Exemplar-Based Emotive Speech Synthesis

Xixin Wu, *Member, IEEE*, Yuewen Cao, Hui Lu, Songxiang Liu, Shiyin Kang, Zhiyong Wu, *Member, IEEE*, Xunying Liu, *Member, IEEE*, and Helen Meng, *Fellow, IEEE*

***Abstract*—Expressive text-to-speech (E-TTS) synthesis is important for enhancing user experience in communication with machines using the speech modality. However, one of the challenges in E-TTS is the lack of a precise description of emotions. Previous categorical specifications may be insufficient for describing complex emotions. The dimensional specifications face the difficulty of ambiguity in annotation. This work advocates a new approach of describing emotive speech acoustics using spoken exemplars. We investigate methods to extract emotion descriptions from the input exemplar of emotive speech. The measures are combined to form two descriptors, based on capsule network (CapNet) and residual error network (RENet). The first is designed to consider the spatial information in the input exemplary spectrogram, and the latter is to capture the contrastive information between emotive acoustic expressions. Two different approaches are applied for conversion from the variable-length feature sequence to fixed-size description vector: (1) dynamic routing groups similar capsules to the output description; and (2) recurrent neural network's hidden states store the temporal information for the description. The two descriptors are integrated to a state-of-the-art sequence-to-sequence architecture to obtain an end-to-end architecture that is optimized as a whole towards the same goal of generating correct emotive speech. Experimental results on a public audiobook dataset demonstrate that the two exemplar-based approaches achieve significant performance improvement over the baseline system in both emotion similarity and speech quality.**

***Index Terms*—Expressive speech synthesis, exemplary emotion descriptor, residual error, speech emotion recognition, capsule.**

I. Introduction

SPEECH-BASED communication is a hallmark of artificial intelligence. In recent years, we have seen the proliferation of applications using speech as the interaction medium, such as virtual assistants (e.g. Apple Siri, Microsoft's Cortana, Amazon's Alexa, etc.), voice search (e.g. Google's and Baidu's voice search) and call centers (e.g. Google Duplex). The spoken responses generated by these systems have diverse content, which calls for various expressive presentations to convey paralinguistic information such as emotions, intentions and speaking styles. Expressive spoken presentations of information can significantly improve user experiences for these applications [1], [2]. For example, when faced with user requests such as "Tell me a joke", "Tell me about romantic dinner venues nearby", and "Recite a poem for me", if the system responds only with a neutral speaking style, the user experience will be severely limited. On the other hand, if the synthetic speech response can convey different speaking styles (e.g. in telling jokes or reciting poems), different emotions (e.g. happy, satisfied, affectionate, etc.) and different intentions (e.g. making an inquiry, request, suggestion, confirmation, complaint etc.), the user experience will be greatly enhanced.

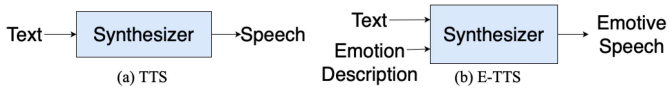


Fig. 1. Architecture of TTS and E-TTS system. (a) TTS: converting text to speech in neutral style; (b) E-TTS: converting text to speech with specified emotion. The E-TTS system needs not only the input text, but also the emotion specification to guide the synthesis model to generate speech with the target emotion.

Previous research efforts in expressive text-to-speech (E-TTS) synthesis aimed at generating natural and expressive speech with specified styles, which relate to emotions [3]–[7], intentions [8]–[11], emphasis (e.g. emphatic versus neutral) [12]–[14] and conversational characteristics [15]–[17]. The study of emotive synthesis attracts much attention due to its diverse applicability and the availability of corpora. Compared to neutral TTS (Fig. 1(a)) where text is the only input, E-TTS (Fig. 1(b)) requires an emotion specification as additional input, e.g. using categorical codes [18] or dimensional values [19], for synthesizing an emotive spoken response. However, specifying emotions is a challenge because both categorical descriptors and quantized dimensional descriptors are difficult to code. Furthermore, a specified emotion may be conveyed via a great variety of acoustic realizations. These challenges motivate our work in investigating approaches for emotion specification for speech synthesis. We advocate an exemplar-based approach to synthesizing emotive speech, as shown in Fig. 2. This approach describes emotion(s) via an utterance exemplar, instead of categorical codes or dimensional values. Emotive information is extracted from the utterance exemplar to be used in speech synthesis. In other words, the synthetic speech is designed

Manuscript received July 22, 2020; revised November 29, 2020 and January 14, 2021; accepted January 15, 2021. Date of publication January 18, 2021; date of current version February 13, 2021. This work was supported by the National Natural Science Foundation of China-Research Grants Council of Hong Kong (NSFC-RGC) joint fund (61531166002, N_CUHK404/15). The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Heiga Zen. *(Corresponding author: Zhiyong Wu.)*

Xixin Wu was with the Chinese University of Hong Kong. He is now with Engineering Department, Cambridge University, CB2 1PZ Cambridge, U.K. (e-mail: xw369@cam.ac.uk).

Yuewen Cao, Hui Lu, Songxiang Liu, Xunying Liu, and Helen Meng are with the Department of Systems Engineering and Engineering Mangement, The Chinese University of Hong Kong (e-mail: ywcao@se.cuhk.edu.hk; lu-h17@mails.tsinghua.edu.cn; sxliu@se.cuhk.edu.hk; xyliu@se.cuhk.edu.hk; hmmeng@se.cuhk.edu.hk).

Shiyin Kang is with Huya Inc., Guangzhou 511442, China (e-mail: shiyinkang@tencent.com).

Zhiyong Wu is with Shenzhen International Graduate School, Tsinghua University, Shenzhen 518055, China (e-mail: zywu@se.cuhk.edu.hk).

Digital Object Identifier 10.1109/TASLP.2021.3052688

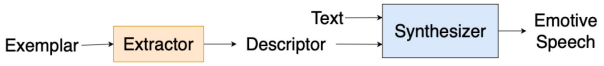


Fig. 2. Architecture of an exemplar-based emotive speech synthesis system. Speech is synthesized to mimic the emotion(s) in the exemplar. The emotive information is extracted from the exemplar and represented by descriptors and used as input for synthesis.

to mimic the emotions in the exemplar and there is no need for coding emotive information (subjectively) using categorical or dimensional descriptors. To realize exemplar-based emotive speech synthesis, we need to address several research questions: (i) What feature representation should be used for the utterance exemplar? (ii) What kind of emotion descriptors should be used? (iii) How can we map the feature description to the emotion descriptors? (iv) How can we synthesize emotive speech based on input descriptor values?

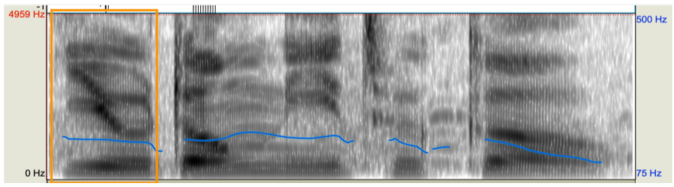
We will address these research questions in the following sections, after a review of previous work in Section II. In Section III, we will present an approach where spectrograms are used to encode spatial and contrastive information related to emotions in the utterance exemplar. We will use categorical codes and neural latent representations as descriptors of emotions. We also present two approaches to extract emotive information, based on capsule networks (CapNets) [20] and residual error networks (RENets) [21]. Experiments are presented and discussed in Section IV. Conclusions are drawn in Section V.

II. Related Work

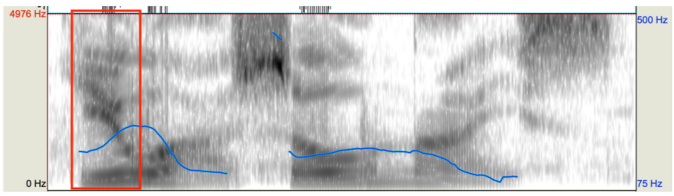
Previous work have described emotions explicitly using categorical codes or dimensional values. The categorical emotions are defined as a closed set of discrete basic emotions, e.g. *Anger*, *Disgust*, *Fear*, *Joy*, *Sadness* and *Surprise* [3], [22]. However, it may not be straightforward to encode complex emotions (e.g. a mix of emotions at different intensity levels) using categorical descriptors. Alternatively, the dimensional approach defines emotions as points within a dimensional space, e.g. the dimensions of pleasure-displeasure, arousal-nonarousal and dominance-submissiveness (PAD) [23]. Charfuelan and Steiner [24] used PAD to describe the expressive styles in audiobook data. Hodari *et al.* [19] used the dimensions of valence, activation and dominance (VAD) to describe emotions. Theoretically, any arbitrary emotion can be described by dimensional values. However, annotating data with these dimensions is difficult. Due to lack of standardized coding schemes and lack of effective methods to minimize subjective variability in human annotations, there can be significant inconsistencies in coding [25]. To address this problem in data annotation, researchers have explored semi-supervised and unsupervised methods to automatically learn emotion specifications from unlabeled data [24], [26]–[33]. For example, the K-means clustering algorithm has been applied to group speech samples into a predefined number of emotion classes [26], [30], [31]. The resulting emotion classes, however, are still a set of discrete categories and further work is needed before the categories can be used to describe complex emotions.

