\\_rui@sutd.edu.sg](mailto:liu_rui@sutd.edu.sg) (R. Liu), [haizhou.li@nus.edu.sg](mailto:haizhou.li@nus.edu.sg) (H. Li).

<sup>1</sup> <https://hltssingapore.github.io/ESD/>.

2010; Sisman, 2019). It is considered that prosody-related features are speaker-independent, that are to be carried forward from the source to the target (Liu et al., 2020a). However, in emotional voice conversion, emotion is inherently supra-segmental and complex that involves both spectrum and prosody (Arias et al., 2020; Schuller and Schuller, 2020). Therefore, it is insufficient to convert the emotion through frame-wise spectral mapping (Erro et al., 2009b). Segmental level prosodic dynamics needs to be properly dealt with (Tao et al., 2006).

The early studies on emotional voice conversion include GMM technique (Aihara et al., 2012; Kawanami et al., 2003), sparse representation technique (Aihara et al., 2014), or an incorporated framework of hidden Markov model (HMM), GMM and fundamental frequency (F0) segment selection method (Inanoglu and Young, 2009). The recent studies on deep learning have seen remarkable performance, such as DNN (Lorenzo-Trueba et al., 2018a; Luo et al., 2016, 2017), highway neural network (Shankar et al., 2019a), deep bi-directional long-short-term memory network (DBLSTM) (Ming et al., 2016a), and sequence-to-sequence model (Robinson et al., 2019; Le Moine et al., 2021). Beyond parallel training data, new techniques have been proposed to learn the translation between emotional domains with CycleGAN (Zhou et al., 2020b; Shankar et al., 2020b) and StarGAN (Rizos et al., 2020), to disentangle the emotional elements from speech with auto-encoders (Zhou et al., 2021b; Gao et al., 2019; Cao et al., 2020; Schnell et al., 2021), and to leverage text-to-speech (TTS) (Kim et al., 2020; Zhou et al., 2021a) or automatic speech recognition (ASR) (Liu et al., 2020b). Such framework generally works well in speaker-dependent tasks. Studies have also revealed that the emotions can be expressed through some universal principles that are shared across different individuals and cultures (Ekman, 1992; Kane et al., 2014; Manokara et al., 2020), that motivates the study of multi-speaker (Shankar et al., 2019, 2020; Liu et al., 2020b), and speaker-independent emotional voice conversion (Zhou et al., 2020c; Choi and Hahn, 2021).

Speaker-independent emotional voice conversion studies call for a large multi-speaker emotional speech database (Zhou et al., 2020c). However, existing databases, such as VCTK database (Yamagishi et al., 2019), CMU-Arctic database (Kominek and Black, 2004), and Voice Conversion Challenge (VCC) corpus (Toda et al., 2016; Lorenzo-Trueba et al., 2018b; Zhao et al., 2020), are designed for speaker voice conversion. The study of emotional voice conversion is facing the lack of a suitable database. To our best knowledge, there is no large-scale, multi-speaker, and open-source emotional speech database just yet. The lack of data greatly limits the emotional voice conversion research (Schuller et al., 2013). In addition, an overview of existing databases in Section 3 identifies that imperfect recording condition and limited lexical variability are other limitations. Based on the review of the current emotional voice conversion research landscape, we propose the development of a novel database to address the needs.

This paper is organized as follows. We first give a comprehensive overview of recent studies on emotional voice conversion in Section 2. We discuss the 19 existing databases (Zhou et al., 2021c) in Section 3. We then formulate the design of a novel *ESD* database for speaker-independent emotional voice conversion, that is also suitable for other speech synthesis tasks, such as mono-lingual or cross-lingual speaker voice conversion and emotional text-to-speech. The *ESD* database consists of a total of 29 h of audio recordings from 10 native English speakers and 10 native Chinese speakers, covering 5 different emotion categories (neutral, happy, angry, sad and surprise). It represents one of the largest emotional speech databases publicly available, in terms of speaker and lexical variability. All the recordings are conducted in the studio with professional devices to guarantee audio quality. In Section 5, we report the performance for the *ESD* database in emotional voice conversion. In Section 6, we show the promise of the *ESD* database on other voice conversion and text-to-speech tasks. Finally, Section 7 concludes the study.

## 2. Overview of emotional voice conversion

Emotional voice conversion is the enabling technology for many applications, such as expressive text-to-speech (Liu et al., 2020c; Tits et al., 2019a), speech emotion recognition (SER) (Bao et al., 2019; Rizos et al., 2020) and conversational agents (Tits et al., 2020). It has attracted much attention in recent years. Nearly one-third of the research papers under the prosody and emotion category at INTERSPEECH 2020 is about emotional voice conversion. As most of the state-of-the-art emotional voice conversion frameworks are based on data-driven techniques, they hugely depend on the training data (Inanoglu and Young, 2009; Zhou et al., 2020c; Elgaar et al., 2020). We start with an introduction to emotion in speech. Then, we give a comprehensive overview of emotional voice conversion studies according to the type of training data.

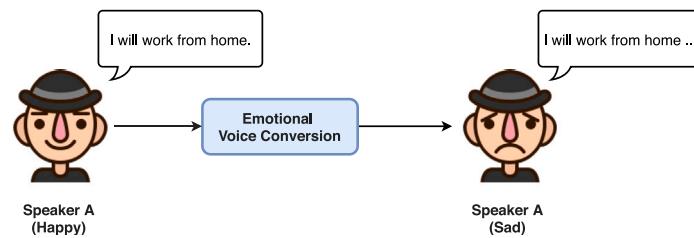
### 2.1. Emotion in speech

It is well known that emotion can be rendered in both the acoustic speech and spoken linguistic content (Schuller and Schuller, 2020). Psychological studies also show that acoustic attitudes, cf. emotional prosody, play a bigger role than spoken words in human-human communication (Mehrabian and Wiener, 1967). Since the emotional prosody is the complex interplay of multiple acoustic attributes, it is not straightforward to precisely characterize speech emotion (Hirschberg, 2004). When modeling the emotion, two research problems are generally considered: (1) how to describe and represent emotions; and (2) how to model the process of emotional expression and perception by humans.

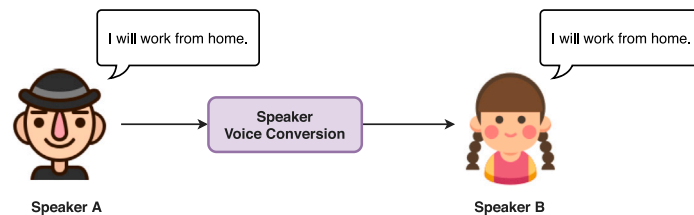
#### 2.1.1. Emotion representation

Emotion can be characterized with either categorical (Whissell, 1989; Ekman, 1992) or dimensional representations (Russell, 1980; Schroder, 2006). With emotion-denoting labels, the emotion category approach is the most straightforward way to represent emotion. One of the most well-known categorical approaches is Ekman's six basic emotions theory (Ekman, 1992), where emotions are grouped into six discrete categories, i.e., angry, disgust, fear, happy, sad, and surprise, that are adopted in many emotional speech synthesis studies (Schröder, 2009; Barra-Chicote et al., 2010; Zhou et al., 2020c). However, such discrete representations do not seek to model the subtle differences between human emotions (Nose et al., 2007; Tao et al., 2006; Schuller and Schuller, 2018) so as to control the rendering of voice. Another way is to model the physical properties of emotion expression. One example is Russell's circumplex model (Russell, 1980; Soleymani et al., 2017), which is defined by arousal, valence, and dominance (Gunes and Schuller, 2013). For example, in the valence-arousal (V-A) representation (Xue et al., 2018), happy speech is characterized by positive valence and arousal values, while sad speech is characterized by all negative values. On the other hand, angry can be divided into hot and cold angry, which correspond to the full-blown angry and milder angry in psychology (Biaassoni et al., 2016).

In general, both categorical and dimensional representations of emotions have been widely used in both emotion recognition (Schuller, 2018) and emotional voice conversion (Luo et al., 2019; Zhou et al., 2020b; Ming et al., 2015; Xue et al., 2018). The study on representation learning (Liu et al., 2020c) represents a new way of emotion representation, that further calls for large scale emotion-annotated speech data.



(a) A typical emotional voice conversion system is trained with speech data from the same speaker with different emotions. At run-time, the system converts the emotional state from one to another [45, 48].



(b) A typical speaker voice conversion system is trained with speech data from different speakers. At run-time, the system converts the speaker identity from one to another [69, 18].

Fig. 1. A comparison between emotional voice conversion and speaker voice conversion.

### 2.1.2. Emotion perception and production

Another research problem is how to model the process of emotional expression and perception. In Brunswik's model (Brunswik, 1956), it is suggested that the perception of emotion is multi-layered (Huang and Akagi, 2008). This model has been widely used in speech emotion recognition (Li and Akagi, 2016; Elbarougy and Akagi, 2014), where emotion categories, semantic primitives and acoustic features constitute the layers from top to bottom, respectively. Suppose that emotion production is an inverse process of emotion perception, an inverse three-layered model was studied for speech emotion production (Xue et al., 2018). The idea is further explored in emotional voice conversion (Xue et al., 2018).

Prosodic features concerning voice quality, speech rate, and fundamental frequency (F0), such as spectral features, duration, F0 contour and energy envelope are commonly used as the acoustic features in emotion-related studies (Schroder, 2006; Erickson, 2005). In emotional voice conversion, we are interested in transforming such acoustic features to render the emotions (Luo et al., 2016; An et al., 2017; Zhou et al., 2021b). In emotion recognition, we also rely on similar acoustic features (Schuller and Schuller, 2020; Latif et al., 2020), for example, expert-crafted features such as those extracted with openSMILE (Eyben et al., 2010, 2013), and learnt acoustic features from the spectrum (Ghosh et al., 2016). In emotional speech synthesis, both rule-based (Benesty et al., 2007; Zovato et al., 2004), and data-driven techniques with statistical modeling (Barra-Chicote et al., 2010; Yamagishi et al., 2005) or deep learning methods (An et al., 2017; Lorenzo-Trueba et al., 2018a) rely on speech databases for emotion analysis and production.

With the advent of deep learning, it is possible to describe different emotional styles in the continuous space with deep features learnt by neural networks (Zhou et al., 2021c; Schuller and Schuller, 2020; Tits et al., 2019b). Unlike human-crafted features, deep emotional features are less dependent on human knowledge (Schuller and Schuller, 2020), thus more suitable for emotional style transfer (Zhou et al., 2021c). Recently, deep emotional features are used to project an emotional style to a target utterance for emotional voice conversion (Zhou et al., 2021c). In this paper, we also visualize and evaluate the deep emotional features extracted from ESD database learnt by neural networks in Section 4.4. The studies on emotion analysis, recognition and synthesis have been inspiring emotional voice conversion studies.

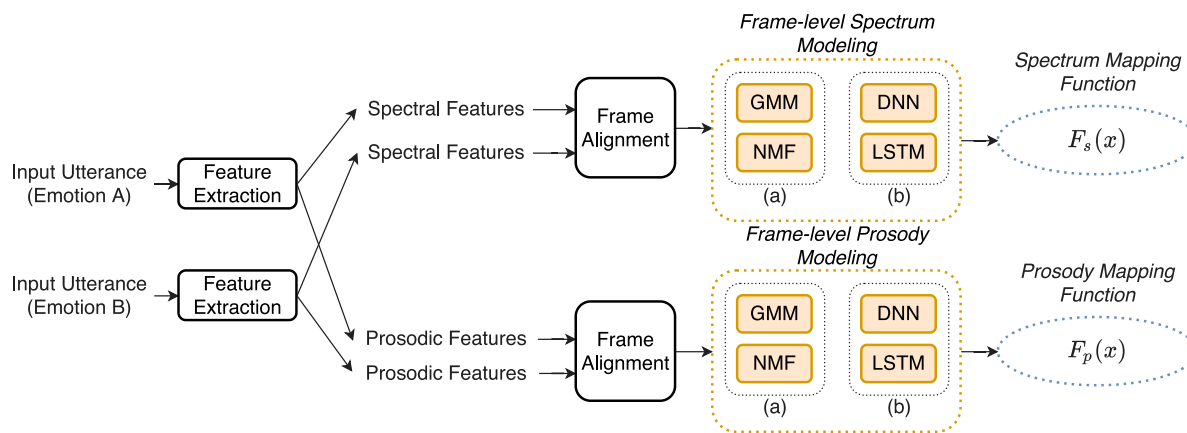
### 2.2. Emotional voice conversion with parallel data

Early studies on emotional voice conversion mostly rely on parallel training data, i.e., a pair of utterances of the same content by the same speaker but with different emotions. During training, the conversion model learns a mapping from the source emotion A to the target emotion B through the paired feature vectors. In general, as shown in Fig. 2, an emotional voice conversion process typically consists of three steps, i.e. feature extraction, frame alignment and feature mapping, where dynamic time warping (DTW) (Müller, 2007) and model-based alignment with speech recognizer (Ye and Young, 2004) or attention mechanism (Zhang et al., 2019b; Tanaka et al., 2019) are commonly used for frame alignment.

#### 2.2.1. Feature extraction and modeling

Given the paired utterances, feature extraction seeks to obtain features that characterize emotion in speech. The resulting paired feature sequences are aligned to obtain frame-level alignment. It is common to use low-dimensional spectral representations from the high-dimensional spectra for modeling and manipulation. The commonly used spectral features include Mel-cepstral coefficients (MCC), linear predictive cepstral coefficients (LPC), and line spectral frequencies (LSF).

As emotion is complex with multiple signal attributes concerning spectrum and prosody, both spectral and prosodic components need the same level of attention in emotional voice conversion. Several prosodic features are usually considered, such as pitch, energy, and duration. We note that F0 is an essential prosodic component that describes intonation, either linguistic or emotional, at different duration ranging from syllable to utterance (Veaux and Rodet, 2011). The methods to model the F0 variants include stylization methods (Teutenberg et al., 2008; Obin and Beliao, 2018) and multi-level modeling (Wang et al., 2017; Obin et al., 2011). As a multi-level modeling method, continuous wavelet transform (CWT) has been widely used to model hierarchical prosodic features, such as F0 (Suni et al., 2013; Ming et al., 2015; Luo et al., 2017) and energy contour (Şişman et al., 2017; Sisman and Li, 2018b; Sisman et al., 2019b). With CWT analysis, a signal can be decomposed into frequency components and represented with different temporal scales. CWT has shown to be effective for modeling the speech prosody (Ming et al., 2016b; Suni et al., 2013), and has been successfully applied into various emotional voice conversion



**Fig. 2.** A frame-level parallel emotional voice conversion framework with separate spectrum and prosody training using: (a) statistical methods, or (b) deep learning methods, to derive spectrum mapping function  $F_s(x)$  and prosody mapping function  $F_p(x)$ . The yellow boxes represent the models involved in the training. The dotted box (a) includes examples of statistical approaches (Aihara et al., 2012, 2014), and (2) includes examples of deep learning approaches (Ming et al., 2015; Luo et al., 2016). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

frameworks such as exemplar-based (Ming et al., 2015) or other deep learning-based (Luo et al., 2019; Ming et al., 2016a; Zhou et al., 2020b) frameworks.

### 2.2.2. Feature mapping

In a typical flow of emotional voice conversion, feature mapping aims to capture the relationship between the source and target features. With aligned paired utterances, feature mapping can be achieved by a statistical model (Inanoglu and Young, 2007). In Tao et al. (2006) and Wu et al. (2009), researchers propose to use a classification and regression tree to decompose the pitch contour of the source speech into a hierarchical structure, then followed by GMM and regression-based clustering methods. In Aihara et al. (2012), a GMM-based emotional voice conversion framework is proposed to learn the spectrum and prosody mapping. An exemplar-based emotional approach is introduced in Aihara et al. (2014), where parallel exemplars are used to encode the source speech signal and synthesize the target speech signal. The idea is further extended to a unified exemplar-based emotional voice conversion framework (Ming et al., 2015) that learns the mappings of spectral features and CWT-based F0 features simultaneously.

The neural network approach represents the state-of-the-art solution to feature mapping. DNN (Lorenzo-Trueba et al., 2018a), deep belief network (DBN) (Luo et al., 2016), highway neural network (Shankar et al., 2019a), and DBLSTM (Ming et al., 2016a) are examples that perform spectrum and prosody mapping with parallel training utterances. It is noted that frame-level feature mapping does not explicitly handle the mapping of duration, which is however an important element of prosody. The sequence-to-sequence encoder–decoder architecture represents a solution to duration mapping (Sutskever et al., 2014; Vaswani et al., 2017). With the attention mechanism, neural networks learn the feature mapping and alignment during training and automatically predict the output duration during run-time inference. The encoder–decoder model (Robinson et al., 2019) is an example, where pitch and duration are jointly modeled. In general, the more accurate alignment would help to build a better feature mapping function, that also explains why these frameworks are all built with parallel data.

### 2.3. Emotional voice conversion with non-parallel data

In practice, parallel speech data is expensive and difficult to collect. Thus, emotional voice conversion techniques with non-parallel data are more suitable for real-life applications (Zhou et al., 2020b, 2021b; Schuller and Schuller, 2020). We refer *non-parallel* data to the multi-emotion utterances, that do not share the same lexical content across

emotions. Neural solutions have enabled many emotional voice conversion frameworks with non-parallel data. Next, we discuss two typical non-parallel methods, that are (1) auto-encoder (Kingma and Welling, 2013), and (2) CycleGAN (Zhu et al., 2017) methods. We summarize the ways to learn from non-parallel training data into three scenarios,

- Emotional domain translation,
- Disentanglement between emotional prosody and linguistic content,
- Leveraging TTS or ASR systems.

#### 2.3.1. Emotional domain translation

Emotion voice conversion with non-parallel training data learns the unique emotion patterns and project them to the new utterances. To achieve this, translation models between the source and target domains have been studied, such as GANs (Emir Ak et al., 2019; Ak et al., 2020a, 2019, 2020b), CycleGANs (Zhu et al., 2017), Augmented CycleGAN (Almahairi et al., 2018) and StarGAN (Choi et al., 2018) in computer vision problems.

As shown in Fig. 3, CycleGAN is based on the concept of adversarial learning (Goodfellow et al., 2014), that is to train a generative model to find an optimal solution through a min–max game between the generator and the discriminator. In general, a CycleGAN is incorporated with three losses, that are (1) adversarial loss, (2) cycle-consistency loss, and (3) identity mapping loss. The adversarial loss measures how distinguishable between the converted and target feature distributions. The cycle-consistency loss eliminates the need for parallel training data through circular conversion, and the identity-mapping loss helps preserve the linguistic content between the input and output (Kaneko and Kameoka, 2017; Fang et al., 2018; Kaneko and Kameoka, 2018).

In emotional domain translation (Zhou et al., 2020b), CycleGANs are used to model the spectrum and prosody mapping between the source and target. For example, an adaptation of CycleGAN blended with implicit regularization is proposed (Shankar et al., 2020b) to predict the F0 contour, where the generators are composed of a trainable and a static component. These frameworks have shown the effectiveness to convert a pair of emotions.

Extending from two domains to multiple domains, StarGAN (Choi et al., 2018) is introduced to learn a multi-domain translation, that works for many-to-many speaker voice conversion (Kameoka et al., 2018; Kaneko et al., 2019). StarGAN is an improvement of CycleGAN, which consists of one generator/discriminator pair parameterized by class, and a domain classifier to verify the class correctness. In this way, StarGAN allows for multi-class to multi-class domain conversion. In Rizos et al. (2020), a StarGAN is adopted to learn spectral



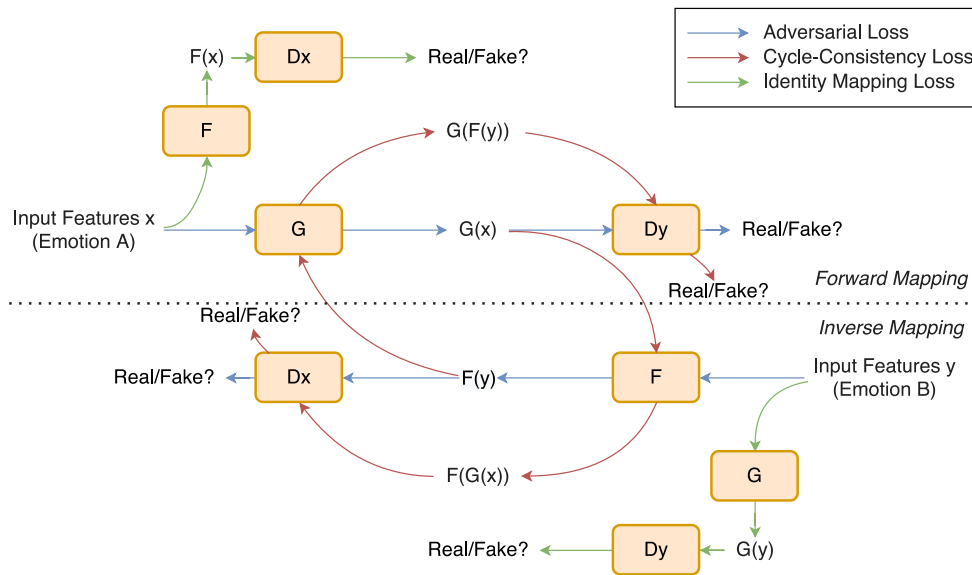


Fig. 3. A CycleGAN-based emotional voice conversion framework (Zhou et al., 2020b; Shankar et al., 2020b) learns both forward and inverse mappings between two different emotion domains through adversarial loss, cycle-consistency loss, and identity-mapping loss.

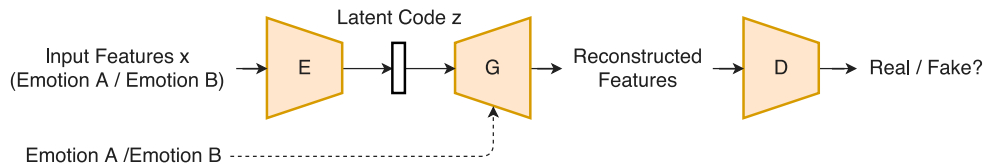


Fig. 4. An emotional voice conversion framework based on VAE-GAN (Cao et al., 2020) or VAW-GAN (Zhou et al., 2020c, 2021b,c) learns to disentangle emotional elements and recompose speech by assigning new emotional state. Emotion information (Emotion A / B) is used as the condition of the generator, which can be an one-hot emotion identity (Zhou et al., 2020c, 2021b; Cao et al., 2020), or a deep feature representation for emotion (Zhou et al., 2021c).

mappings between multiple emotional domains, and validated by data augmentation of emotion recognition. These emotional voice conversion studies represent the successful attempts to learn an emotion domain translation without the need for parallel data.

### 2.3.2. Disentanglement between emotional prosody and linguistic content

In the context of emotional voice conversion, speech can be considered as a composition of linguistic content and speaking style. If we were able to disentangle the emotional style from the linguistic content, we could modify the emotional style independently without changing the linguistic content. Auto-encoder (Kingma and Welling, 2013) is one of the techniques which is commonly used for speech disentanglement and reconstruction. Studies have shown the effectiveness of auto-encoder (Hsu et al., 2016) and its variants (Hsu et al., 2017; Chou et al., 2019; Wu and Lee, 2020) in disentangling the speaker information from the content, thus they are widely used in speaker voice conversion (Sisman et al., 2020) and singing voice conversion (Lu et al., 2020).

As shown in Fig. 4, a typical auto-encoder-based emotional voice conversion mainly consists of three components, those are (1) encoder, (2) generator, and (3) discriminator. The encoder learns to generate a latent code  $z$  frame-by-frame that is supposed to contain emotion-agnostic information such as linguistic content and speaker identity. The generator learns to reconstruct the input features from the latent code  $z$  and the emotion identity, the discriminator learns to distinguish the reconstructed features with the real features. Auto-encoder is easy to train owing to its simple architecture (Elgaar et al., 2020; Cao et al., 2020; Gao et al., 2019).

In practice, we can train an emotion classifier to derive deep feature representation from speech, i.e., deep emotional feature. The generator is then conditioned on such deep emotional feature or one-hot emotion

identity during the training. In this way, we learn the frame-level content-related representations from the speech in an unsupervised manner (Cao et al., 2020; Zhou et al., 2020c, 2021b). VAE-GAN (Cao et al., 2020) and VAW-GAN (Hsu et al., 2017; Zhou et al., 2020c) are successful attempts. Another idea is for an auto-encoder to learn an emotion-invariant latent code and an emotion-related style code in latent space (Gao et al., 2019). In doing so, a pair of content encoder and style encoder is used to disentangle the emotional style from the speaking content. In sum, auto-encoder is proven effective in learning disentangled representations, which makes training with non-parallel data possible.

### 2.3.3. Leveraging TTS or ASR systems

The voice conversion task shares many common attributes with the TTS task. For example, both require a high-quality speech vocoder. Studies also show that voice conversion benefits from the knowledge about linguistic content in the speech. For example, speaker voice conversion successfully leverages TTS (Zhang et al., 2019; Huang et al., 2019; Zhang et al., 2021) or ASR systems (Sisman et al., 2018a; Tian et al., 2019) that are phonetically-informed and trained on large speech corpus.

Inspired by the success in speaker voice conversion, multi-task learning between emotional voice conversion and text-to-speech is studied (Kim et al., 2020). In this framework, a single sequence-to-sequence model is trained to optimize both VC and TTS, in which the VC system benefits from the latent phonetic representation learnt by TTS during the training. At the run-time inference, given a reference speech, the system can transfer the emotional style of the input utterances from one to another. Experimental results also show that the VC and TTS system with multi-task learning outperforms a VC-alone system.

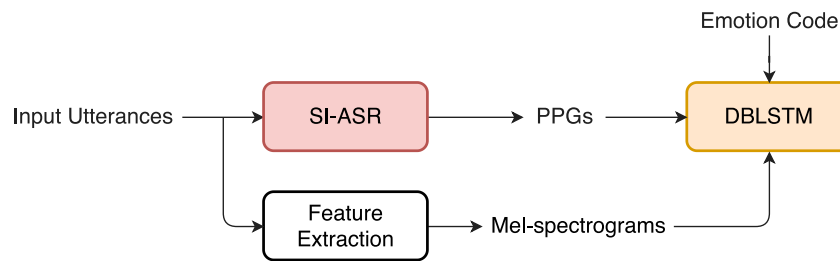


Fig. 5. An example of PPG-based emotional voice conversion framework that leverages from ASR (Liu et al., 2020b). A pre-trained speaker-independent automatic speech recognition (SI-ASR) is used to extract PPGs from the input utterances. Red box represents the model that is pre-trained. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

It is known that the prosody rendering of emotional speech consists of both speaker-dependent and speaker-independent, which makes it challenging to only learn emotion-related information from a multi-speaker database (Sisman and Li, 2018b; Zhou et al., 2020c). As shown in Fig. 5, phonetic posteriorgrams (PPGs) are utilized as the auxiliary input features for emotional voice conversion (Liu et al., 2020b). It is noted that PPGs are derived from an automatic speech recognition system, that are assumed to be speaker-independent and emotion-independent. With the phonetic information from PPGs, the proposed emotional voice conversion system demonstrates a better generalization ability with multi-speaker and multi-emotion speech data. In general, these frameworks have shown the effectiveness of multi-task learning or transfer learning, which represents a potential direction for future studies.

### 3. Databases in emotional voice conversion research

In this section, we present a survey of emotional speech databases available for development and evaluation for emotional voice conversion. We first give an overview of existing emotional speech databases in Table 1. Even though the given list of databases in Table 1 may not be exhaustive, it covers the most widely-used emotional speech databases to our best knowledge. There have been some surveys of existing emotional databases for speech emotion recognition that can be found in El Ayadi et al. (2011) and Akçay and Oğuz (2020). In this section, we would like to give a comprehensive overview of existing emotional speech databases from the standpoint of emotional voice conversion rather than emotion recognition. In order to motivate our ESD database, we provide an in-depth comparison of the state-of-the-art emotional speech databases in terms of lexical variability, language variability, speaker variability, confounding factors, and recording environment, as reported in Table 1. We also provide discussions and insights about the characteristics, strengths and weaknesses of these existing emotional speech databases in the following subsections.

#### 3.1. Lexical variability

It is known that the performance of speech synthesis and voice conversion systems is strongly dependent on the condition of the speech samples that are used during the training (Sisman et al., 2020). For both voice conversion and text-to-speech research, lexical content is especially important as it can affect the generalization ability of the speech synthesis system. In general, speech synthesis systems could benefit from the training data with a large lexical variety. However, in practice, such databases with a large variety of emotional utterances are not widely available, and most databases only contain limited utterances for each emotion.