Given a specific emotion, there are many strategies in acoustic realizations in speech. The strategies may vary among speakers [34]–[36]. For example, among the two speakers investigated in [34], one used a different rhythm, i.e. the timing of syllables and number of silent segments, to express the emotions of *Angry*, *Surprise* and *Disgust*. Contrastively, the other speaker used other acoustic realizations, i.e. high energy for *Angry*, high pitch for *Surprise* and low pitch for *Disgust*, but kept the same rhythmic pattern across emotions. Still other strategies may be used in different interaction contexts, e.g. smaller F0 ranges in interviews; versus larger ranges in sports commentaries [37]. Hence, descriptors of emotions need precise encoding of emotions and their realization strategies. Previous efforts have been devoted to defining acoustic parameter rules to describe various acoustic realizations [34], [38]. For example, compared to the neutral emotion, the F0 mean of *Joy* needs to be increased by 50%, and the tempo is expected to be 30% faster [6]. The derived acoustic parameters can be used as emotion descriptors for the synthesis model. Meng *et al.* [39] perturbed the acoustic features of neutral speech to convey focus in the output expressive speech. Perturbations were applied to the F0 maximum, F0 minimum, F0 mean, F0 range, F0 slope, mean of RMS energy, and duration of each phone. For example, the maximum F0 of voiced phones is increased to generate emphatic speech. However, it is difficult, if not impossible, to specify all strategies for all emotions with manual derived parameters. Therefore, automatic methods based on neural networks were introduced, including unconditional control vectors [40], [41], bottleneck features [10], residual features [10], style token [42], [43], latent variables [44]–[46] and variational embedding [47]–[49].

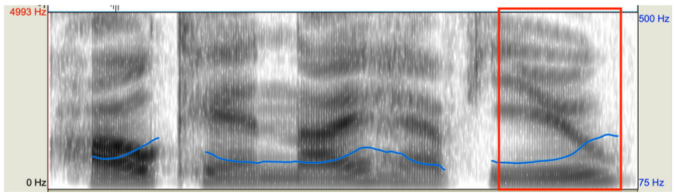
Recently, deep learning has demonstrated impressive effectiveness in E-TTS [18], [40], [41]. Neural expressive synthesis is as flexible as HMM-based expressive synthesis, while at the same time superior in modeling the correlation among the input contextual features. This is because the training data with different contextual features are not fragmented through clustering [50]. Neural E-TTS models can be classified into two types: (i) the two-step architecture [40], [41], [51] and (ii) the sequence-to-sequence (seq2seq) architecture [18], [32], [52]–[56]. The two-step architecture involves mapping the linguistic features augmented with an emotion specification to acoustic features in two separately optimized steps — namely, duration prediction and frame-level acoustic feature generation. Contrastively, the seq2seq architecture jointly optimizes both duration prediction and acoustic feature generation in one model with an attention mechanism. The attention weights are calculated based on the input sequence and the previously generated output sequence. The attention weights, which measure how much attention should be devoted to each step in the input sequence, can be considered as implicit duration modeling, or alignments between the input linguistic sequence and the output acoustic sequence. In this work, we adopt the seq2seq architecture, where the emotion specification, e.g. categorical codes, are passed to the seq2seq model to control the synthesis of emotive speech. Moreover,



(a) "You don't need to go." (neutral statement)

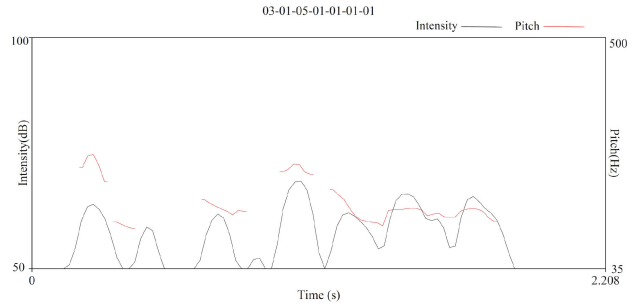


(b) "<emphatic You> are so nice!"

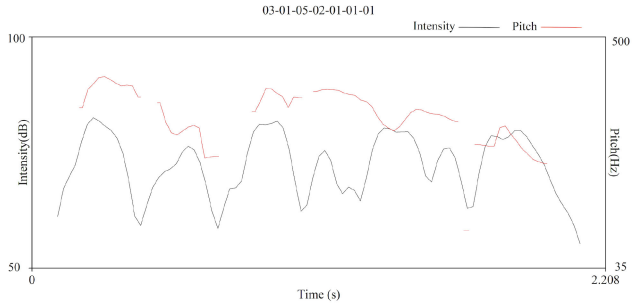


(c) "How can I help you?"

Fig. 3. Spectrogram examples with blue lines explicitly showing pitch contours corresponding to voiced parts. Utterance (a) is a neutral statement utterance without salient pitch rise. Utterance (b) shows pitch rise at the beginning to emphasize the word "you". Utterance (c) is a question that shows pitch rise at the end.



(a)



(b)

Fig. 4. Two utterances uttered by a single speaker with the text content of "The kids are talking by the door!". Both utterances are annotated with the emotion of *Anger*. However, we can observe different acoustic realizations – the pitch and intensity contour of utterance (a) has lower mean; while the contours of utterance (b) has larger mean and variance.

exemplary descriptors that are extracted from spoken utterances will be integrated into the seq2seq architecture to control the synthesis.

III. Approach

This section presents the approach for emotive speech synthesis that attempts to mimic the emotion(s) specified through an input utterance exemplar. The approach involves four stages, each addressing a research question presented in the introductory section.

A. Feature Representation of the Utterance Exemplar

We consider the spectrogram to be a desirable representation of the utterance exemplar that can preserve the time-frequency analysis in its entirety. The analysis is important for representing emotive information. For example, consider the statement, exclamation or question in the utterances shown respectively in Fig. 3(a) to 3(c). Fig. 3(a) is a neutral statement with a flat pitch contour. Fig. 3(b) is an exclamation with salient intonation rise in the word "you" at the beginning of the utterance. Fig. 3(c) is a question with salient pitch rise at the end of the utterance. Such spatial information can be captured well by the spectrogram. The utilization of prosodic features of pitch, energy and duration has also been investigated in previous work to improve the controllability of synthesis models [57]–[59]. However, the timbre information in spectrogram also plays an important role in emotions, e.g. harshness in the voice, harmonic saturation and formant distribution, etc.

B. Emotive Descriptors of the Utterance Exemplar

Next, we seek to find efficient descriptors for encoding emotive information from the utterance exemplars. One possible option is to use categorical codes, e.g. *Anger*, *Happiness* and *Sadness*, etc. However, there are several limitations in categorical codes. First, the definitions for the codes are vague due to the lack of standardized coding schemes. This brings challenges to synthesizing the corresponding emotive speech accurately. Second, it is not straightforward to encode complex emotions, e.g. mix of emotions at different intensity levels. Third, the acoustic realizations, corresponding to one categorical code, can be highly different, and the codes are insufficient for specifying the acoustic variation information that is necessary for the synthesis model. As an illustration, Fig. 4 shows two utterances that are uttered by a single speaker with same text content from RAVDESS corpus [60]. The emotion of the two utterances are both annotated as *Anger*, but the acoustic realizations of the two utterances are highly different. Both the pitch contour and intensity contour of utterance (a) have lower mean and variance values than those of utterance (b). This demonstrates that the highly varied acoustic realizations cannot be represented by categorical codes. To precisely synthesize these acoustic variations, a more accurate descriptor is desired. In this work, we also use the neural latent representation that is learnt from such

realization variations automatically. The latent representation is optimzied towards the final goal of synthesizing the highly varied acoustic realizations accurately.

1) SER-Based Categorical Descriptors: As mentioned above, categorical codes are widely used to specify emotions. It will be desirable to derive categorical codes from the utterance exemplar using speech emotion recognition (SER) techniques, which can predict the probabilities of different pre-defined emotion categories based on the input utterance. We will use capsule networks (CapNet) for SER, which offer special advantages in capturing the spatial information in spectrograms for speech emotion recognition (Sec. III-C1). The CapNet outputs a set of probabilities corresponding to the categorical emotions, which is converted into a one-hot emotion code vector (EC) by identifying the emotion with maximum probability. An alternative descriptor for the exemplar is to directly use the probability values output by the SER model for the various emotions, which we will denote as EP. Another option is to use the raw logit values (before softmax layers) generated by the SER model. We will denote the logit-based descriptor as EL.

2) Neural Descriptor: While SER trained well on data labeled with categorical codes should provide reasonable performance, the use of single emotion categories may not be sufficient for describing utterances carrying complex emotions (e.g. a mix of emotions at varying degrees). One may consider the use of a dimensional descriptor as an alternative, but the dimensional values are difficult to annotate with consistency. Hence, we propose the use of a latent representation which is automatically derived from the utterance exemplar using neural methods. Residual error networks (RENets) are adopted because they present the advantages of explicitly capturing the contrastive information across emotions in spectrograms and automatically optimizing a latent representation (Sec. III-C2). This automatically derived latent emotive representation will be denoted as EA. The methods for generating EC, EP, EL and EA will be elaborated in the next subsection.