We note that many of the existing emotional speech databases cover most of the emotion categories only with limited lexical variability. For example, the RAVDESS database (Livingstone and Russo, 2018) contains emotional speech data from 24 actors, and the actors are asked to read only 2 different sentences in a spoken and sung way.

Each utterance is repeated with 2 different intensities (except for the neutral emotion). Similar to the RAVDESS database, the CREMA-D database (Cao et al., 2014) contains 12 different sentences recorded by 91 actors. The Berlin Emotional Speech Dataset (EmoDB) (Burkhardt et al., 2005) collects 10 sentences read by 10 speakers with 7 emotions. The Surrey Audio-Visual Expressed Emotion (SAVEE) (Jackson and Haq, 2014) database covers 14 speakers with 7 different emotions but only contains 480 utterances in total. JL-Corpus (James et al., 2018) is a collection of 4 New-Zealand English speakers uttering 15 sentences in 5 primary emotions with 2 repetitions, and another 10 sentences in 5 secondary emotions. In general, these databases are all of good quality and can be ideal to build a rule-based emotional voice conversion framework. However, they might not be suitable for building a comprehensive model for emotional speech, especially with current deep learning approaches, due to their limited variety of sentences.

There are also some emotional speech databases designed for lexical control. In the Toronto Emotional Speech Set (TESS) (Pichora-Fuller and Dupuis, 2020), 2 speakers speak 200 words with 7 different emotions in the carrier phrase (“Say the word ...”). A recent VESUS database (Sager et al., 2019) is designed and released with over 250 distinct phrases, each read by ten actors in five emotional states. These databases mark valuable practice to understand the emotion variance in the word or phrase level, but may not be suitable to build a state-of-the-art emotional voice conversion framework that is usually data-driven. It is also known that emotion is hierarchical in nature and can be varying from different prosody levels, we believe it would be more ideal for emotional voice conversion research to study the emotion variance within the utterances (Zhou et al., 2020b; Xu, 2011; Latorre and Akamine, 2008). In general, due to the limited lexical variability, these databases may not be ideal to build an emotional voice conversion framework, especially a speaker-independent one, with state-of-the-art deep learning approaches. Therefore, it is needed to have an emotional speech database with a large lexical variability to boost the related studies, such as emotional voice conversion in a speaker-independent scenario.

#### 3.2. Language variability

Another limitation is the imbalanced language representation that can be observed from Table 1. We note that 9 databases out of 19 listed in Table 1 only contain English speech, while the others include German, French, Danish, Italian, Japanese and Chinese. Among them, 2 databases contain multilingual content of English and French, those are AmuS (El Haddad et al., 2017) and EmoV-DB (Adigwe et al., 2018). AmuS is a database which is designed to study the expression of amusement across English and French, and contains 3 h speech data in total. EmoV-DB contains emotional speech data performed by four English speakers and one Belgium French speaker in five emotion categories. Besides, there is an attempt to build multi-lingual emotional speech databases covering different languages including Chinese, English, Japanese, Korean and Russian (Wang et al., 2012). These

**Table 1**  
An overview of existing emotional speech databases.

Database	Language	Emotions	Size	Modalities	Content type	Remarks
RAVDESS (Livingstone and Russo, 2018)	EN	Neutral, happy, angry, sad, calm, fear, disgust	24 actors (12 male and 12 female) x 2 sentences x 8 emotions x 2 intensities (normal, strong)	Audio/Visual	Acted	Limited lexical variability
CREMA-D (Cao et al., 2014)	EN	Neutral, happy, angry, sad, fear, disgust	91 actors (48 male and 43 female) x 12 sentences x 6 emotions, 1 sentence is expressed with 3 intensities (low, medium, high), others are not specified	Audio/Visual	Acted	Limited lexical variability
EmoDB (Burkhardt et al., 2005)	GE	Neutral, happy, angry, disgust, bored	10 speakers (5 male and 5 female) x 10 sentences x 7 emotions	Audio/Visual	Acted	Limited lexical variability
MSP-IMPROV (Busso et al., 2016)	EN	Neutral, happy, angry, sad	12 actors, 9 h of data	Audio/Visual	Acted/Improvised	Contains external noise, and overlapping speech
IEMOCAP (Busso et al., 2008)	EN	Neutral, happy, angry, sad, frustrate	10 speakers (5 male and 5 female), 12 h of data	Audio/Visual	Acted/Improvised	Contains external noise, and overlapping speech
DES (Engberg et al., 1997)	DA	Neutral, happy, angry, sad, surprise	4 speakers (2 male and 2 female), 10 min of speech	Audio	Acted	Limited lexical variability
CHEAVD (Li et al., 2017)	CN	Neutral, happy, angry, sad, surprise, fear	219 speakers, two hours emotional segments from films and TV shows	Audio/Visual	Improvised	Contains external noise
CASIA (Zhang et al., 2008)	CN	Neutral, happy, angry, sad, surprise, fear	4 speakers (2 male and 2 female) x 500 utterances x 6 emotions	Audio/Visual	Acted	Limited speaker variability Not free to use
AmuS (El Haddad et al., 2017)	EN, FR	Neutral, amuse	4 speakers (3 male and 1 female), about 3 h of data	Audio	Acted	Only single emotion
EmoV-DB (Adigwe et al., 2018)	EN, FR	Neutral, amuse, angry, sleepy, disgust	4 English speakers (2 male and 2 female), 1 French speaker (male), about 5 h of data	Audio	Acted	Contains other non-verbal expressions (laughter, sigh) Incomplete public version Limited speaker variability
SAVEE (Jackson and Haq, 2014)	EN	Neutral, happy, sad, angry, surprise, disgust, fear	4 speakers (all male), 480 utterances in total	Audio/Visual	Acted	Limited speaker variability (all male) Limited lexical variability
VESUS (Sager et al., 2019)	EN	Neutral, happy, angry, sad, fear	10 actors (5 male and 5 female) x 250 distinct phrases	Audio	Acted	Phrases, not complete sentences
JL-Corpus (James et al., 2018)	EN	Neutral, happy, angry, sad, excite	4 speakers (2 male and 2 female) x 5 emotions x 15 sentences x 2 repetitions; 4 speakers (2 male and 2 female) x 6 variants (questions/interrogative, apologetic, encouraging/reassuring, concerned, assertive, anxiety) x 10 sentences x 2 repetitions	Audio	Acted	Limited speaker variability Limited lexical variability
eINTERFACE'05 (Martin et al., 2006)	EN	Angry, disgust, fear, happy, sad, surprise	42 speakers (34 male and 8 female) from 14 nationalities, 1116 video clips in total	Audio/Visual	Improvised	With different accents Limited lexical variability
TESS (Pichora-Fuller and Dupuis, 2020)	EN	Neutral, angry, sad, fear, disgust, surprise, pleased	2 speakers (female) x 200 words x 7 emotions (in the carrier phrase "Say the word _")	Audio	Acted	Limited speaker variability Limited lexical variability
DEMOs (Parada-Cabaleiro et al., 2020)	IT	Neutral, angry, happy, sad, fear, disgust, surprise, guilty	9365 emotional and 332 neutral samples produced by 68 native speakers (23 females, 45 males)	Audio	Improvised	Imbalanced setting for each emotion
EMOVO (Costantini et al., 2014)	IT	Neutral, disgust, fear, anger, joy surprise, sad	6 actors (3 male and 3 female) x 14 sentences x 7 emotions	Audio	Acted	Limited speaker variability Limited lexical variability
VAM Corpus (Grimm et al., 2008)	GE	Valence, activation, dominance	12 h of audio-visual recordings of a German TV talk show	Audio/Visual	Improvised	Contain other non-verbal expressions and external noise
JTES (Takeishi et al., 2016)	JP	Neutral, sad, joy, angry	23 h and 31 min of 50 spoken sentences of 5 emotions acted by 100 speakers (50 male and 50 female)	Audio	Acted	Not publicly available

Language abbreviations are used as follows: *EN* stands for English, *GE* stands for German, *DA* stands for Danish, *IT* stands for Italian, *JP* stands for Japanese, *FR* stands for French and *CN* stands for Chinese (Mandarin).

databases mark the valuable practice that allow for the study of emotion expression across different languages within the same database. There are also some emotional speech databases designed for those languages which are under-explored, such as Italian (Costantini et al., 2014), Finnish (Seppänen et al., 2003), Basque (Saratxaga et al., 2006) and Polish (Staroniewicz and Majewski, 2009). However, the imbalance of the emotional speech data between English and other languages still makes it tough to learn a cross-lingual emotion representation for emotional voice conversion.

In general, the omnipresence of English in existing databases is mainly due to historical reasons, and many other languages are considered as low-resource (Tits et al., 2019a). Since the lexical content is also constrained by the language, there have been some efforts to build a cross-lingual (Sisman et al., 2019a) or multi-lingual (Nekvinda and Dušek, 2020) speech synthesis system in current literature. However, the lack of multi-lingual emotional speech databases in the literature makes the related studies in emotional speech synthesis or emotional voice conversion more difficult, thus motivates us to build a multi-lingual emotional speech database that can achieve a balance between English and other languages, and suitable for a study of emotional expression across different languages.

### 3.3. Speaker variability

Emotional expression presents both speaker-dependent and speaker-independent characteristics at the same time (Arnold, 1960). Previous studies on emotion recognition (Kotti and Paternò, 2012) and emotional voice conversion (Zhou et al., 2020c) have shown a better generalization ability by learning a speaker-independent emotion representation over a large multi-speaker emotional corpus. These studies indicate the importance of a database with a large speaker variety for emotional speech synthesis (Choi et al., 2019) and emotional voice conversion (Zhou et al., 2020c).

We note that many existing emotional speech databases lack speaker variability, thus not optimal for the study of speaker-independent emotional voice conversion. For example, DES (Engberg et al., 1997) contains emotional speech data from four Danish speakers. Similar to DES, JL-Corpus (James et al., 2018) consists of emotional speech data from four New Zealand English speakers, and the EmoV-DB database (Adigwe et al., 2018) is built with about 5 h of multi-lingual speech data covering 5 different emotions and recorded by 4 English speakers and a French speaker. SAVEE (Jackson and Haq, 2014) has emotional utterances spoken by 14 speakers, but all of them are male. These databases contain high-quality emotional speech data that can be used to build an emotional voice conversion for a single speaker, that namely speaker-dependent one. However, they may not be optimal when building a speaker-independent emotional voice conversion framework, due to their limited speaker variability.

### 3.4. Confounding factor

It is known that emotion can be varying from individuals and their backgrounds (Kotti and Paternò, 2012; Fersini et al., 2009; Arnold, 1960; Dai et al., 2009). As a para-linguistic factor, emotion is complex with multiple speech characteristics, such as age and accent. In emotional voice conversion, we would like to convert the emotional state of the individual, and carry other para-linguistic characteristics from the source to the target. In the study of speaker-independent emotional voice conversion, we always expect an emotional speech database with a large variety of speakers, and other confounding factors should be controlled at the same time.

We note that there are some emotional speech databases with a large variety of speakers but not suitable to build a speaker-independent emotional voice conversion framework, due to their confounding factors. For example, the eINTERFACE'05 (Martin et al., 2006) contains English speech data of 42 different speakers from 14 nationalities.

However, the underlying accents in the speech are unwanted for emotional voice conversion. IEMOCAP (Busso et al., 2008) is collected from conversations, and contains confounding factors such as laughter and sigh. These confounding factors can affect the conversion performance, thus they are unwanted in designing an emotional voice conversion database.

### 3.5. Recording environment

Unlike the robustness of emotion recognition systems could benefit from the external noises, the performance of speech synthesis is hugely dependent on the quality of the audio samples. The speech samples with low signal-to-noise ratio (SNR) or low sampling rate could harm the quality of the synthesized audios. Therefore, we always require a typical indoor environment with professional recording devices during the audio collection for speech synthesis tasks. We note that there are some emotional speech databases with a large amount of speech data from a large variety of speakers. The speech data is usually collected from dynamic conversations, films or TV shows, and evaluated by subjects. However, these databases might not be optimal to build speech synthesis systems because of their recording environments.

As listed in Table 1, there are some databases containing a large number of diverse sentences of emotional data. For example, IEMOCAP (Busso et al., 2008) contains audio-visual recordings of 5 sessions, each of which is displayed by a pair of speakers (male and female) in scripted and improvised scenarios. In total, the database contains 10 speakers and 12 h of data. Similar to IEMOCAP, MSP-IMPROV (Busso et al., 2016) consists of 6 sessions and 9 h of audio-visual data from 12 actors. These databases are both evaluated in terms of emotion category (Ekman, 1992) and emotional dimensions (Posner et al., 2005). CHEAVD (Li et al., 2017) is a large-scale emotional speech database for Mandarin Chinese that contains 219 speakers. In CHEAVD, all the audio-visual data are segmented from films and TV shows. We note that these databases have been widely used for speech emotion recognition because they have a large variability of speakers and lexical content. However, they are not suitable for synthesis purpose due to the overlapping speech and some external noise. Their recording setups could affect the performance of synthesized audios, thus these databases are not ideal choices for emotional speech synthesis and emotional voice conversion.

## 4. ESD database

The ESD database was recorded with the aim to provide the community with a large emotional speech database with a sufficient variety of speaker and lexical coverage. In this section, we first explain how the ESD database overcomes the weakness of existing emotional speech databases. We then present the design details and provide a comprehensive analysis of the ESD database. We also evaluate the ESD database with a speech emotion recognition framework, and report its classification performance, and visualize the emotion embeddings that are learnt from the network.

### 4.1. Addressing the research need

According to the content type, emotional speech databases can be divided into two types: acted and improvised, as shown in Table 1. We note that both types of databases have their pros and cons (Juslin and Laukka, 2001). For example, the actors may create stereotypical portrayals of emotions in acted emotional speech (Bachorowski and Owren, 1995; Kappas and Hess, 2011). As for improvised speech databases, emotions are induced in the speaker under naturalistic conditions. However, it is hard to induce strong and well-differentiated emotions in this manner (Johnstone and Scherer, 2000; Banse and Scherer, 1996). We note that acted speech databases are widely used in emotional speech synthesis (Livingstone and Russo, 2018; Burkhardt



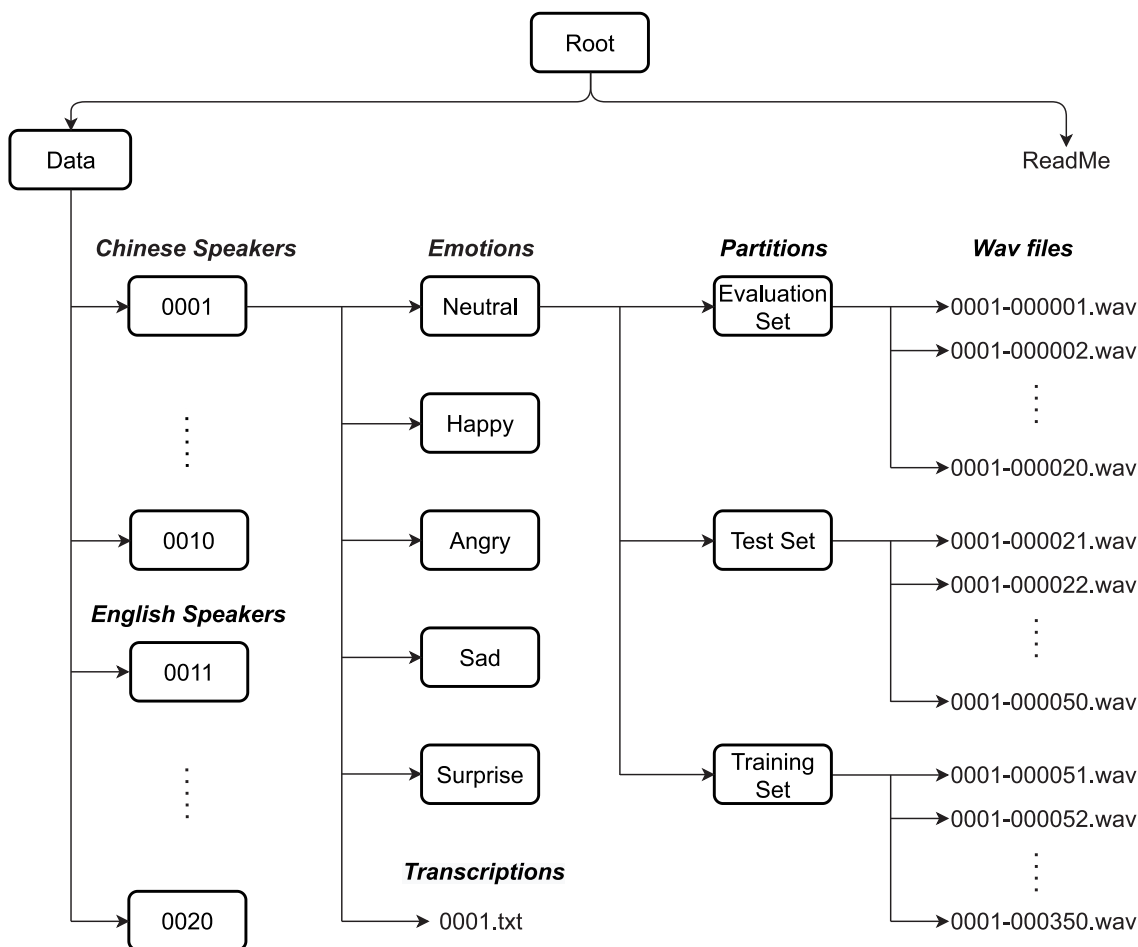


Fig. 6. Schematic representation of directory organization of the ESD database.

et al., 2005; Adigwe et al., 2018), and improvised speech databases are frequently used in speech emotion recognition (Busso et al., 2008; Li et al., 2017). To be consistent with the literature on emotional speech synthesis, we would like to design a database with parallel and acted emotional speech. During the database design, we require the speakers to act with different emotions with the same linguistic content.

The ESD database is designed to fill the gaps left by the existing emotional speech databases as discussed in Section 3, in terms of *lexical variability*, *language variability*, *speaker variability*, *confounding factors*, and *recording environment*. Next, we will explain how ESD database addresses the above five aspects.

#### 1. Lexical variability

As mentioned in Section 3, most the existing emotional speech databases do not have sufficient lexical coverage. We also note that most of voice conversion databases, such as VCC 2016 (Toda et al., 2016), usually contain hundreds of utterances for each speaker for lexical variability. With parallel utterances, it also makes possible to compare the same sentence given by the speakers expressing different emotions. Therefore, we propose to record 350 parallel utterances for each emotion from each speaker in the ESD database.

#### 2. Language variability

The ESD database includes Chinese and English speech data at the same time, which is therefore naturally multi-lingual. The balanced setting of both languages allows us to study emotional expression across different languages and cultures. The ESD database also can be easily used to build a cross-lingual emotional voice conversion or speaker voice conversion framework.

#### 3. Speaker variability

We propose to have emotional speech data containing 10 Chinese speakers and 10 English speakers with a balanced gender setting, that has a larger number of speakers than previous databases (El Haddad et al., 2017; Adigwe et al., 2018; James et al., 2018). Furthermore, the speakers are instructed to speak hundreds of parallel utterances in different emotions. In this way, we expect the ESD database to facilitate the speaker-independent emotional voice conversion studies.

#### 4. Confounding factor

To control the confounding factors in speech, such as age, accents as discussed in Section 3, we propose to have all speakers in the same age group of 25–35. The speakers are instructed to refrain from any confounding expressions, such as laughter or sigh. All English speakers speak North American English, while all Chinese speakers speak standard Mandarin.

#### 5. Recording environment

To ensure an ideal recording environment, all the speech data in ESD database is recorded in a typical indoor environment with an SNR of above 20 dB and a sampling frequency of 16 kHz. It ensures that the ESD database are suitable to build a state-of-the-art emotional voice conversion framework.

### 4.2. Design

The ESD database is recorded by 10 native English speakers and 10 native Chinese speakers. Each speaker contributes 350 utterances,

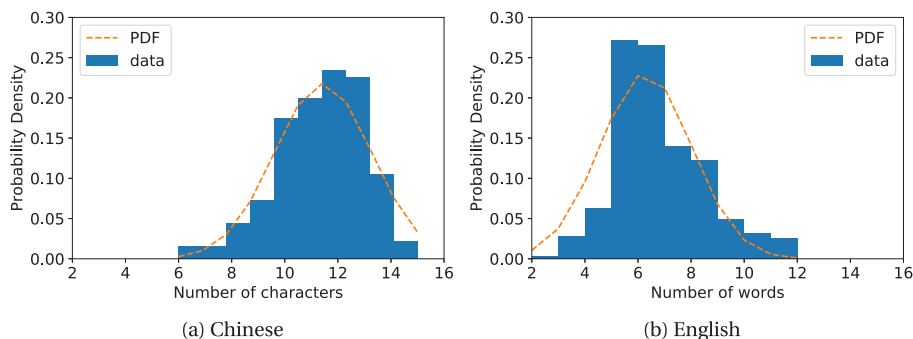


Fig. 7. The histograms of probability density of the number of characters/words per utterance for Chinese and English speech data in *ESD* database. The statistics of characters are reported for Chinese, and the statistics of words are reported for English. Probability density function (PDF) is estimated with Gaussian density function.

Table 2

There are 350 English or Chinese utterances for each speaker, and in each emotion. We partition each speaker-emotion set into an evaluation set, a test set, and a training set.

Set	Number of utterances	
	English	Chinese
Evaluation set	20	20
Test set	30	30
Training set	300	300

covering 5 emotion categories, namely *Neutral*, *Happy*, *Angry*, *Sad*, and *Surprise*. It is made available for research purpose upon request.<sup>2</sup>

#### 4.2.1. Partition

We propose to partition the *ESD* database as shown in Table 2. We partition the 350 parallel utterances of a speaker by emotion into three non-overlapping sets, which are evaluation, test, and training set. We use the first 20 utterances for the evaluation set, the following 30 utterances for the test set, and the rest of the utterances for the training set.

#### 4.2.2. Organization

The directory organization of *ESD* database is shown in Fig. 6. The *Data* folder and *ReadMe* are under the *Root* directory. In the *Data* folder, we have 20 sub-folders for speakers including 10 Chinese speakers (0001–0010) and 10 English speakers (0011–0020). In each speaker’s folder, we have a text file for transcriptions, and 5 sub-folders corresponding to five emotion categories. For each emotion folder, we partition all the speech data into three sub-folders, that are *Evaluation Set*, *Test Set*, and *Training Set*.

### 4.3. Statistics

A summary of the statistics of the *ESD* database is presented in Table 3. In order to give a comprehensive analysis of *ESD* database, the statistics of lexical variability, duration, and F0 are further presented and discussed in the following.

#### 4.3.1. Lexicon

We report the statistics of English words for English speech data, and Chinese characters for Chinese speech data respectively. For each language, 350 parallel utterances are collected in five different emotions, resulting in a total of 1750 utterances for each speaker. We collect speech from 10 speakers for both Chinese and English.

For Chinese, the 1750 utterances have a total character count of 20,025 Chinese characters and 939 unique Chinese characters. They

are commonly used in daily communication. On average, there are 11.5 characters in each Chinese utterance, as shown in Fig. 7(a).

For English, the 1750 utterances have a total word count of 11,015 words and 997 unique lexical words. They are the commonly used English words in daily communication. On average, there are about 6.31 words in each English utterance, as shown in Fig. 7(b).

In our work, 350 utterances are chosen to make sure that the text scripts have a large lexical coverage while reducing the repetition of word or character. The lexicon size in the *ESD* database is close to that for a person to speak and understand a language (Sagar-Fenton and McNeill, 2018).

#### 4.3.2. Duration

Speech duration is the manifestation of speaking rhythm. In Fig. 8, the probability density histograms of the utterance duration of *ESD* database are presented by emotion and language. For both English and Chinese, we observe that the utterance duration of those emotions with lower arousal and negative valence values such as sad tends to have a higher mean and a higher variance than that of emotions with higher arousal and positive valence such as happy (excited). This suggests that there could be some common codes of emotional expression across different languages, which motivates a multi-lingual study on emotional voice conversion.

It is reported that the average utterance and word duration is 2.76 s and 0.44 s respectively for English speech data, and 3.22 s and 0.28 s respectively for Chinese. We notice that Chinese speech has a longer average utterance duration but shorter character duration than English speech, that is mainly because the number of characters in a Chinese sentence is usually larger than the number of words in an English one, as observed in Fig. 7. We also note that the average utterance duration of *ESD* database is similar to that of most publicly available voice conversion databases in the literature (Toda et al., 2016; Lorenzo-Trueba et al., 2018b; Zhao et al., 2020).

#### 4.3.3. Fundamental frequency (F0)

Fundamental frequency (F0) is an essential prosodic descriptor of speech. In Table 4, we summarize the mean and standard variance of F0 over all the utterances by emotion and by gender. The F0 values are extracted by WORLD vocoder (Morise et al., 2016).

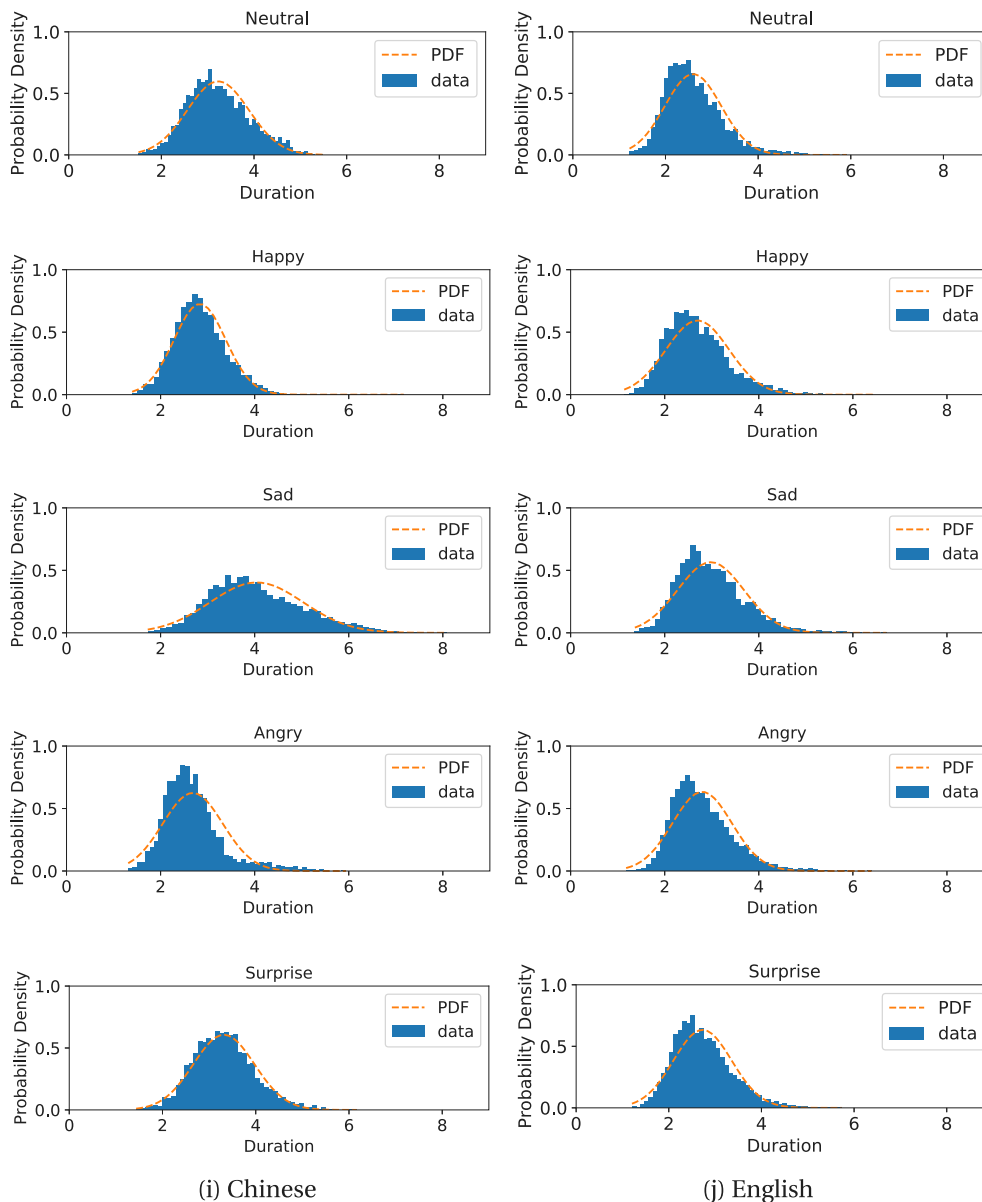
As shown in Table 4, we observe that the female speakers typically have a higher F0 value and standard variance than the male speakers within each emotion category. For both male and female speakers, happy and surprise tend to have a higher mean and variance of F0 than the others. The F0 of the angry category is generally higher than neutral and sad, but it has a lower variance than happy and surprise. The F0 of the sad and neutral categories have lower values, while sad has a slightly larger F0 variance than neutral. These observations generally conform with the findings in the prior studies (Xue et al., 2018).