C. Mapping Signal Representation to Emotive Descriptors

This section elaborates on CapNet, i.e. capsule networks for speech emotion recognition [20], [61], [62] and how they are superior to the commonly used convolutional neural networks (CNNs) for capturing spatial information related to emotions in the utterance exemplars. We will also elaborate on RENet [21], i.e. residual error networks for deriving an emotive latent representation from the utterance exemplar by explicitly capturing the constrastive information across emotions from spectrograms.

1) CapNet: While CNNs have been a popular network structure for extracting information from spectrograms, we consider that they have two limitations in capturing spatial and contrastive information for speech emotion recognition (SER). First, the neurons in the convolutional layers output a scalar, which only provides the probability that the feature pattern (e.g. intonation rise) matches the kernel. However, the more detailed instantiation parameters (e.g. position) are ignored. As shown in Fig. 5, the shared kernel is applied to various parts of the spectrogram. When the feature pattern (e.g. intonation rise) matches

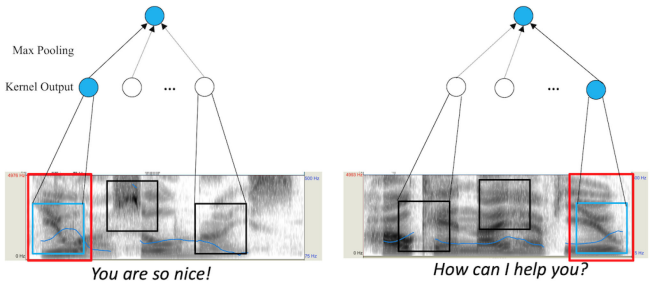


Fig. 5. CNN perception of intonation rise (highlighted with red rectangles) at different positions. The shared kernel is applied to various parts of the input, as shown by the squares. When the kernel is applied to the intonation rises, the output will be activated (highlighted as blue). Although the activated outputs come from different parts of the input spectrogram, the max-pooling operation produces the same result, which hinders the accurate classification of the emotional (left) and neutral (right) utterances.

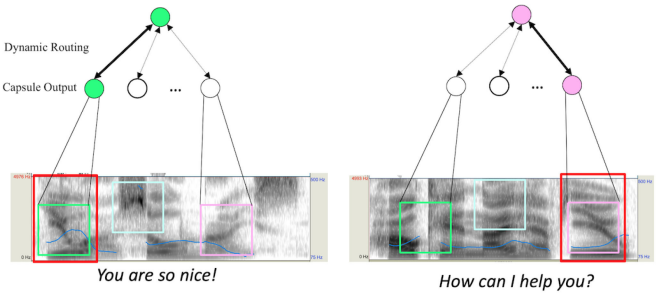


Fig. 6. Capsule perception of intonation rise (highlighted with red rectangles) at different positions. The intonation rises at different positions produce different capsule outputs, which are highlighted as green and purple to reflect the position difference. The dynamic routing then passes the capsule outputs containing the position information to the upper layer. Hence, the final distinctive outputs containing the position information of intonation rise support the accurate emotion classification.

the kernel, the output is activated (and highlighted in blue in the figure), regardless of the positional information of feature (e.g. at the beginning versus the end). Second, the max-pooling layer discards all but the most activated neuron, and the spatial relationship across neurons are lost when the activated neurons are passed to the upper layers. As illustrated in Fig. 5, activated outputs in different positions are selected by max-pooling. The final CNN outputs of the exclamation and the question (see left and right parts of Fig. 5) are the same and thus the emotive and neutral characteristics are not distinguished. To address this issue of spatial information loss, we propose to use CapNets – the neuron that output a scalar in CNN is replaced with a capsule, i.e. group of neurons which output a vector containing the instantiation information with the pose and position of the recognized pattern. Furthermore, the max-pooling layer is replaced with a dynamic-routing algorithm, which routes all capsules that are in various positions to upper layers without information loss. As shown in Fig. 6, intonation rise at different spatial positions produce different capsule outputs, as highlighted in green and purple. The position-aware capsule outputs are passed to the upper layer via dynamic routing. Distinguishing the positional information in the final output is important for accurate classification of emotions.

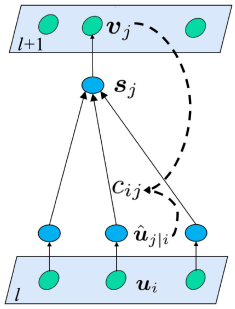


Fig. 7. Capsule structure. Each capsule is a group of neurons. The connections between layers are established by dynamic routing algorithm.

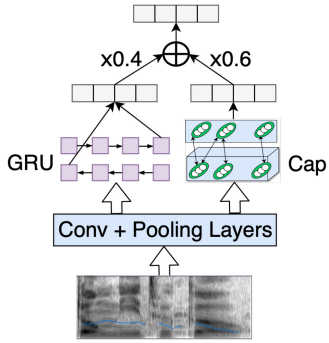


Fig. 8. Architecture of the CapNet-based speech emotion recognition system.

To connect between capsule layers, the dynamic routing algorithm (also known as routing-by-agreement), is applied to learn the hierarchical relationship between features in consecutive layers. The algorithm works as follows: Assume that the i-th capsule in layer l is denoted as $\boldsymbol{u}_i$, and the j-th capsule in layer $l+1$ as $\boldsymbol{v}_j$, referring to Fig. 7. The $\boldsymbol{u}_i$ is first projected to the space of $\boldsymbol{v}_j$ by

$$\hat{\boldsymbol{u}}_{j|i} = \boldsymbol{W}_{ij}\boldsymbol{u}_i + \boldsymbol{b}_{ij}, \tag{1}$$

where $\boldsymbol{W}_{ij}$ and $\boldsymbol{b}_{ij}$ are weight matrix and bias vector, and they are both position-aware and trainable. To obtain the capsule $\boldsymbol{v}_j$ in the upper layer, the procedure described by (2)–(5) is iterated for a predefined number of times n, with the initial value of $d_{ij} = 0$:

$$c_{ij} = \frac{\exp(d_{ij})}{\sum_k \exp(d_{ik})}, \tag{2}$$

$$\boldsymbol{s}_j = \sum_i c_{ij}\hat{\boldsymbol{u}}_{j|i}, \tag{3}$$

$$\boldsymbol{v}_j = \frac{||\boldsymbol{s}_j||^2}{1+||\boldsymbol{s}_j||^2}\frac{\boldsymbol{s}_j}{||\boldsymbol{s}_j||} \tag{4}$$

and

$$d_{ij} \leftarrow d_{ij} + \hat{\boldsymbol{u}}_{j|i} \cdot \boldsymbol{v}_j, \tag{5}$$

where $\cdot$ denotes the dot product. The c_{ij} is the coupling coefficient that measures the agreement between $\boldsymbol{v}_j$ in the upper layer and $\hat{\boldsymbol{u}}_{j|i}$ projected from $\boldsymbol{u}_i$. Hence, this algorithm is also called routing-by-agreement.

We have built a CapNet-based system for speech emotion recognition (SER) from utterance exemplars. The recognized emotion is used subsequently in the descriptors EC, EP and EL [62]. The SER system is illustrated in Fig. 8. The input frame sequence (i.e. spectrogram) is first sliced into overlapping windows, and the shared capsule layers are applied to each window for parameter reduction. In each window, several separated convolutional layers shared across windows are applied to the input to obtain primary capsules. The neurons of different channels at the same position along the width- and height-axes of output feature map of the convolutional layer are grouped together to form a capsule. The primary capsules are routed to generate window-level capsules $\{\boldsymbol{v}_t\}_{t=1}^{M}$. The utterance-level capsules are then obtained with utterance-level routing based on the output vectors $\boldsymbol{o}$ of these windows, which are defined as:

$$\boldsymbol{o} = [\boldsymbol{v}_1^\top, \ldots, \boldsymbol{v}_M^\top, ||\boldsymbol{v}_1||, \ldots, ||\boldsymbol{v}_M||]. \tag{6}$$

The window output vector consists of the orientations and lengths of all the M capsules in one window, since both are important for utterance-level emotion recognition. Though the length information $||\boldsymbol{v}||$ is redundant given the vectors $\boldsymbol{v}$, we intend to provide this information explicitly to save learning effort of the network. To further capture temporal information, we add a branch of gated recurrent unit (GRU) on the convolutional layers. The two branches of GRU and capsules are combined with heuristic weights (set at 0.6).