<sup>2</sup> *ESD* database: <https://github.com/HLTSingapore/Emotional-Speech-Data>.

**Table 3**  
A summary of the statistics of ESD database.

Parameter	Chinese					All	English					All
	Neu	Ang	Sad	Hap	Sur		Neu	Ang	Sad	Hap	Sur	
# speakers	10	10	10	10	10	10	10	10	10	10	10	10
# utterances per speaker	350	350	350	350	350	1750	350	350	350	350	350	1750
# unique utterances	350	350	350	350	350	350	350	350	350	350	350	350
# characters/words per speaker	4005	4005	4005	4005	4005	20,025	2203	2203	2203	2203	2203	11,015
# unique characters/words	939	939	939	939	939	939	997	997	997	997	997	997
Avg. utterance duration [s]	3.23	2.68	4.04	2.84	3.32	3.22	2.61	2.80	2.98	2.70	2.73	2.76
Avg. character/word duration [s]	0.28	0.23	0.35	0.25	0.29	0.28	0.41	0.44	0.47	0.43	0.43	0.44
Total duration [s]	11,305	9380	14,140	9940	11,620	56,385	9135	9800	10,430	9450	9555	48,370

Emotion abbreviations are used as follows: *Neu* stands for neutral, *Ang* stands for angry, *Sad* stands for sad, *Hap* stands for happy and *Sur* stands for surprise. The statistics of characters are reported for Chinese, and the statistics of words are reported for English.



**Fig. 8.** Probability density histograms of the utterance duration [s] of five emotions for Chinese and English speech data. The histograms are reported with fifty bins. Probability density function (PDF) is estimated with Gaussian density function.

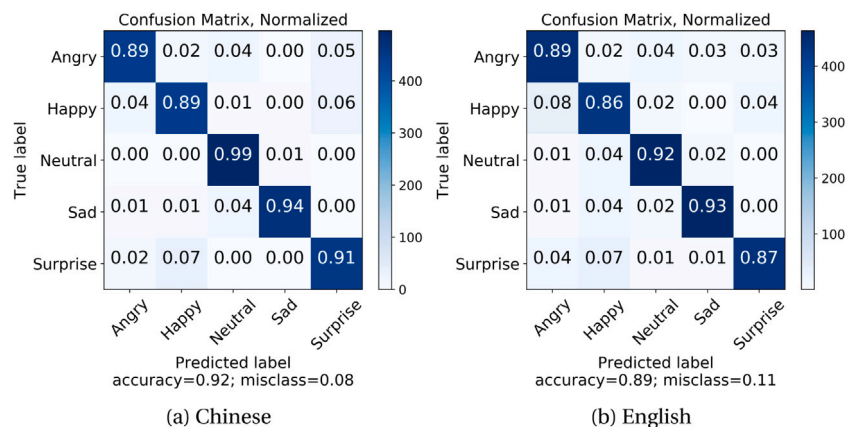
#### 4.4. Emotion classification

To validate the quality of emotional expression in ESD database, we develop a state-of-the-art speaker-independent speech emotion recognition system. We follow the same data partition presented in Fig. 6.

We use the training data from all the speakers to develop a speech emotion recognition system for Chinese and English respectively. We first extract frame-level features with the sliding window of 25 ms and a shift of 10 ms. For each frame, we then extract 312-dimensional feature vectors with openSMILE (Eyben et al., 2010), including zero-crossing

**Table 4**  
A summary of mean and standard variance (std) of F0.

Parameter	Neutral		Angry		Sad		Happy		Surprise	
	Male	Female	Male	Female	Male	Female	Male	Female	Male	Female
<i>(English speech data)</i>										
F0 mean [Hz]	107.60	148.92	134.88	167.28	114.64	154.39	142.86	206.98	154.21	211.52
F0 std [Hz]	26.02	37.64	45.38	54.08	37.46	31.69	51.77	68.69	61.26	82.19
<i>(Chinese speech data)</i>										
F0 mean [Hz]	91.22	156.59	144.34	193.23	102.18	199.79	142.27	221.45	156.09	226.63
F0 std [Hz]	19.88	29.63	41.26	53.57	21.67	40.98	46.99	64.47	50.52	71.07



**Fig. 9.** Confusion matrix of Chinese and English test data in a SER experiment on the *ESD* database. The diagonal entries represent the recall rates of each emotion.

**Table 5**  
A comparison of the MCD [dB] values of Zero effort, CycleGAN-EVC (Zhou et al., 2020b), and VAWGAN-EVC (Zhou et al., 2021b) for four emotion conversion pairs. The lower value of MCD indicates the better conversion performance.

MCD [dB]	Neutral-to-Angry	Neutral-to-Happy	Neutral-to-Sad	Neutral-to-Surprise
Zero effort	6.47	6.64	6.22	6.49
CycleGAN-EVC	4.57	4.46	4.32	4.68
VAWGAN-EVC	5.89	5.73	5.31	5.67

rate, voicing probability, mel-frequency cepstral coefficients (MFCCs) and mel-spectrum.

The SER model consists of a LSTM layer, followed by a ReLU-activated fully connected (FC) layer with 256 nodes. Dropout is applied on the LSTM layer with a keep probability of 0.5. Finally, the resulting 256-features vector is fed to a softmax classifier, which is a FC layer with 5 outputs. We follow the data partition in Table 2 in this experiment, and present a confusion matrix to report the SER results. As shown in Fig. 9, we observe an overall SER accuracy of 92.0% and 89.0% respectively for Chinese and English systems. The results suggest that the speech in the *ESD* database is rendered with high level of emotion consistency.

Deep emotional features are intermediate activations of the SER system, that characterize emotion states (Liu et al., 2020; Zhou et al., 2021c). We randomly select 20 utterances for each emotion from the test set and visualize their deep emotional features using the t-distributed stochastic neighbor embedding (t-SNE) algorithm (Maaten and Hinton, 2008) in a two-dimensional plane, as shown in Fig. 10. The results suggest that the deep emotional features derived from *ESD* serve as an excellent descriptor of speech emotion states. Encouraged by this, we will study the use of deep emotional features as the emotion code for emotional voice conversion.

## 5. Emotional voice conversion on *ESD*

To provide a reference benchmark, we conduct emotional voice conversion experiments on the *ESD* database using state-of-the-art techniques.

### 5.1. Experimental setup

We implement two state-of-the-art emotional voice conversion frameworks, that are (1) *CycleGAN-EVC* (Zhou et al., 2020b) (see Fig. 3), and (2) *VAWGAN-EVC* (Zhou et al., 2021b) (see Fig. 4). The former learns the pair-wise emotion domain translation through the CycleGAN network itself, while the latter relies on the emotion codes to control the conversion. We made available the implementation of the two frameworks for ease of reference.<sup>34</sup> It is noted that *ESD* also can be easily used for speaker-independent EVC.

We choose the emotional speech data of one male speaker ('0013') from *ESD* database to carry out all the experiments. We conduct emotion conversion from neutral to angry, happy, sad and surprise respectively. For each emotion, there are 350 parallel utterances in total. In all experiments, we follow the same partition of evaluation, test, and training set described in Section 4.2.1. For each baseline framework, we train one model for each emotion pair, i.e., neutral to angry ('*Neu-Ang*'), neutral to happy ('*Neu-Hap*'), neutral to sad ('*Neu-Sad*'), and neutral to surprise ('*Neu-Sur*'), with the same set of training data. It is noted that there is no frame-alignment procedure in all the experiments, thus both frameworks do not require parallel training data.

<sup>3</sup> **CycleGAN-EVC:** <https://github.com/KunZhou9646/emotional-voice-conversion-with-CycleGAN-and-CWT-for-Spectrum-and-F0>.

<sup>4</sup> **VAWGAN-EVC:** <https://github.com/KunZhou9646/Speaker-independent-emotional-voice-conversion-based-on-conditional-VAW-GAN-and-CWT>.



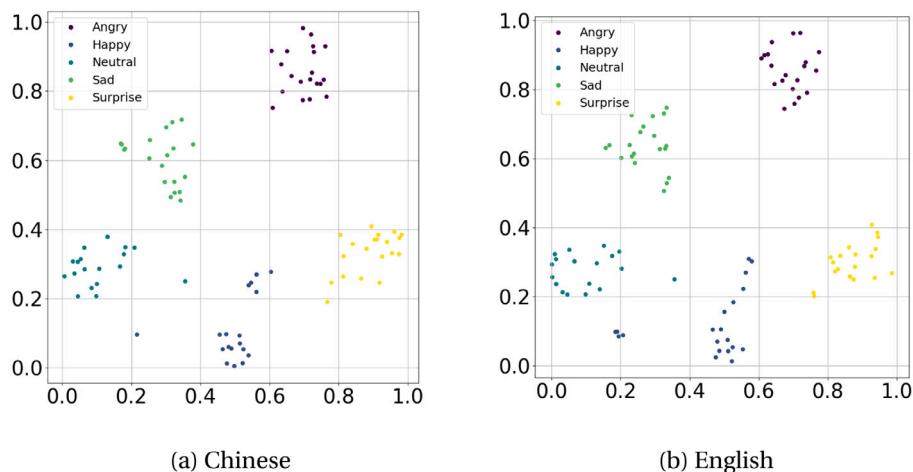


Fig. 10. Visualizations of the deep emotional features for 20 utterances.

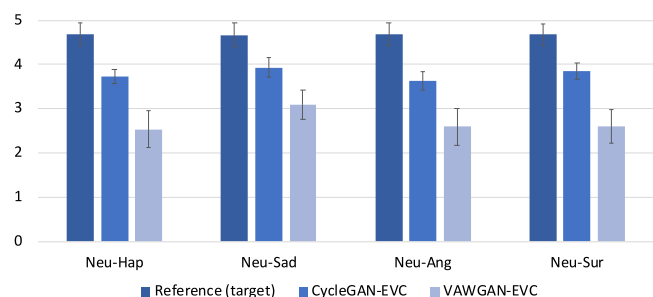


Fig. 11. MOS test results for speech quality with 95% confidence interval. This test aims to compare the speech quality of reference target audio and the converted audio with CycleGAN-EVC and VAWGAN-EVC.

## 5.2. Objective evaluation

For both frameworks, we derive 24-dimensional Mel-cepstral coefficients (MCEPs) from the spectrum, calculate their Mel-cepstral distortion (MCD) values (Kubichek, 1993) between the converted and target MCEPs, and report in Table 5. In Table 5, *Zero effort* denotes the case where we directly compare the source with the target without any conversion. We report *Zero effort* as a contrastive reference. As shown in Table 5, we observe that both CycleGAN-EVC and VAWGAN-EVC achieve lower MCD values than *Zero effort*, which is encouraging. Moreover, the CycleGAN-EVC consistently outperforms VAWGAN-EVC for all conversion pairs.

## 5.3. Subjective evaluation

We further conduct listening experiments to assess the performance of CycleGAN-EVC and VAWGAN-EVC in terms of speech quality and emotion similarity. 11 subjects participate in all the listening experiments, and each of them listens to 72 utterances in total.

We first report the mean opinion score (MOS) of reference target speech, CycleGAN-EVC and VAWGAN-EVC to assess the overall speech quality, which covers two aspects: (1) How the linguistic information and speaker identity is preserved, and (2) the naturalness of the converted speech. During the listening test, the listeners are asked to rate the quality of speech audio with a 5-point scale: 5 for excellent, 4 for good, 3 for fair, 2 for poor, and 1 for bad. As shown in Fig. 11, CycleGAN-EVC outperforms VAWGAN-EVC for all the emotion conversion pairs.

We also conduct an XAB preference test to assess the emotion similarity, where the subjects are asked to choose the speech sample

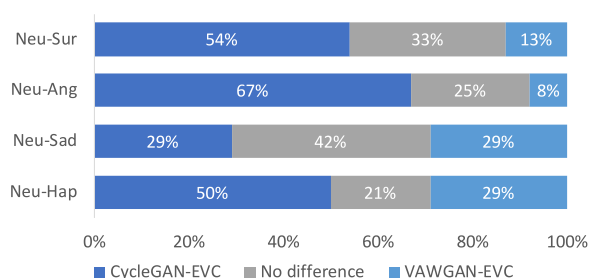


Fig. 12. XAB preference test results for emotion similarity. This test is conducted to compare the performance of CycleGAN-EVC and VAWGAN-EVC in terms of the emotion similarity with the reference target audio.

which sounds closer to the reference in terms of emotional expression. As shown in Fig. 12, we observe that CycleGAN-EVC outperforms VAWGAN-EVC for Neu-Sur, Neu-Ang and Neu-Hap, and achieves comparable results with VAWGAN-EVC for Neu-Sad.

## 5.4. Discussion

We have evaluated two recent proposed EVC frameworks on the ESD database. We observe that CycleGAN-EVC achieves better performance than VAWGAN-EVC with a limited amount of data, especially in terms of speech quality. However, with the same amount of training data, the training of CycleGAN-EVC takes nearly 72 h on one GeForce GTX 1080 Ti, while that of VAWGAN-EVC only takes about 7 h with the same experimental setup. On the other hand, VAWGAN-EVC (Zhou et al., 2020c) allows for the conversion of many-to-many emotions, which is more controllable than CycleGAN-EVC, that can be extended for speaker-independent emotional voice conversion.

To evaluate the performance of the emotional expression in the converted speech, we regard the target speech emotion as the ground truth in emotional similarity experiment. We note that there exists a mismatch between the listener assessment and the speaker's intention for emotion, which will be studied in our future work.

Through the EVC experiments, we seek to evaluate the usability of the ESD database for emotional voice conversion studies. To our delight, both objective and subjective evaluations confirm that the ESD database provides an excellent shared task of emotional voice conversion under a controlled environment. The speech samples are publicly available.<sup>5</sup>

<sup>5</sup> Speech Samples: <https://hlsingapore.github.io/ESD/demo.html>.

## 6. Other voice conversion and text-to-speech on ESD

We have studied the use of the *ESD* database for emotional voice conversion. We also would like to highlight that the database can be used in other voice conversion and text-to-speech studies as well.