2) RENet: The residual error network (RENet) is designed to encode residual error information into the emotion representation vector. The residual error is defined as the difference between acoustic feature sequence (i.e. spectrogram) of emotive utterance $\boldsymbol{y}^{(e)} = \{\boldsymbol{y}_1^{(e)}, \boldsymbol{y}_2^{(e)}, \ldots, \boldsymbol{y}_{T_e}^{(e)}\}$ and that of neutral utterance $\boldsymbol{y}^{(n)} = \{\boldsymbol{y}_1^{(n)}, \boldsymbol{y}_2^{(n)}, \ldots, \boldsymbol{y}_{T_n}^{(n)}\}$, corresponding to the same lingusitic feature sequence (i.e. textual) $\boldsymbol{x} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N\}$. In this way, the RENet is very desirable for extracting contrastive information between neutral and emotive speech. However, there is difficulty in finding an utterance pair with the same textual content but different contrastive acoustic realizations in the training dataset. Also, if the two utterances have different durations, we will need to find their alignment before calculating the residual error based on the aligned feature sequences. To circumvent these two difficulties, we first generate parallel data using an externally trained seq2seq neutral TTS model. However, it is challenging to obtain sufficient neutral data with the same recording condition as the emotive data for training. In this work, we use an average-emotion TTS model to generate the "neutral" data. The model is optimized to generate emotive utterances given only the text without any emotion specification. Hence the model tends to generate an average of the emotions in the emotive training data. This model generates every step based on the attention alignment between the input linguistic features $\boldsymbol{x}$ and the previous generated acoustic features $\hat{\boldsymbol{y}}_{\{1:t-1\}}^{(n)}$,

$$\hat{\boldsymbol{y}}_t^{(n)} = f\left(\hat{\boldsymbol{y}}_{\{1:t-1\}}^{(n)}, \boldsymbol{x}, \alpha\left(\hat{\boldsymbol{y}}_{\{1:t-1\}}^{(n)}, \boldsymbol{x}\right)\right), \tag{7}$$

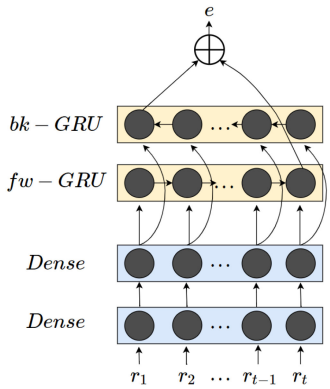


Fig. 9. Structure of the RENet-based extractor.

where $\alpha(\cdot)$ denotes the attention alignments and $f(\cdot)$ the seq2seq structure. To make the generated neutral utterance aligned to the emotive utterance, one option is to feed the emotive feature sequence $\boldsymbol{y}^{(e)}$ to the seq2seq model as

$$\hat{\boldsymbol{y}'}_t^{(n)} = f\left(\boldsymbol{y}_{\{1:t-1\}}^{(e)}, \boldsymbol{x}, \alpha\left(\boldsymbol{y}_{\{1:t-1\}}^{(e)}, \boldsymbol{x}\right)\right). \tag{8}$$

Eq. (8) is also referred to as teacher-forcing generation in some literature [63], [64]. In this way, we obtain two sequences $\boldsymbol{y}^{(e)}$ and $\hat{\boldsymbol{y}'}^{(n)}$ that are aligned to each other. Another option is to obtain the alignments from the emotive utterance $\boldsymbol{y}^{(e)}$ and use them during inference to obtain the average emotion utterances

$$\hat{\boldsymbol{y}''}_t^{(n)} = f\left(\hat{\boldsymbol{y}''}_{\{1:t-1\}}^{(n)}, \boldsymbol{x}, \alpha\left(\boldsymbol{y}_{\{1:t-1\}}^{(e)}, \boldsymbol{x}\right)\right). \tag{9}$$

since the alignment is calculated from the emotive utterance, therefore the generated average-emotion utterances $\hat{\boldsymbol{y}'}^{(n)}$ and $\hat{\boldsymbol{y}''}^{(n)}$ are aligned to $\boldsymbol{y}^{(e)}$. With the aligned utterances, the residual error $\boldsymbol{r}$ can then be calculated frame-by-frame as

$$\boldsymbol{r}_t = \boldsymbol{y}_t^{(e)} - \hat{\boldsymbol{y}'}_t^{(n)}, t = 1, 2, \ldots, T_e. \tag{10}$$

The RENet, as shown in Fig. 9, is then used to encode the residual error sequence $\boldsymbol{r}$ into a residual error embedding (REE) $\boldsymbol{e}$. The RENet consists of multiple dense layers and one bi-directional GRU layer from bottom to top. The dense layers are designed to transform the input residual error into the feature space that is more appropriate for providing emotive information for the synthesizer model. The GRU layer is used to capture temporal information in the residual error sequence. To improve the robustness against noise in the residual error sequence, the outputs of the dense layers are dropped out with a certain rate (e.g. 0.5), as applied similarly in [53]. The hidden state values of the forward direction GRU (fw-GRU) at the last time step and those of the backward GRU (bk-GRU) at the first time step are concatenated to obtain the REE. The REE based on residual error sequence calculated using teacher forcing in Eq. (8) and that using emotive alignments in Eq. (9) will be used as the EA and EAli descriptor, respectively.

As an illustration of how the learned REE control the synthesized prosodic variations, we calculate Pearson correlation coefficients between each dimension of the embeddings and the mean F0 of all training samples. As shown in Fig. 10, the 23-rd dimension of the embedding has strong negative correlation with the mean F0 values. To verify the controllability of the embedding, we compare the F0 trajectories of the synthetic speech when the embedding varies only in the 23-rd dimension. We first extract an embedding vector, denoted as emb-0, from a random reference sample in the training set. Then we add $+0.2$, -0.2, -0.4 and -0.6 to the 23-rd dimension of emb-0 to obtain four new embeddings, denoted as $emb + 0.2$, emb-0.2, emb-0.4 and emb-0.6. We then use these embeddings to generate audio samples. The smoothed F0 trajectories (linear interpolation in unvoiced frames) of synthetic speeches are shown in Fig. 11. We can observe that the F0 trajectories increase as the value of the 23-rd dimension decreases.[1] To control the F0 trajectory heights described by REE, we can tune the value in the 23-rd dimension, which effectuates the synthesis of the corresponding trajectories.

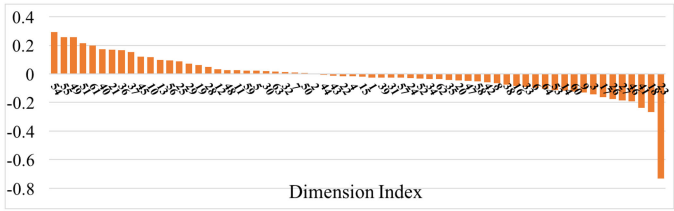


Fig. 10. Pearson correlation coefficients between each dimension of the style embeddings and the mean F0 values of training data.

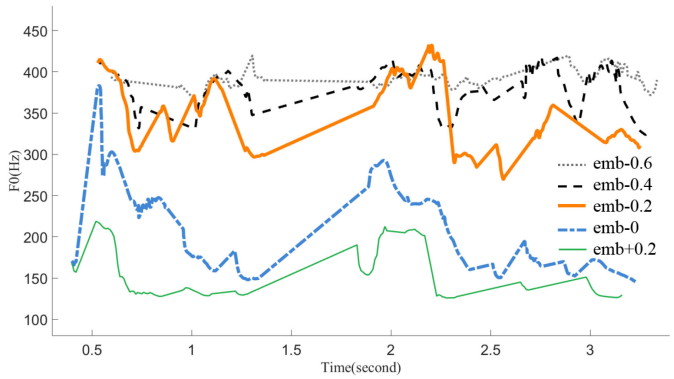


Fig. 11. Smoothed F0 trajectories of the manipulated embeddings.

D. Emotive Speech Synthesis

Having extracted emotive descriptors, including the categorical descriptors (i.e. EC, EP and EL) from SER, and the automatically derived descriptor (i.e. EA and EAli) from RENet, we follow through with emotive speech synthesis.

By setting the input exemplar to be the same as the target utterance, we train the system to synthesize output speech with the emotion(s) specified by the exemplar. In this way, the input emotive descriptor and the output speech are consistent in the emotion and the strategy of expression. The training process is

[1]Samples are available in "http://www1.se.cuhk.edu.hk/~wuxx/TASLP/ExemplarTTS.html "

to optimize the target seq2seq TTS model using the loss function

$$\mathcal{L}^{(e)} = \frac{1}{T_e}\sum_{t=1}^{T_e} \parallel \boldsymbol{y}_t - \hat{\boldsymbol{y}}_t^{(e)} \parallel_2^2, \tag{11}$$

where $\boldsymbol{y}$ and $\hat{\boldsymbol{y}}^{(e)}$ are the ground-truth emotive acoustic sequence and the corresponding generated sequence. In EC-TTS, EP-TTS and EL-TTS, the CapNet and the target model are trained separately on two different corpora. The CapNet is first trained on SER corpus, and then applied to the utterance exemplar to extract the descriptors. The target model is then optimized to generate emotive speech based on the extracted decriptors using another TTS corpus. In EA-TTS and EAli-TTS, the RENet is jointly trained with the neutral TTS model and the target TTS model using the same TTS corpus, in order to improve the alignment accuracy. The joint loss function is

$$\mathcal{L} = \mathcal{L}^{(n)} + \mathcal{L}^{(e)}, \tag{12}$$

where $\mathcal{L}^{(n)}$ is the loss function term corresponding to the average-emotion TTS model:

$$\mathcal{L}^{(n)} = \frac{1}{T_n}\sum_{t=1}^{T_n} \parallel \boldsymbol{y}_t - \hat{\boldsymbol{y}}_t^{(n)} \parallel_2^2 . \tag{13}$$

The average-emotion model only receives textual input without additional emotive information. Hence, it tends to generate acoustic features with the average emotion [65].