### 6.1. Speaker vs. emotional voice conversion

Typically voice conversion is used to describe speaker voice conversion, which seeks to convert speaker identity from one to another without changing the linguistic content (Mohammadi and Kain, 2017; Sisman et al., 2020), as shown in Fig. 1(b). In general, speaker voice conversion includes mono-lingual and cross-lingual voice conversion (Sisman et al., 2020). Cross-lingual voice conversion is a more challenging task because the source and target speakers have different native languages (Abe et al., 1990, 1991; Erro and Moreno, 2007; Du et al., 2020). In VCC 2020, cross-lingual voice conversion becomes one of the main tasks (Zhao et al., 2020). It is noted that a multi-lingual speech database is necessary for cross-lingual voice conversion.

We note that speaker voice conversion and emotional voice conversion are also different in many ways. For example, the former considers prosody-related features as speaker-independent and mostly focuses on spectrum conversion. However, emotion is the interplay of acoustic attributes concerning spectrum and prosody (Arnold, 1960). Therefore, in the latter, prosody conversion requires the same level of attention as spectral conversion. The *ESD* database offers an invaluable addition to extend the scope of the current voice conversion study.

### 6.2. Voice conversion databases vs. ESD

Unlike emotional voice conversion, there are abundant speech databases for speaker voice conversion. For example, VCC 2016 database (Toda et al., 2016) contains parallel speech data from 5 male speakers and 5 female speakers, with each speaker speaking 216 utterances in total. VCC 2018 database (Lorenzo-Trueba et al., 2018b) contains both parallel and non-parallel utterances from 6 male speakers and 6 female speakers. VCC 2020 database (Zhao et al., 2020) is a multi-lingual database that contains 4 different languages speech data. Other databases, such as the CMU-Arctic database (Kominek and Black, 2004) and the VCTK databases (Veaux et al., 2016) are also popular for voice conversion research. There are other large speech databases as well, such as LibriTTS (Zen et al., 2019) and VoxCeleb (Nagrani et al., 2020). LibriTTS corpus has 585 h of transcribed speech data uttered by a total of 2456 speakers, while the VoxCeleb database is further larger and consists of about 2800 h of untranscribed speech from over 6000 speakers. These two databases are mostly used to train a noise-robust speaker encoder to improve the speaker voice conversion performance.

We note that the speech in the above databases is emotion-free, which is not suitable for emotional voice conversion. The multi-speaker and cross-lingual nature of *ESD* database make possible new studies beyond the existing speaker voice conversion databases.

### 6.3. Emotional text-to-speech on ESD

For text-to-speech studies, there are many large-scale and high-quality speech databases, such as the LJ-Speech database (Ito and Johnson, 2017) and the Blizzard Challenge databases (Zhou et al., 2020a; Wu et al., 2019; King and Karaiskos, 2013). However, the emotional text-to-speech studies have been suffering from the lack of suitable emotional speech database as discussed in Section 3. Due to the lack of data, researchers mostly use internal databases for emotional text-to-speech studies (Um et al., 2020; Lei et al., 2021; Kwon et al., 2019; Zhu et al., 2019). We note that these internal databases are not publicly available and limit the related research in emotional text-to-speech.

The *ESD* database is designed to fill the gap and can be easily used for various text-to-speech tasks, especially for emotional text-to-speech. To our delight, there are already a few successful attempts (Zhou et al., 2021a; Liu et al., 2021) to build emotional text-to-speech frameworks on the *ESD* database. In Zhou et al. (2021a), researchers propose a joint training framework of emotional voice conversion and emotional text-to-speech. In Liu et al. (2021), an emotional text-to-speech framework with reinforcement learning is built and evaluated through the comparative study with several state-of-the-art emotional text-to-speech frameworks (Li et al., 2021; Cai et al., 2021) on the *ESD* database. These studies have all shown the effectiveness, which further show the promise of the *ESD* database on the emotional text-to-speech task.

## 7. Conclusion

This paper provides a comprehensive overview of the recent research and existing emotional speech databases for emotional voice conversion. To our best knowledge, this paper is the first overview paper that covers emotional voice conversion research and databases in recent years. Moreover, we release the *ESD* database and make it publicly available. The *ESD* database represents one of the largest emotional speech databases in the literature. By reporting several experiments on the *ESD* database, this paper provides a reference benchmark for emotional voice conversion studies that represent the state-of-the-art.

### CRediT authorship contribution statement

**Kun Zhou:** Conception and design of study, Acquisition of data, Analysis and/or interpretation of data, Drafting the manuscript, Revising the manuscript critically for important intellectual content. **Berrak Sisman:** Conception and design of study, Acquisition of data, Revising the manuscript critically for important intellectual content. **Rui Liu:** Drafting the manuscript, Revising the manuscript critically for important intellectual content. **Haizhou Li:** Revising the manuscript critically for important intellectual content.

### Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

### Acknowledgments

The research by Berrak Sisman and Rui Liu is funded by SUTD Start-up Grant Artificial Intelligence for Human Voice Conversion (SRG ISTD 2020 158) and SUTD AI Grant - Thrust 2 Discovery by AI (SG-PAIRS1821).

The research by Kun Zhou and Haizhou Li is supported by the Science and Engineering Research Council, Agency of Science, Technology and Research (A\*STAR), Singapore, through the National Robotics Program under Human-Robot Interaction Phase 1 (Grant No. 192 25 00054); Human Robot Collaborative AI under its AME Programmatic Funding Scheme (Project No. A18A2b0046); National Research Foundation Singapore under its AI Singapore Programme (Award Grant No: AISG-100E-2018-006), and by A\*STAR under its RIE2020 Advanced Manufacturing and Engineering Domain (AME) Programmatic Grant (Grant No. A1687b0033, Project Title: Spiking Neural Networks).

Approval of the version of the manuscript to be published.

## References

- Abe, M., Shikano, K., Kuwabara, H., 1990. Cross-language voice conversion. In: *International Conference on Acoustics, Speech, and Signal Processing*. IEEE, pp. 345–348.
- Abe, M., Shikano, K., Kuwabara, H., 1991. Statistical analysis of bilingual speaker's speech for cross-language voice conversion. *J. Acoust. Soc. Am.* 90 (1), 76–82.
- Adigwe, A., Tits, N., Haddad, K.E., Ostadabbas, S., Dutoit, T., 2018. The emotional systems database: Towards controlling the emotion dimension in voice generation systems. *arXiv preprint arXiv:1806.09514*.
- Aihara, R., Takashima, R., Takiguchi, T., Ariki, Y., 2012. Gmm-based emotional voice conversion using spectrum and prosody features. *Amer. J. Signal Process.*
- Aihara, R., Ueda, R., Takiguchi, T., Ariki, Y., 2014. Exemplar-based emotional voice conversion using non-negative matrix factorization. In: *APSIPA ASC*. IEEE.
- Ak, K.E., Lim, J.H., Tham, J.Y., Kassim, A.A., 2019. Attribute manipulation generative adversarial networks for fashion images. In: *Proceedings of the IEEE International Conference on Computer Vision*.
- Ak, K.E., Lim, J.H., Tham, J.Y., Kassim, A.A., 2020a. Semantically consistent text to fashion image synthesis with an enhanced attentional generative adversarial network. *Pattern Recognit. Lett.*
- Ak, K.E., Sun, Y., Lim, J.H., 2020b. Learning cross-modal representations for language-based image manipulation. In: *Proceedings of the IEEE ICIP*.
- Akçay, M.B., Oğuz, K., 2020. Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. *Speech Commun.*
- Almahairi, A., Rajeswar, S., Sordoni, A., Bachman, P., Courville, A., 2018. Augmented cyclegan: Learning many-to-many mappings from unpaired data. *arXiv preprint arXiv:1802.10151*.
- An, S., Ling, Z., Dai, L., 2017. Emotional statistical parametric speech synthesis using lstm-rnns. In: *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, pp. 1613–1616.
- Arias, P., Rachman, L., Liuni, M., Aucouturier, J.-J., 2020. Beyond correlation: acoustic transformation methods for the experimental study of emotional voice and speech. *Emot. Rev.* 1754073920934544.
- Arnold, M.B., 1960. *Emotion and Personality*. Columbia University Press.
- Bachorowski, J.-A., Owren, M.J., 1995. Vocal expression of emotion: Acoustic properties of speech are associated with emotional intensity and context. *Psychol. Sci.* 6 (4), 219–224.
- Banse, R., Scherer, K.R., 1996. Acoustic profiles in vocal emotion expression. *J. Personal. Soc. Psychol.* 70 (3), 614.
- Bao, F., Neumann, M., Vu, N.T., 2019. Cyclegan-based emotion style transfer as data augmentation for speech emotion recognition. In: *INTERSPEECH*. pp. 2828–2832.
- Barra-Chicote, R., Yamagishi, J., King, S., Montero, J.M., Macias-Guarasa, J., 2010. Analysis of statistical parametric and unit selection speech synthesis systems applied to emotional speech. *Speech Commun.* 52 (5), 394–404.
- Benesty, J., Sondhi, M.M., Huang, Y., 2007. *Springer Handbook of Speech Processing*. Springer.
- Biassoni, F., Balzarotti, S., Giamporcaro, M., Ciceri, R., 2016. Hot or cold anger? Verbal and vocal expression of anger while driving in a simulated anger-provoking scenario. *Sage Open* 6 (3), 2158244016658084.
- Brunswick, E., 1956. Historical and thematic relations of psychology to other sciences. *Sci. Mon.* 83 (3), 151–161.
- Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W.F., Weiss, B., 2005. A database of german emotional speech. In: *Ninth European Conference on Speech Communication and Technology*.
- Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J.N., Lee, S., Narayanan, S.S., 2008. Iemocap: Interactive emotional dyadic motion capture database. *Lang. Resour. Eval.* 42 (4), 335.
- Busso, C., Parthasarathy, S., Burmania, A., AbdelWahab, M., Sadoughi, N., Provost, E.M., 2016. Msp-IMPROV: An acted corpus of dyadic interactions to study emotion perception. *IEEE Trans. Affect. Comput.* 8 (1), 67–80.
- Cai, X., Dai, D., Wu, Z., Li, X., Li, J., Meng, H., 2021. Emotion controllable speech synthesis using emotion-unlabeled dataset with the assistance of cross-domain speech emotion recognition. In: *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 5734–5738.
- Cao, H., Cooper, D.G., Keutmann, M.K., Gur, R.C., Nenkova, A., Verma, R., 2014. Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE Trans. Affect. Comput.* 5 (4), 377–390.
- Cao, Y., Liu, Z., Chen, M., Ma, J., Wang, S., Xiao, J., 2020. Nonparallel emotional speech conversion using VAE-GAN. In: *Proc. Interspeech 2020*, pp. 3406–3410.
- Chen, L.-H., Ling, Z.-H., Liu, L.-J., Dai, L.-R., 2014. Voice conversion using deep neural networks with layer-wise generative training. *IEEE/ACM Trans. Audio Speech Lang. Process.* 22 (12), 1859–1872.
- Childers, D.G., Wu, K., Hicks, D., Yegnanarayana, B., 1989. Voice conversion. *Speech Commun.* 8 (2), 147–158.
- Choi, Y., Choi, M., Kim, M., Ha, J.-W., Kim, S., Choo, J., 2018. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8789–8797.
- Choi, H., Hahn, M., 2021. Sequence-to-sequence emotional voice conversion with strength control. *IEEE Access* 9, 42674–42687.
- Choi, H., Park, S., Park, J., Hahn, M., 2019. Multi-speaker emotional acoustic modeling for cnn-based speech synthesis. In: *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 6950–6954.
- Chou, J.-c., Yeh, C.-c., Lee, H.-y., 2019. One-shot voice conversion by separating speaker and content representations with instance normalization. *arXiv preprint arXiv:1904.05742*.
- Çişman, B., Li, H., Tan, K.C., 2017. Sparse representation of phonetic features for voice conversion with and without parallel data. In: *2017 IEEE Automatic Speech Recognition and Understanding Workshop*.
- Costantini, G., Iaderola, I., Paoloni, A., Todisco, M., 2014. Emovo corpus: an Italian emotional speech database. In: *International Conference on Language Resources and Evaluation (LREC 2014)*. European Language Resources Association (ELRA), pp. 3501–3504.
- Crumpton, J., Bethel, C.L., 2016. A survey of using vocal prosody to convey emotion in robot speech. *Int. J. Soc. Robot.* 8 (2), 271–285.
- Dai, K., Fell, H., MacAuslan, J., 2009. Comparing emotions using acoustics and human perceptual dimensions. In: *CHI'09 Extended Abstracts on Human Factors in Computing Systems*.
- Desai, S., Black, A.W., Yegnanarayana, B., Prahallad, K., 2010. Spectral mapping using artificial neural networks for voice conversion. *IEEE Trans. Audio Speech Lang. Process.* 18 (5), 954–964.
- Du, Z., Zhou, K., Sisman, B., Li, H., 2020. Spectrum and prosody conversion for cross-lingual voice conversion with cyclegan. In: *2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, pp. 507–513.
- Ekman, P., 1992. An argument for basic emotions. *Cogn. Emot.*
- El Ayadi, M., Kamel, M.S., Karray, F., 2011. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognit.* 44 (3), 572–587.
- El Haddad, K., Torre, I., Gilmartin, E., Çakmak, H., Dupont, S., Dutoit, T., Campbell, N., 2017. Introducing amus: The amused speech database. In: *International Conference on Statistical Language and Speech Processing*. Springer, pp. 229–240.
- Elbarougy, R., Akagi, M., 2014. Improving speech emotion dimensions estimation using a three-layer model of human perception. *Acoust. Sci. Technol.* 35 (2), 86–98.
- Elgaar, M., Park, J., Lee, S.W., 2020. Multi-speaker and multi-domain emotional voice conversion using factorized hierarchical variational autoencoder. In: *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 7769–7773.
- Emir Ak, K., Hwee Lim, J., Yew Tham, J., Kassim, A., 2019. Semantically consistent hierarchical text to fashion image synthesis with an enhanced-attentional generative adversarial network. In: *Proceedings of the IEEE International Conference on Computer Vision Workshops*.
- Engberg, I.S., Hansen, A.V., Andersen, O., Dalsgaard, P., 1997. Design, recording and verification of a danish emotional speech database. In: *Fifth European Conference on Speech Communication and Technology*.
- Erickson, D., 2005. Expressive speech: Production, perception and application to speech synthesis. *Acoust. Sci. Technol.* 26 (4), 317–325.
- Erro, D., Moreno, A., 2007. Frame alignment method for cross-lingual voice conversion. In: *Eighth Annual Conference of the International Speech Communication Association*.
- Erro, D., Moreno, A., Bonafonte, A., 2009a. Voice conversion based on weighted frequency warping. *IEEE Trans. Audio Speech Lang. Process.* 18 (5), 922–931.
- Erro, D., Navas, E., Hernández, I., Saratxaga, I., 2009b. Emotion conversion based on prosodic unit selection. *IEEE Trans. Audio Speech Lang. Process.* 18 (5), 974–983.
- Eyben, F., Weninger, F., Gross, F., Schuller, B., 2013. Recent developments in opensmile, the munich open-source multimedia feature extractor. In: *Proceedings of the 21st ACM International Conference on Multimedia*, pp. 835–838.
- Eyben, F., Wöllmer, M., Schuller, B., 2010. Opensmile: the munich versatile and fast open-source audio feature extractor. In: *Proceedings of the 18th ACM International Conference on Multimedia*, pp. 1459–1462.
- Fang, F., Yamagishi, J., Echizen, I., Lorenzo-Tрубеa, J., 2018. High-quality nonparallel voice conversion based on cycle-consistent adversarial network. In: *IEEE ICASSP*. pp. 5279–5283.
- Fersini, E., Messina, E., Arosio, G., Archetti, F., 2009. Audio-based emotion recognition in judicial domain: A multilayer support vector machines approach. In: *International Workshop on Machine Learning and Data Mining in Pattern Recognition*. Springer.
- Gao, J., Chakraborty, D., Tembine, H., Olaleye, O., 2019. Nonparallel emotional speech conversion. In: *Proc. Interspeech 2019*, pp. 2858–2862.
- Ghosh, S., Laksana, E., Morency, L.-P., Scherer, S., 2016. Representation learning for speech emotion recognition. In: *Interspeech*. pp. 3603–3607.
- Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014. Generative adversarial nets. In: *Advances in Neural Information Processing Systems*. pp. 2672–2680.
- Grimm, M., Kroschel, K., Narayanan, S., 2008. The vera am mittag german audio-visual emotional speech database. In: *2008 IEEE International Conference on Multimedia and Expo*. IEEE, pp. 865–868.
- Gunes, H., Schuller, B., 2013. Categorical and dimensional affect analysis in continuous input: Current trends and future directions. *Image Vis. Comput.* 31 (2), 120–136.