IV. Experiments

A. Corpus

We train the SER models using the interactive emotive dyadic motion capture (IEMOCAP) database [66], which consists of five sessions, with two speakers in each session. We adopt five-fold cross validation as [67] — 8 speakers from four sessions in the corpus are used as training data. One speaker from the remaining session is used as validation data, and the other one as test data. Only the improvised data is used. The spectrograms are extracted with 40-ms Hanning window, 10-ms shift and DFT of length 1600 (for 10 Hz grid resolution). In this work, we use the four emotion categories of *Happy*, *Angry*, *Sad* and *Neutral*.

The proposed E-TTS systems are evaluated on the audiobook corpus from Blizzard Challenge 2016, which is recorded by a native female speaker [68]. The speaker tries to utter in different styles in the recording, including emotions, mimicked role characters' voice. There are 50 books in the audiobook data. We use the book "A Midsummer Night's Dream" as testing data (around 0.35 hours), and the other 49 books as training data (around 4.79 hours). We extract the 1025-dimension Logarithmic magnitude linear-scale spectrograms and 80-band Mel-scale spectrograms with 50-ms Hanning window, 12.5-ms shift, and 2048-point Fourier transform [45].

B. Evaluation Criteria

1) SER Evaluation: We use two common evaluation metrics for performance comparison across various systems:

- Weighted Accuracy (WA) – the accuracy of all samples in the test data.
- Unweighted Accuracy (UA) – the average of class accuracies in the test set.

This reflects the accuracy of the extracted categorical descriptor:

$$\text{WA} = \frac{\sum_{i=1}^{K} P_i}{\sum_{i=1}^{K} U_i}, \tag{14}$$

$$\text{UA} = \frac{\sum_{i=1}^{K} P_i/U_i}{K}, \tag{15}$$

where P_i is the number of utterances with correct prediction of emotion i, U_i is the number of utterances with actual emotion i, and K is the number of emotions tested.

2) E-TTS Evaluation: For the objective evaluation of E-TTS systems, we calculate the mean squared error (MSE) between the teacher-forced generated Mel-spectrograms with the actual Mel-spectrograms on the test set.

For subjective evaluation, we use mean opinion scores (MOS) for evaluating speech quality and the emotive expressions of the above systems. 16 utterances are randomly selected from the testing data and synthesized by the seven systems respectively, thus we have 112 utterances to be evaluated. For the six systems requiring exemplar for emotion specification, another utterance is randomly selected from the testing data as the exemplar. We invite 19 participants without listening impairment to participate in the tests.[2] To get MOS on speech quality, each subject listens to each utterance and scores using a 5-point Likert scale on speech quality (5: excellent, 4: good, 3: fair, 2: poor, 1: bad). For evaluation on emotion similarity, the subjects are required to listen to a set of seven utterances generated respectively by the seven systems, in addition to the utterance exemplar, and then provide a 5-point score for the emotion similarity between the generated speech and the utterance exemplar (5: very similar, 4: similar, 3: somewhat similar, 2: different, 1: very different). The order of the seven generated utterances is randomized.

C. Network Configurations

This section describes the network configurations of the experimental systems, including two baseline systems and the five proposed E-TTS systems.

1) Baseline Systems: The baseline model Tacotron is a seq2seq-based system that consists of a CBHG encoder, an RNN decoder, an attention module, a Pre-net module and a Post-net component [53]–[55], as shown in Fig. 12. Tacotron does not require emotion specification and generates speech with the average emotion.

The encoder consists of three parts: the embedding lookup layer, the dense layers and the CBHG module. The embedding lookup layer is used to transform the input character sequence of one-hot vectors into an embedding sequence of continuous vectors. The embeddings are retrieved from the embedding lookup table by multiplying the table with the corresponding one-hot

[2]Subjective evaluations conducted on Amazon Mechanical Turk can be found in our previous work in [21]

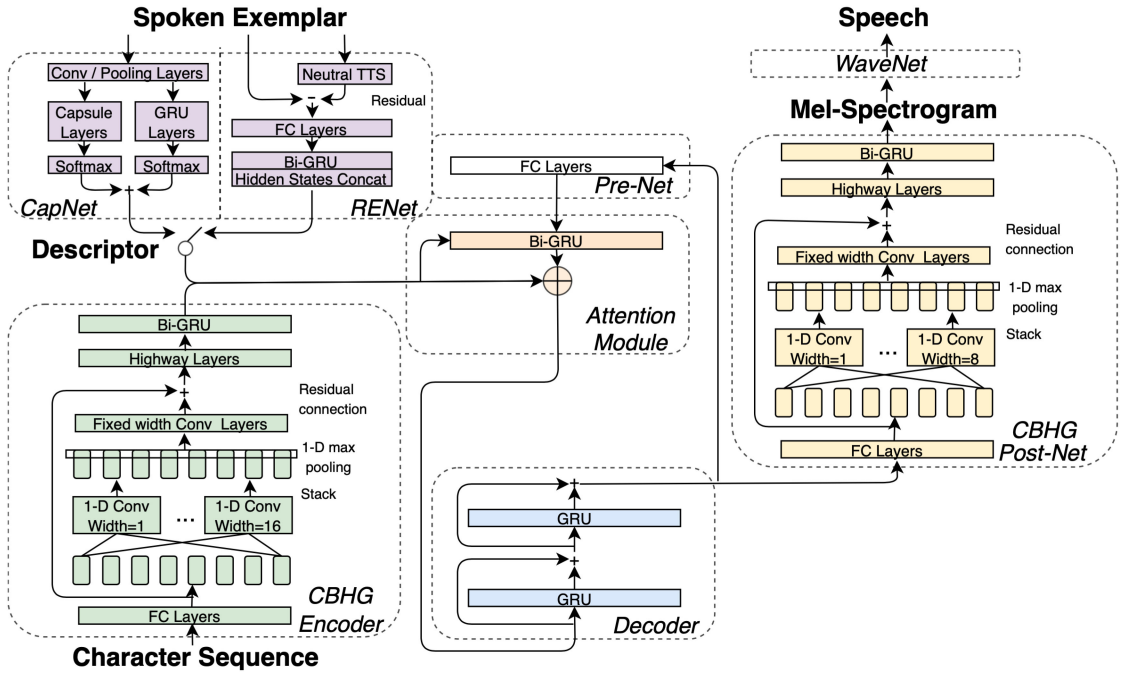


Fig. 12. System architecture based on the seq2seq structure Tacotron and exemplary emotion descriptors, which can be obtained from CapNet or RENet.

vector. The table is trained together with the whole model using backpropagation. Two dense layers with 256 units activated by ReLU activation, and dropped out with the rate of 0.5, are used to further transform the embeddings into hidden representations. The CBHG module consists of a bank of 1-dimension convolutional filters, followed by highway networks [69] and a bidirectional GRU layer. The filter bank contains 16 sets of filters with various widths from 1 to 16. Different widths of filters are used to capture different lengths of context. The convolution outputs are stacked together and further max-pooled along time to increase local invariance. The pooling window stride is set to 1 to preserve the original time resolution. The pooling outputs are fed to fixed-width 1-dimension convolutional layers and added with the original input sequence using residual connections. The outputs from the residual connections are fed to a 4-layer highway network to extract high-level features. A bidirectional GRU layer with 128 cells per direction is stacked upon the highway network to capture the temporal information.

The decoder is composed of two unidirectional GRU layers with 256 cells. The outputs of the GRU layers are added with the original inputs via residual connections [70]. The input to the decoder is the attention-based weighted summation of the encoder outputs, i.e. the bidirectional GRU layer outputs. The attention weights are calculated based on the encoder outputs and the Pre-net outputs. The Pre-net, which is used to transform the output Mel-scale spectrogram into hidden representation sequence for attention calculation, consists of two dense layers. The two dense layers with 256 and 128 units are activated by ReLU activation, and dropped out with the rate of 0.5.

The Post-net is utilized to further improve the Mel-spectrogram output from the decoder. The waveform is generated with the obtained Mel-spectrogram using the parallel WaveNet [71], [72]. The Post-net is another CBHG module with 8 sets of filters. The filter widths are from 1 to 8.

Another baseline system is the incorporation of global style token (GST) into Tacotron, denoted as GST-Tacotron [42]. The GSTs are a set of embeddings that are combined using attention mechanisms to generate emotive embeddings and jointly trained with the Tacotron. The GSTs are randomly initialized before training. The utterance exemplar is fed to a reference encoder that consists of a stack of six 2-dimensional convolutional layers and a GRU layer. The last hidden state of the GRU layer is projected to the space of GSTs with an content-based tanh attention module. The projected embedding, i.e. the weighted combination of GSTs with weights generated from the attention module, is used the emotive descriptor. We use a 4-head attention as [42].

2) EC-TTS, EP-TTS and EL-TTS: In the CapNet-based SER model, 8 separated convolutional layers with kernel size of 5×5 and channel number of 8 are applied to the CNN component consisting of 4 convolutional layers [62]. Then, for each position in the outputs of the 8 convoutional layers, the units along each channel direction are combined together to obtain capsules with size of 8 (i.e. the channel number). These capsules are then routed to the consequent window-level capsule layer with 8 capsules of size 8 in each window. An utterance-level routing is applied to the window output vectors to produce 4 utterance-level capsules with size of 16. The window used to slice the input matrix is set at size 40 input steps with shifts of 20 steps. The iteration number of the routing algorithm is set to 3. The outputs of the GRU layers and those of the capsule components are fed to separate sets of dense layers and softmax layer.

The categorical emotion descriptors extracted from the utterance exemplar are integrated to the seq2seq model that has the

TABLE I
WA AND UA OF SPEECH EMOTION RECOGNITION ACROSS VARIOUS SYSTEMS

System	WA(%)	UA(%)
CNN [67]	66.1	56.6
CNN_LSTM [67]	68.80	59.40
RNN-ELM [74]	62.85	63.89
CNN_TF_Att.pooling [75]	71.75	**68.06**
CapNet	**72.73**	59.71

TABLE II
PARAMETER NUMBERS OF CAPNETS COMPARED TO THE BASELINE OF CNN_GRU

Systems	WA(%)	UA(%)	Parameter Number
CNN_GRU	67.02	51.84	833,012
CNN_Cap	69.86	56.71	703,540
CNN_GRU-Cap	72.73	59.71	1,523,448

same structure as the Tacotron system. The emotion descriptors of EC, EP and EL are repeated and added to each step of the encoder outputs as [42], [73], called "broadcast," as in Fig. 12.

3) EA-TTS and EAli-TTS: The residual error encoder is implemented as two dense layers with 128 units per layer, activated by ReLU, dropped out with rate of 0.5, and one bi-directional GRU layer with 32 memory blocks in each direction. This generates the 64 dimensional EA and EAli descriptor, with 32 dimensions for each direction of the GRU layer. The seq2seq model has the same configuration as the Tacotron baseline.

D. Results

1) CapNet-Based SER Performance: The CapNet-based SER model achieves performance comparable to state-of-the-art systems, as shown in Table I. This validates the effectiveness of capsule structure in capturing the spatial features in spectrograms. Table II shows the performances and parameter numbers of the baseline system of CNN_GRU and two possible CapNet structures: CNN_Cap and CNN_GRU-Cap. As shown in Fig. 8, CNN_GRU-Cap combines the two branches of GRU and Cap on top of the convolutional and pooling layers. The CNN_Cap outperforms the CNN_GRU, but requires less parameters. This demonstrates the effectiveness of the capsule structure. The combination of GRU and capsule incorporates the advantages of both structures and achieves better results than each individual.

2) Exemplary Feature Comparison: In this section, we compare two different types of exemplary features, the spectrogram and the prosodic features of pitch, energy and duration. Based on the GST-Tacotron system, we build four systems using the following features as reference: (i) spectrogram (GST-Tacotron); (ii) pitch contour (P-GST-Tacotron); (iii) pitch and energy contours (PE-GST-Tacotron); and (iv) pitch contour, energy contour and word duration sequence (PED-GST-Tacotron).[3] The architectures of (ii)–(iv) are the same as (i), except that separate reference encoders are applied to pitch, energy and duration features respectively and the encoder outputs are concatenated to obtain the utterance-level embedding. The convolutional layers in the reference encoders of pitch, energy and duration are changed to 1 dimension. We use the above four systems,

[3]The duration sequence is obtained using Montreal Forced Aligner. https://github.com/MontrealCorpusTools/Montreal-Forced-Aligner

TABLE III
PEARSON CORRELATION COEFFICIENTS BETWEEN MEAN ENERGY OF THE EXEMPLARY UTTERANCES AND THE SYNTHESIZED UTTERANCES THAT ARE SYNTHESIZED BY GST-TACOTRON SYSTEMS USING DIFFERENT EXEMPLARY FEATURES (RANGES INDICATE 95% CONFIDENTIAL INTERVALS)

Exemplary Feature	Correlation	p-value	Emotion Similarity
Spectrogram	0.45	0.0089	3.35±0.19
Pitch	0.10	0.5927	2.39±0.18
Pitch+Energy	0.49	0.0035	2.54±0.19
Pitch+Energy+Duration	0.56	0.0007	2.36±0.20

TABLE IV
OBJECTIVE AND SUBJECTIVE EVALUATION RESULTS OF VARIOUS SYSTEMS (RANGES INDICATE 95% CONFIDENTIAL INTERVALS)

Systems	MSE(dB)	Speech Quality	Emotion Similarity
Tacotron	0.610	3.10±0.19	2.25±0.25
GST-Tacotron	0.591	3.55±0.17	3.09±0.17
EC-TTS	0.601	3.68±0.17	2.82±0.22
EL-TTS	0.575	**3.88±0.18**	2.86±0.24
EP-TTS	0.585	3.64±0.17	3.07±0.20
EAli-TTS	0.897	2.26±0.20	2.56±0.17
EA-TTS	**0.549**	3.50±0.17	**3.25±0.19**

GST-Tacotron, P-GST-Tacotron, PE-GST-Tacotron and PED-GST-Tacotron, to synthesize speech with the same text content, while conditioned on different utterance exemplars. The Pearson correlation coefficients between the mean energy values of the utterance exemplars and that of the generated utterances are calculated and shown in Table III, where the p-values reflect the test for non-correlation. We also conduct subjective evaluation to compare the emotion similarity performances of the sytems using different exemplary features. As can be observed from Table III, feeding energy information to GST-Tacotron can significantly improve the correlation between the synthesized utterances and the exemplary utterances, in both the explicit way of energy contour and the implicit way of spectrogram. However, the system using spectrogram demonstrates advantages in emotion similarity of the synthesized speech. Recent work has shown possible ways to improve the emotion similarity performance with these explicit prosodic features, for example, via variational auto-encoder (VAE) [57] or secondary attention mechanism [76]. In the following, we will focus on the systems using spectrograms as exemplary features.

3) Objective Evaluation on E-TTS: For the E-TTS systems, we first investigate the objective evaluation based on the criteria of teacher-forcing output Mel-spectrogram MSE. Results can be found in Table IV. All the E-TTS systems, except EAli-TTS, outperform the baseline Tacotron, which demonstrates that the utterance exemplars help the model generate spectrograms that are closer to the target spectrograms. The EAli-TTS generates poor quality speech, and the corresponding MSE value is much larger than Tacotron and the other systems, due to the mismatch between training and inference. The EP-TTS and EL-TTS system provide smaller MSE than the EC-TTS system. A possible reason is that the EP and EL carries richer information (i.e. the SER model's confidence of recognizing the four emotions) than the EC that only provides the identity of the most confident emotion. The EA-TTS system achieves the best performance compared to the other two systems with SER-based exemplar specifications, which reflects the superiority of the neural latent

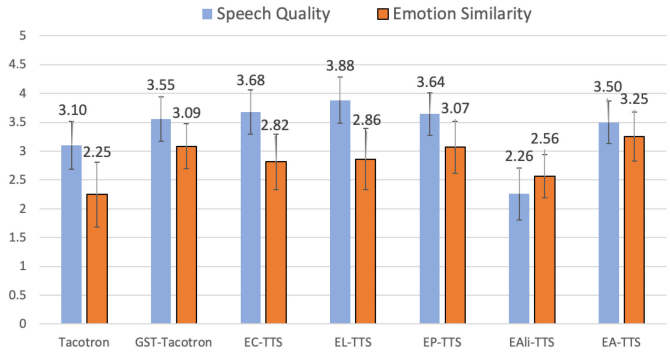


Fig. 13. Subjective evaluation of various systems.