- Helander, E., Virtanen, T., Nurminen, J., Gabbouj, M., 2010. Voice conversion using partial least squares regression. *IEEE Trans. Audio Speech Lang. Process.*
- Hirschberg, J., 2004. Pragmatics and intonation. *Handb. Pragmat.* 515–537.
- Hsu, C.-C., Hwang, H.-T., Wu, Y.-C., Tsao, Y., Wang, H.-M., 2016. Voice conversion from non-parallel corpora using variational auto-encoder. In: 2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA). IEEE, pp. 1–6.
- Hsu, C.-C., Hwang, H.-T., Wu, Y.-C., Tsao, Y., Wang, H.-M., 2017. Voice conversion from unaligned corpora using variational autoencoding wasserstein generative adversarial networks. *arXiv preprint arXiv:1704.00849*.
- Huang, C.-F., Akagi, M., 2008. A three-layered model for expressive speech perception. *Speech Commun.* 50 (10), 810–828.
- Huang, W.-C., Hayashi, T., Wu, Y.-C., Kameoka, H., Toda, T., 2019. Voice transformer network: Sequence-to-sequence voice conversion using transformer with text-to-speech pretraining. *arXiv preprint arXiv:1912.06813*.
- Inanoglu, Z., Young, S., 2007. A system for transforming the emotion in speech: Combining data-driven conversion techniques for prosody and voice quality. In: Eighth Annual Conference of the International Speech Communication Association.
- Inanoglu, Z., Young, S., 2009. Data-driven emotion conversion in spoken english. *Speech Commun.*
- Ito, K., Johnson, L., 2017. The LJ speech dataset. <https://keithito.com/LJ-Speech-Dataset/>.
- Jackson, P., Haq, S., 2014. Surrey Audio-Visual Expressed Emotion (Savee) Database. University of Surrey, Guildford, UK.
- James, J., Tian, L., Watson, C.I., 2018. An open source emotional speech corpus for human robot interaction applications. In: *Interspeech*. pp. 2768–2772.
- Johnstone, T., Scherer, K.R., 2000. Vocal communication of emotion. *Handb. Emot.* 2, 220–235.
- Juslin, P.N., Laukka, P., 2001. Impact of intended emotion intensity on cue utilization and decoding accuracy in vocal expression of emotion. *Emotion* 1 (4), 381.
- Kain, A., Macon, M.W., 1998. Spectral voice conversion for text-to-speech synthesis. In: *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'98 (Cat. No. 98CH36181)*, Vol. 1. IEEE, pp. 285–288.
- Kameoka, H., Kaneko, T., Tanaka, K., Hojo, N., 2018. Stargan-vc: Non-parallel many-to-many voice conversion using star generative adversarial networks. In: 2018 IEEE Spoken Language Technology Workshop (SLT). IEEE, pp. 266–273.
- Kane, J., Aylett, M., Yanushevskaya, I., Gobl, C., 2014. Phonetic feature extraction for context-sensitive glottal source processing. *Speech Commun.* 59, 10–21.
- Kaneko, T., Kameoka, H., 2017. Parallel-data-free voice conversion using cycle-consistent adversarial networks. *arXiv preprint arXiv:1711.11293*.
- Kaneko, T., Kameoka, H., 2018. Cyclegan-vc: Non-parallel voice conversion using cycle-consistent adversarial networks. In: 2018 26th European Signal Processing Conference (EUSIPCO). IEEE, pp. 2100–2104.
- Kaneko, T., Kameoka, H., Tanaka, K., Hojo, N., 2019. Stargan-VC2: Rethinking conditional methods for stargan-based voice conversion. *arXiv preprint arXiv:1907.12279*.
- Kappas, A., Hess, U., 2011. Nonverbal aspects of oral communication. In: *Aspects of Oral Communication*. De Gruyter, pp. 169–180.
- Kawanami, H., Iwami, Y., Toda, T., Saruwatari, H., Shikano, K., 2003. Gmm-based voice conversion applied to emotional speech synthesis.
- Kim, T.-H., Cho, S., Choi, S., Park, S., Lee, S.-Y., 2020. Emotional voice conversion using multitask learning with text-to-speech. In: *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 7774–7778.
- King, S., Karaiskos, V., 2013. The blizzard challenge 2013. In: *Proc. Blizzard Challenge Workshop*, Vol. 2013.
- Kingma, D.P., Welling, M., 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.
- Kominek, J., Black, A.W., 2004. The cmu arctic speech databases. In: *Fifth ISCA Workshop on Speech Synthesis*.
- Kotti, M., Paternò, F., 2012. Speaker-independent emotion recognition exploiting a psychologically-inspired binary cascade classification schema. *Int. J. Speech Technol.* 15 (2), 131–150.
- Kubichek, R., 1993. Mel-cestral distance measure for objective speech quality assessment. In: *Proceedings of IEEE Pacific Rim Conference on Communications Computers and Signal Processing*, Vol. 1. IEEE, pp. 125–128.
- Kwon, O., Jang, I., Ahn, C., Kang, H.-G., 2019. An effective style token weight control technique for end-to-end emotional speech synthesis. *IEEE Signal Process. Lett.* 26 (9), 1383–1387.
- Latif, S., Rana, R., Khalifa, S., Jurdak, R., Qadir, J., Schuller, B.W., 2020. Deep representation learning in speech processing: Challenges, recent advances, and future trends. *arXiv preprint arXiv:2001.00378*.
- Latorre, J., Akamine, M., 2008. Multilevel parametric-base f0 model for speech synthesis. In: *Ninth Annual Conference of the International Speech Communication Association*.
- Le Moine, C., Obin, N., Roebel, A., 2021. Towards end-to-end f0 voice conversion based on dual-gan with convolutional wavelet kernels.
- Lei, Y., Yang, S., Xie, L., 2021. Fine-grained emotion strength transfer, control and prediction for emotional speech synthesis. In: 2021 IEEE Spoken Language Technology Workshop (SLT). IEEE, pp. 423–430.
- Li, X., Akagi, M., 2016. Multilingual speech emotion recognition system based on a three-layer model. In: *INTERSPEECH*. pp. 3608–3612.
- Li, J., Sun, X., Wei, X., Li, C., Tao, J., 2019. Reinforcement learning based emotional editing constraint conversation generation. *arXiv preprint arXiv:1904.08061*.
- Li, Y., Tao, J., Chao, L., Bao, W., Liu, Y., 2017. Cheavd: a Chinese natural emotional audio-visual database. *J. Ambient Intell. Humaniz. Comput.* 8 (6), 913–924.
- Li, T., Yang, S., Xue, L., Xie, L., 2021. Controllable emotion transfer for end-to-end speech synthesis. In: 2021 12th International Symposium on Chinese Spoken Language Processing (ISCSLP). IEEE, pp. 1–5.
- Liu, S., Cao, Y., Kang, S., Hu, N., Liu, X., Su, D., Yu, D., Meng, H., 2020a. Transferring source style in non-parallel voice conversion. In: *Proc. Interspeech 2020*, pp. 4721–4725.
- Liu, S., Cao, Y., Meng, H., 2020b. Multi-target emotional voice conversion with neural vocoders. *arXiv preprint arXiv:2004.03782*.
- Liu, R., Sisman, B., Gao, G., Li, H., 2020c. Expressive tts training with frame and style reconstruction loss. *arXiv preprint arXiv:2008.01490*.
- Liu, R., Sisman, B., Li, H., 2021. Reinforcement learning for emotional text-to-speech synthesis with improved emotion discriminability. *arXiv preprint arXiv:2104.01408*.
- Livingstone, S.R., Russo, F.A., 2018. 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), e0196391.
- Lorenzo-Trueba, J., Henter, G.E., Takaki, S., Yamagishi, J., Morino, Y., Ochiai, Y., 2018a. Investigating different representations for modeling and controlling multiple emotions in dnn-based speech synthesis. *Speech Commun.*
- Lorenzo-Trueba, J., Yamagishi, J., Toda, T., Saito, D., Villavicencio, F., Kinnunen, T., Ling, Z., 2018b. The voice conversion challenge 2018: Promoting development of parallel and nonparallel methods. *arXiv preprint arXiv:1804.04262*.
- Lu, J., Zhou, K., Sisman, B., Li, H., 2020. Vaw-gan for singing voice conversion with non-parallel training data. *arXiv preprint arXiv:2008.03992*.
- Luo, Z., Chen, J., Takiguchi, T., Ariki, Y., 2017. Emotional voice conversion with adaptive scales f0 based on wavelet transform using limited amount of emotional data. In: *INTERSPEECH*. pp. 3399–3403.
- Luo, Z., Chen, J., Takiguchi, T., Ariki, Y., 2019. Emotional voice conversion using dual supervised adversarial networks with continuous wavelet transform f0 features. *IEEE/ACM Trans. Audio Speech Lang. Process.* 27 (10), 1535–1548.
- Luo, Z., Takiguchi, T., Ariki, Y., 2016. Emotional voice conversion using deep neural networks with mcc and f0 features. In: 2016 IEEE/ACIS 15th ICIS.
- Maaten, L.v.d., Hinton, G., 2008. Visualizing data using t-SNE. *J. Mach. Learn. Res.* 9 (Nov), 2579–2605.
- Manokara, K., Đurić, M., Fischer, A., Sauter, D., 2020. Do people agree on how positive emotions are expressed? A survey of four emotions and five modalities across 11 cultures.
- Martin, O., Kotsia, I., Macq, B., Pitas, I., 2006. The enterface'05 audio-visual emotion database. In: 22nd International Conference on Data Engineering Workshops (ICDEW'06). IEEE, p. 8.
- Mehrabian, A., Wiener, M., 1967. Decoding of inconsistent communications. *J. Personal. Soc. Psychol.* 6 (1), 109.
- Ming, H., Huang, D., Dong, M., Li, H., Xie, L., Zhang, S., 2015. Fundamental frequency modeling using wavelets for emotional voice conversion. In: 2015 International Conference on Affective Computing and Intelligent Interaction (ACII). IEEE, pp. 804–809.
- Ming, H., Huang, D., Xie, L., Wu, J., Dong, M., Li, H., 2016a. Deep bidirectional lstm modeling of timbre and prosody for emotional voice conversion. In: *Interspeech 2016*. pp. 2453–2457.
- Ming, H., Huang, D., Xie, L., Zhang, S., Dong, M., Li, H., 2016b. Exemplar-based sparse representation of timbre and prosody for voice conversion. In: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 5175–5179.
- Mohammadi, S.H., Kain, A., 2017. An overview of voice conversion systems. *Speech Commun.* 88, 65–82.
- Morise, M., Yokomori, F., Ozawa, K., 2016. World: a vocoder-based high-quality speech synthesis system for real-time applications. *IEICE Trans. Inf. Syst.* 99 (7), 1877–1884.
- Müller, M., 2007. Dynamic time warping. *Inf. Retr. Music Motion* 69–84.
- Nagrani, A., Chung, J.S., Xie, W., Zisserman, A., 2020. Voxceleb: Large-scale speaker verification in the wild. *Comput. Speech Lang.* 60, 101027. <http://dx.doi.org/10.1016/j.csl.2019.101027>, URL <http://www.sciencedirect.com/science/article/pii/S0885230819302712>.
- Nakashika, T., Takiguchi, T., Ariki, Y., 2014. High-order sequence modeling using speaker-dependent recurrent temporal restricted boltzmann machines for voice conversion. In: *Fifteenth Annual Conference of the International Speech Communication Association*.
- Nekvinda, T., Dušek, O., 2020. One model, many languages: Meta-learning for multilingual text-to-speech. In: *Proc. Interspeech 2020*, pp. 2972–2976.
- Nose, T., Yamagishi, J., Masuko, T., Kobayashi, T., 2007. A style control technique for hmm-based expressive speech synthesis. *IEICE Trans. Inf. Syst.* 90 (9), 1406–1413.
- Obin, N., Beliao, J., 2018. Sparse coding of pitch contours with deep auto-encoders. In: *Speech Prosody*.
- Obin, N., Lacheret, A., Rodet, X., 2011. Stylization and trajectory modelling of short and long term speech prosody variations.