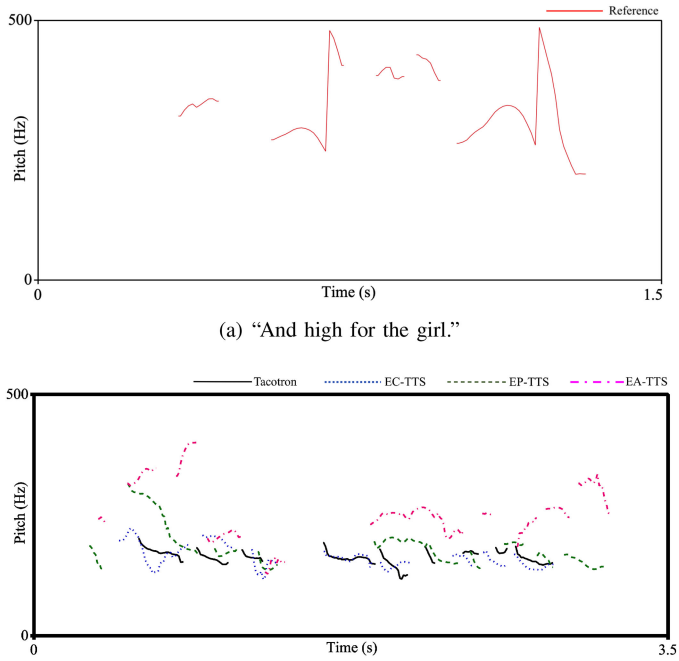
emotion specification. It should be noted that the MSE values mainly reflect relative performance across systems, which is more relevant for model tuning. The final performance evaluation needs to rely on the further subjective evaluations.

4) Subjective Evaluation on Speech Quality: Results of the MOS test on speech quality are shown in Table IV and Fig. 13. All the E-TTS systems, except EAli-TTS, outperform the average emotion system Tacotron ($p < 0.01$). One possible reason is that the training data carries distributions of various emotions. Without the specification from the exemplar, the Tacotron system tends to predict the averaged distribution. While the E-TTS systems utilize the emotion information provided by the utterance exemplar and fit the network weights for each distribution more accurately. The EC-TTS, EP-TTS and EA-TTS systems achieve slightly better (not significantly) or comparable perforamce with the baseline GST-Tacotron system, and the EL-TTS system significantly outperforms the GST-Tacotron system ($p < 0.01$).

5) Subjective Evaluation on Emotion Similarity: Table IV shows the results for MOS on emotion similarity. All E-TTS systems, except EAli-TTS, outperform the baseline ($p < 0.01$). which demonstrates the effectiveness of exemplar-based emotion specification for generating emotive speech. The EA-TTS system achieves the best performance, which demonstrates the advantage of using the neural latent specification to explicitly model the residual difference. An interesting observation is that the EC, EL and EP specifications using the SER model led to better speech quality but worse emotion similarity, compared with the EA specification. A possible reason is that the training corpus for SER is different from that of E-TTS. This change in corpora tends to lead to degradation in emotion recognition precision, resulting in inferior emotion similarity. Also, the EA specification is jointly optimized with the target E-TTS system towards the goal of sufficiently capturing the output acoustic variations. The effect of better variation capturing can be observed from the lower MSE error in Table IV. A case study is presented in the following section.

E. Pitch Contour Analysis

We conduct further case study on synthetic pitch contours to investigate the ability of the proposed E-TTS systems to generate speech with various acoustic realizations (pitch variations) specified by the utterance exemplars.



(a) "And high for the girl."

(b) "A few hours later, three weddings had taken place."

Fig. 14. (a) An exemplar with the text content of "And high for the girl". The pitch contour is high and has two peaks at around 0.7 s and 1.2 s. The recognized emotion code is *Neutral* and detailed probabilities of the four classes *Neutral*, *Angry*, *Happy* and *Sad* are (0.57, 0.06, 0.37, 0). (b) The speech synthesized by Tacotron, EC-TTS, EP-TTS and EA-TTS using the exemplar in (a). EA-TTS can mimic the emotion in the exemplar better than the SER-based systems EC-TTS and EP-TTS. The pitch contour generated by EA-TTS is higher and contains two peaks.

The first question is why the SER-based systems, i.e EC-TTS and EP-TTS is inferior to EA-TTS in emotion similarity, as shown in the above subjective evaluations. Fig. 14(a) presents an exemplar with the text content of "And high for the girl." The pitch contour is high with two peaks at around 0.7 and 1.2 s. Feeding this exemplar to the SER system obtains the emotion class of *Neutral*, and the predicted probablities are (0.57, 0.06, 0.37, 0) for the four classes of *Neutral*, *Angry*, *Happy* and *Sad*. Fig. 14(b) shows the utterances synthesized by Tacotron, EC-TTS, EP-TTS and EA-TTS using the exemplar in Fig. 14(a). As can be found that since the recognized emotion class is Neutral, EC-TTS systhesize an utterance that is similar to the average-emotion utterance generated by Tacotron, but not similar to the exemplar. The probability values provide EP-TTS more emotive information, and the synthesized pitch contour is more similar to the exemplar—the pitch contour is higher and contains a peak. The EA-TTS system generates a even more similar pitch contour that is higher than the other systems and contains two peaks. This shows the importance of feeding enough emotive information to the E-TTS systems.

We analyze the pitch contours of the speech synthesized by EA-TTS to examine the similarity between the pitch range of the synthesized utteances and that of the exemplary utterances. We use two utterances from the audio corpus that have different pitch ranges for emotion specification, as shown in Fig. 15 and 16. Exemplar *A* in Fig. 15 has a lower pitch contour, mostly

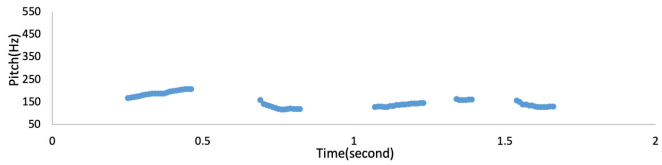


Fig. 15. Pitch contour of the exemplar *A* with textual content of "Soon, he was sound asleep."

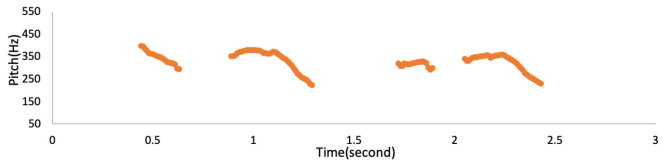


Fig. 16. Pitch contour of the exemplar *B* with textual content of "Oh Juliet, my Juliet!."

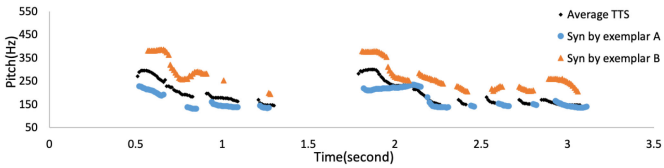


Fig. 17. Pitch contours of the two speech utterances synthesized by the exemplars *A* and *B*, respectively. The two utterances have the same textual content of "A few hours later, three weddings had taken place."

between 117 Hz and 209 Hz, while exemplar *B* in Fig. 16 has a higher pitch contour between 223 Hz and 398 Hz. We generate three utterances with the same textual content for comparison, as shown in Fig. 17. The EA-TTS system is used to generate two speech utterances with the two emotion specifications from the two utterance exemplars. A third speech utterance with average emotion is generated by the Tacotron system. As shown in Fig. 17, the utterance synthesized with exemplar *A* also has a lower pitch contour, between 132 Hz and 236 Hz, below the neutral contour. The utterance synthesized with exemplar *B* has a higher pitch contour, between 195 Hz and 387 Hz, above the neutral one. This demonstrates the ability of the EA-TTS to generate similar pitch contour as specified by the exemplar, in terms of pitch ranges.

V. Conclusion

In this paper, we present a novel approach that advocates the use of exemplar-based emotive speech synthesis. The approach aims to bypass the step of emotion specification using categorical codes or dimensional values which present difficulties not only in annotation, but also in enforcing annotation consistency in face of human subjectivity. The proposed approach circumvents these difficulties by advocating the use of an utterance exemplar that carries the target emotive information.

This paper has addressed four research questions related to the exemplar-based emotive speech synthesis approach, relating to feature representation of the utterance exemplar, emotion descriptor, mapping between the features and the emotion descriptors, and the use of the descriptors in emotive speech synthesis. We adopt conventional categorical codes as emotive descriptors of the utterance exemplar through the use of a speech emotion recognizer (SER). Hence the descriptor can take the form of a one-hot vector with a single recognized emotion (denoted as EC), a set of confidence values across all emotion categories (denoted as EP), or a set of logit values of all emotion categories before the softmax layers in the SER (denoted as EL). We are mindful that conventional categorical codes as emotive descriptors may fall short in describing complex emotions (with a mixture of categories) and the highly varied acoustic realizations. Hence, we propose to use a neural latent representation that can be *automatically* derived from the utterance exemplar as the emotion descriptor (denoted as EA). To map the feature representation (i.e. the spectrogram) into the emotion descriptors, we use two kinds of neural networks, namely capsule networks (CapNets) and residual error networks (RENets). CapNets can capture and preserve spatial information in the time-frequency analyses in the spectrogram. ReNets can capture contrastive information between neutral and emotive acoustic realizations based on the same linguistic (i.e. textual) input. Finally the descriptor values are used to augment textual input for emotive speech synthesis using a sequence-to-sequence architecture.