- Parada-Cabaleiro, E., Costantini, G., Batliner, A., Schmitt, M., Schuller, B.W., 2020. Demos: an Italian emotional speech corpus. *Lang. Resour. Eval.* 54 (2), 341–383.
- Pichora-Fuller, M.K., Dupuis, K., 2020. Toronto emotional speech set (TESS). <http://dx.doi.org/10.5683/SP2/E8H2MF>.
- Pittermann, J., Pittermann, A., Minker, W., 2010. *Handling Emotions in Human-Computer Dialogues*. Springer.
- Posner, J., Russell, J.A., Peterson, B.S., 2005. The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. *Dev. Psychopathol.* 17 (3), 715.
- Rizos, G., Baird, A., Elliott, M., Schuller, B., 2020. Stargan for emotional speech conversion: Validated by data augmentation of end-to-end emotion recognition. In: *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 3502–3506.
- Robinson, C., Obin, N., Roebel, A., 2019. Sequence-to-sequence modelling of F0 for speech emotion conversion. In: *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 6830–6834.
- Russell, J.A., 1980. A circumplex model of affect. *J. Personal. Soc. Psychol.* 39 (6), 1161.
- Sagar-Fenton, B., McNeill, L., 2018. How many words do you need to speak a language?. <https://www.bbc.com/news/world-44569277>.
- Sager, J., Shankar, R., Reinhold, J., Venkataraman, A., 2019. Vesuvius: A crowd-annotated database to study emotion production and perception in spoken English. In: *INTERSPEECH*. pp. 316–320.
- Saratkaga, I., Navas, E., Hernández, I., Luengo, I., 2006. Designing and recording an emotional speech database for corpus based synthesis in Basque. In: *LREC*. pp. 2126–2129.
- Schnell, B., Huybrechts, G., Perz, B., Drugman, T., Lorenzo-Trueba, J., 2021. Emocat: Language-agnostic emotional voice conversion. [arXiv:2101.05695](https://arxiv.org/abs/2101.05695).
- Schroder, M., 2006. Expressive degree of activation in synthetic speech. *IEEE Trans. Audio Speech Lang. Process.* 14 (4), 1128–1136.
- Schröder, M., 2009. Expressive speech synthesis: Past, present, and possible futures. In: *Affective Information Processing*. Springer, pp. 111–126.
- Schuller, B.W., 2018. Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends. *Commun. ACM* 61 (5), 90–99.
- Schuller, D., Schuller, B.W., 2018. The age of artificial emotional intelligence. *Computer* 51 (9), 38–46.
- Schuller, D.M., Schuller, B.W., 2020. A review on five recent and near-future developments in computational processing of emotion in the human voice. *Emot. Rev.* 1754073919898526.
- Schuller, B., Steidl, S., Batliner, A., Burkhardt, F., Devillers, L., Müller, C., Narayanan, S., 2013. Paralinguistics in speech and language—State-of-the-art and the challenge. *Comput. Speech Lang.* 27 (1), 4–39.
- Seppänen, T., Toivanen, J., Väyrynen, E., 2003. Mediateam speech corpus: a first large Finnish emotional speech database. In: *Proceedings of the Proceedings of XV International Conference of Phonetic Science*. Citeseer, pp. 2469–2472.
- Shankar, R., Hsieh, H.-W., Charon, N., Venkataraman, A., 2019a. Automated emotion morphing in speech based on diffeomorphic curve registration and highway networks. In: *INTERSPEECH*. pp. 4499–4503.
- Shankar, R., Hsieh, H.-W., Charon, N., Venkataraman, A., 2020. Multi-speaker emotion conversion via latent variable regularization and a chained encoder-decoder-predictor network. In: *Proc. Interspeech 2020*, pp. 3391–3395.
- Shankar, R., Sager, J., Venkataraman, A., 2019. A multi-speaker emotion morphing model using highway networks and maximum likelihood objective. In: *Proc. Interspeech 2019*.
- Shankar, R., Sager, J., Venkataraman, A., 2020b. Non-parallel emotion conversion using a deep-generative hybrid network and an adversarial pair discriminator. [arXiv:2007.12932](https://arxiv.org/abs/2007.12932).
- Sisman, B., 2019. *Machine Learning for Limited Data Voice Conversion* (Ph.D. thesis, Ph. D. thesis).
- Sisman, B., Lee, G., Li, H., 2018a. Phonetically aware exemplar-based prosody transformation. In: *Odyssey*. pp. 267–274.
- Sisman, B., Li, H., 2018b. Wavelet analysis of speaker dependent and independent prosody for voice conversion. In: *Interspeech*.
- Şişman, B., Li, H., Tan, K.C., 2017. Transformation of prosody in voice conversion. In: *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, pp. 1537–1546.
- Sisman, B., Yamagishi, J., King, S., Li, H., 2020. An overview of voice conversion and its challenges: From statistical modeling to deep learning. *IEEE/ACM Trans. Audio Speech Lang. Process.*
- Sisman, B., Zhang, M., Dong, M., Li, H., 2019a. On the study of generative adversarial networks for cross-lingual voice conversion. *IEEE ASRU*.
- Sisman, B., Zhang, M., Li, H., 2018b. A voice conversion framework with tandem feature sparse representation and speaker-adapted wavenet vocoder. In: *Interspeech*. pp. 1978–1982.
- Sisman, B., Zhang, M., Li, H., 2019b. Group sparse representation with wavenet vocoder adaptation for spectrum and prosody conversion. *IEEE/ACM Trans. Audio Speech Lang. Process.* 27 (6), 1085–1097.
- Sisman, B., Zhang, M., Sakti, S., Li, H., Nakamura, S., 2018c. Adaptive wavenet vocoder for residual compensation in gan-based voice conversion. In: *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, pp. 282–289.
- Soleymani, M., Garcia, D., Jou, B., Schuller, B., Chang, S.-F., Pantic, M., 2017. A survey of multimodal sentiment analysis. *Image Vis. Comput.* 65, 3–14.
- Staroniewicz, P., Majewski, W., 2009. Polish emotional speech database—recording and preliminary validation. In: *Cross-Modal Analysis of Speech, Gestures, Gaze and Facial Expressions*. Springer, pp. 42–49.
- Suni, A., Aalto, D., Raitio, T., Alku, P., Vainio, M., 2013. Wavelets for intonation modeling in hmm speech synthesis. In: *Eighth ISCA Workshop on Speech Synthesis*.
- Sutskever, I., Vinyals, O., Le, Q.V., 2014. Sequence to sequence learning with neural networks. In: *Advances in Neural Information Processing Systems*. pp. 3104–3112.
- Takeishi, E., Nose, T., Chiba, Y., Ito, A., 2016. Construction and analysis of phonetically and prosodically balanced emotional speech database. In: *2016 Conference of the Oriental Chapter of International Committee for Coordination and Standardization of Speech Databases and Assessment Techniques (O-COCOSDA)*. IEEE, pp. 16–21.
- Tanaka, K., Kameoka, H., Kaneko, T., Hojo, N., 2019. Atts2s-vc: Sequence-to-sequence voice conversion with attention and context preservation mechanisms. In: *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 6805–6809.
- Tao, J., Kang, Y., Li, A., 2006. Prosody conversion from neutral speech to emotional speech. *IEEE Trans. Audio Speech Lang. Process.*
- Teutenberg, J., Watson, C., Riddle, P., 2008. Modelling and synthesising f0 contours with the discrete cosine transform. In: *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, pp. 3973–3976.
- Tian, X., Chng, E.S., Li, H., 2019. A speaker-dependent WaveNet for voice conversion with non-parallel data. In: *Interspeech*. pp. 201–205.
- Tits, N., El Haddad, K., Dutoit, T., 2019a. Exploring transfer learning for low resource emotional tts. In: *Proceedings of SAI Intelligent Systems Conference*. Springer, pp. 52–60.
- Tits, N., Haddad, K.E., Dutoit, T., 2020. Ice-talk: an interface for a controllable expressive talking machine. [arXiv preprint arXiv:2008.11045](https://arxiv.org/abs/2008.11045).
- Tits, N., Wang, F., El Haddad, K., Pagel, V., Dutoit, T., 2019b. Visualization and interpretation of latent spaces for controlling expressive speech synthesis through audio analysis. In: *Proc. Interspeech 2019*, pp. 4475–4479.
- Toda, T., Black, A.W., Tokuda, K., 2007. Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory. *IEEE Trans. Audio Speech Lang. Process.* 15 (8), 2222–2235.
- Toda, T., Chen, L.-H., Saito, D., Villavicencio, F., Wester, M., Wu, Z., Yamagishi, J., 2016. The voice conversion challenge 2016. In: *Interspeech*. pp. 1632–1636.
- Um, S.-Y., Oh, S., Byun, K., Jang, I., Ahn, C., Kang, H.-G., 2020. Emotional speech synthesis with rich and granularized control. In: *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 7254–7258.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. *Adv. Neural Inf. Process. Syst.* 30, 5998–6008.
- Veaux, C., Rodet, X., 2011. Intonation conversion from neutral to expressive speech. In: *Twelfth Annual Conference of the International Speech Communication Association*.
- Veaux, C., Yamagishi, J., MacDonald, K., et al., 2016. Cstr vctk corpus: English multi-speaker corpus for CSTR voice cloning toolkit.
- Wang, Y., Ding, J., Meng, Q., Jiang, X., 2012. Multilingual emotion analysis and recognition based on prosodic and semantic features. In: *2011 International Conference in Electronics, Communication and Automatic Control Proceedings*. Springer, pp. 1483–1490.
- Wang, X., Takaki, S., Yamagishi, J., 2017. An RNN-based quantized F0 model with multi-tier feedback links for text-to-speech synthesis. In: *INTERSPEECH*. pp. 1059–1063.
- Whissell, C.M., 1989. The dictionary of affect in language. In: *The Measurement of Emotions*. Elsevier, pp. 113–131.
- Wu, C.-H., Hsia, C.-C., Lee, C.-H., Lin, M.-C., 2009. Hierarchical prosody conversion using regression-based clustering for emotional speech synthesis. *IEEE Trans. Audio Speech Lang. Process.* 18 (6), 1394–1405.
- Wu, D.-Y., Lee, H.-y., 2020. One-shot voice conversion by vector quantization. In: *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 7734–7738.
- Wu, Z., Xie, Z., King, S., 2019. The blizzard challenge 2019. In: *Proc. Blizzard Challenge Workshop, Vol. 2019*.
- Xu, Y., 2011. Speech prosody: A methodological review. *J. Speech Sci.* 1 (1), 85–115.
- Xue, Y., Hamada, Y., Akagi, M., 2018. Voice conversion for emotional speech: Rule-based synthesis with degree of emotion controllable in dimensional space. *Speech Commun.*
- Yamagishi, J., Onishi, K., Masuko, T., Kobayashi, T., 2005. Acoustic modeling of speaking styles and emotional expressions in hmm-based speech synthesis. *IEICE Trans. Inf. Syst.* 88 (3), 502–509.
- Yamagishi, J., Veaux, C., MacDonald, K., et al., 2019. Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit (version 0.92).
- Ye, H., Young, S., 2004. Voice conversion for unknown speakers. In: *Eighth International Conference on Spoken Language Processing*.
- Zen, H., Dang, V., Clark, R., Zhang, Y., Weiss, R.J., Jia, Y., Chen, Z., Wu, Y., 2019. LibriTTS: A Corpus derived from LibriSpeech for text-to-speech.
- Zhang, J.-X., Ling, Z.-H., Dai, L.-R., 2019a. Non-parallel sequence-to-sequence voice conversion with disentangled linguistic and speaker representations. *IEEE/ACM Trans. Audio Speech Lang. Process.* 28, 540–552.

- Zhang, J.-X., Ling, Z.-H., Liu, L.-J., Jiang, Y., Dai, L.-R., 2019b. Sequence-to-sequence acoustic modeling for voice conversion. *IEEE/ACM Trans. Audio Speech Lang. Process.* 27 (3), 631–644.
- Zhang, J., Tao, F., Liu, M., Jia, H., 2008. Design of speech corpus for mandarin text to speech. In: *The Blizzard Challenge 2008 Workshop*.
- Zhang, M., Wang, X., Fang, F., Li, H., Yamagishi, J., 2019. Joint training framework for text-to-speech and voice conversion using multi-source tacotron and wavenet. In: *Proc. Interspeech 2019*, pp. 1298–1302.
- Zhang, M., Zhou, Y., Zhao, L., Li, H., 2021. Transfer learning from speech synthesis to voice conversion with non-parallel training data. *IEEE/ACM Trans. Audio Speech Lang. Process.* 29, 1290–1302.
- Zhao, Y., Huang, W.-C., Tian, X., Yamagishi, J., Das, R.K., Kinnunen, T., Ling, Z., Toda, T., 2020. Voice conversion challenge 2020: Intra-lingual semi-parallel and cross-lingual voice conversion. *arXiv preprint arXiv:2008.12527*.
- Zhou, X., Ling, Z.-H., King, S., 2020. The Blizzard Challenge 2020. In: *Proc. Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge 2020*, pp. 1–18.
- Zhou, K., Sisman, B., Li, H., 2020. Transforming spectrum and prosody for emotional voice conversion with non-parallel training data. In: *Proc. Odyssey 2020 the Speaker and Language Recognition Workshop*, pp. 230–237.
- Zhou, K., Sisman, B., Li, H., 2021a. Limited data emotional voice conversion leveraging text-to-speech: Two-stage sequence-to-sequence training. *arXiv preprint arXiv:2103.16809*.
- Zhou, K., Sisman, B., Li, H., 2021b. Vaw-gan for disentanglement and recombination of emotional elements in speech. In: *2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, pp. 415–422.
- Zhou, K., Sisman, B., Liu, R., Li, H., 2021c. Seen and unseen emotional style transfer for voice conversion with a new emotional speech dataset. In: *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 920–924.
- Zhou, K., Sisman, B., Zhang, M., Li, H., 2020. Converting anyone's emotion: Towards speaker-independent emotional voice conversion. In: *Proc. Interspeech 2020*, pp. 3416–3420.
- Zhu, J.-Y., Park, T., Isola, P., Efros, A.A., 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2223–2232.
- Zhu, X., Yang, S., Yang, G., Xie, L., 2019. Controlling emotion strength with relative attribute for end-to-end speech synthesis. In: *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, pp. 192–199.
- Zovato, E., Pacchiotti, A., Quazza, S., Sandri, S., 2004. Towards emotional speech synthesis: A rule based approach. In: *Fifth ISCA Workshop on Speech Synthesis*.