All four types of emotion descriptors (EC, EP, EL and EA), derived from the utterance exemplar, proved to be effective for emotive speech synthesis in generating outputs with superior speech quality and emotion similarity (with the target reference) compared with a baseline, average emotion TTS system based on Tacotron. Analysis of synthesized pitch contours also validates the capability of generating similar acoustic variations as specified by the utterance exemplars. The neural descriptor derived using RENets achieves better emotion similarity than those derived with CapNets, by leveraging joint training of the RENets and the seq2seq architecture for synthesis. Future work will investigate speaker independence in the use of utterance exemplars for this novel emotive speech synthesis framework.

References

[1] J. Hirschberg, "Communication and prosody: Functional aspects of prosody," *Speech Commun.*, vol. 36, no. 1-2, pp. 31–43, 2002.
[2] A. Vinciarelli Hammal *et al.*, "Open challenges in modelling, analysis and synthesis of human behaviour in human-human and human-machine interactions," *Cogn. Computat.*, vol. 7, no. 4, pp. 397–413, 2015.
[3] J. E. Cahn, "Generating expression in synthesized speech," Ph.D. dissertation, Dept. Architecture, Massachusetts Inst. Technol., 1989.
[4] M. Bulut, S. S. Narayanan, and A. K. Syrdal, "Expressive speech synthesis using a concatenative synthesizer," in *Proc. Int. Conf. Spoken Lang. Process.*, 2002.
[5] I. R. Murray and J. L. Arnott, "Implementation and testing of a system for producing emotion-by-rule in synthetic speech," *Speech Commun.*, vol. 16, pp. 369–390, 1995.
[6] M. Schröder, "Emotional speech synthesis: A review," in *Proc. Eur. Conf. Speech Commun. Technol.*, 2001, pp. 561–564.
[7] J. Yamagishi, K. Onishi, T. Masuko, and T. Kobayashi, "Modeling of various speaking styles and emotions for HMM-based speech synthesis," in *Proc. Eur. Conf. Speech Commun. Technol.*, 2003, pp. 2461–2464.
[8] P. Bell, T. Burrows, and P. Taylor, "Adaptation of prosodic phrasing models," in *Proc. Speech Prosody*, 2006.
[9] V. Strom, R. A. Clark, and S. King, "Expressive prosody for unit-selection speech synthesis," in *Proc. Int. Conf. Spoken Lang. Process.*, 2006, pp. 1296–1299.

[10] X. Wu *et al.*, "Feature based adaptation for speaking style synthesis," in *Proc. Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 5304–5308.

[11] F. Burkhardt, A. Steinhilber, and B. Weiss, "Ironic speech - Evaluating acoustic correlates by means of speech synthesis," *ESSV*, pp. 342–350, 2018.

[12] A. W. Black, "Unit selection and emotional speech," in *Proc. Eur. Conf. Speech Commun. Technol.*, 2003, pp. 1649–1652.

[13] Q. T. Do, S. Sakti, G. Neubig, T. Toda, and S. Nakamura, "Preserving word-level emphasis in speech-to-speech translation," in *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 3, pp. 544–556, 2017.

[14] K. Yu, F. Mairesse, and S. Young, "Word-level emphasis modelling in HMM-based speech synthesis," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Dallas, TX, USA, 2010, pp. 4238–4241.

[15] S. Andersson, J. Yamagishi, and R. A. Clark, "Synthesis and evaluation of conversational characteristics in HMM-based speech synthesis," *Speech Commun.*, vol. 54, no. 2, pp. 175–188, 2012.

[16] N. Campbell, "Approaches to conversational speech rhythm: Speech activity in two-person telephone dialogues," *Int. Congr. Phonetic Sci.*, pp. 343–348, 2007.

[17] K. Lenzo and A. Black, "Customized synthesis: Blending and tiering," *Avios*, 2002.

[18] Y. Lee, A. Rabiee, and S.-Y. Lee, "Emotional end-to-end neural speech synthesizer," *Adv. Neural Inf. Process. Syst.*, 2017.

[19] Z. Hodari, O. Watts, S. Ronanki, and S. King, "Learning interpretable control dimensions for speech synthesis by using external data," *Interspeech*, pp. 32–36, 2018.

[20] S. Sabour, N. Frosst, and G. Hinton, "Dynamic routing between capsules," *Adv. Neural Inf. Process. Syst.*, pp. 3856–3866, 2017.

[21] X. Wu *et al.*, "Rapid style adaptation using residual error embedding for expressive speech synthesis," in *Proc. Interspeech*, pp. 3072–3076, 2018.

[22] R. Barra-Chicote, J. Yamagishi, S. King, J. M. Montero, and J. Macias-Guarasa, "Analysis of statistical parametric and unit selection speech synthesis systems applied to emotional speech," *Speech Commun.*, vol. 52, no. 5, pp. 394–404, 2010.

[23] A. Mehrabian, "Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament," *Curr. Psychol.*, vol. 14, no. 4, pp. 261–292, 1996.

[24] M. Charfuelan and I. Steiner, "Expressive speech synthesis in mary tts using audiobook data and emotionml," in *Proc. Interspeech*, pp. 1564–1568, 2013.

[25] H. Gunes, B. Schuller, M. Pantic, and R. Cowie, "Emotion representation, analysis and synthesis in continuous space: A survey," *Face Gesture*, pp. 827–834, 2011.

[26] F. Eyben, S. Buchholz, and N. Braunschweiler, "Unsupervised clustering of emotion and voice styles for expressive TTS," in *Proc. Int. Conf. Acoust., Speech Signal Process.*, 2012, pp. 4009–4012.

[27] T. Kenter, V. Wan, C.-A. Chan, R. Clark, and J. Vit, "CHiVE: Varying prosody in speech synthesis with a linguistically driven dynamic hierarchical conditional variational network," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 3331–3340.

[28] E. Szèkely, J. P. Cabral, P. Cahill, and J. Carson-Berndsen, "Clustering expressive speech styles in audiobooks using glottal source parameters," in *Proc. Annu. Conf. ISCA*, 2011, pp. 2409–2412.

[29] L. Wang, Y. Zhao, M. Chu, Y. Chen, F. Soong, and Z. Cao, "Exploring expressive speech space in an audio-book," *Speech Prosody*, 2006.

[30] M. Wöllmer, F. Eyben, B. Schuller, E. Douglas-Cowie, and R. Cowie, "Data-driven clustering in emotional space for affect recognition using discriminatively trained LSTM networks," in *Proc. Interspeech*, 2009.

[31] X. Wu, Z. Wu, J. Jia, H. Meng, L. Cai, and W. Li, "Automatic speech data clustering with human perception based weighted distance," in *Proc. Int. Symp. Chinese Spoken Lang. Process.*, 2014, pp. 216–220.

[32] G. E. Henter, J. Lorenzo-Trueba, X. Wang, and J. Yamagishi, "Deep encoder-decoder models for unsupervised learning of controllable speech synthesis," 2018, *arXiv:1807.11470*.

[33] R. Habib *et al.*, "Semi-supervised generative modeling for controllable speech synthesis," in *Proc. Int. Conf. Learn. Representations*, 2020.

[34] I. Jauk and A. Bonafonte Càvez, "Prosodic and spectral Ivectors for expressive speech synthesis," *ISCA Workshop Speech Synth.*, pp. 59–63, 2016.

[35] J. M. Montero, J. Gutièrrez-Arriola, J. Colàs, E. Enriquez, and J. M. Pardo, "Analysis and modelling of emotional speech in Spanish," in *Proc. Int. Congr. Phonetic Sci.*, vol. 2, 1999, pp. 957–960.

[36] M. Schröder, "Can emotions be synthesized without controlling voice quality," *Phonus*, vol. 4, pp. 35–50, 1999.

[37] J. Llisterri, "Speaking styles in speech research," *Workshop Integrating Speech Natural Lang.*, pp. 1–28, 1992.

[38] I. Iriondo *et al.*, "Validation of an acoustical modelling of emotional expression in spanish using speech synthesis techniques," *ISCA Tut. Res. Workshop Speech Emotion*, pp. 161–166, 2000.

[39] F. Meng, H. Meng, Z. Wu, and L. Cai, "Synthesizing expressive speech to convey focus using a perturbation model for computer-aided pronunciation training," *Second Lang. Studies: Acquisition, Learn., Educ. Technol.*, 2010.

[40] O. Watts, Z. Wu, and S. King, "Sentence-level control vectors for deep neural network speech synthesis," in *Proc. Annu. Conf. ISCA*, 2015, pp. 2217–2221.

[41] G. E. Henter, J. Lorenzo-Trueba, X. Wang, and J. Yamagishi, "Principles for learning controllable tts from annotated and latent variation," in *Proc. Interspeech*, 2017, pp. 3956–3960.

[42] Y. Wang *et al.*, "Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 5180–5189.

[43] D. Stanton, Y. Wang, and R. Skerry-Ryan, "Predicting expressive speaking style from text in end-to-end speech synthesis," *IEEE Spoken Lang. Technol. Workshop*, 2018, pp. 595–602.

[44] Y. Wang *et al.*, "Uncovering latent style factors for expressive speech synthesis," *ML4Audio Workshop*, NIPS, 2017.

[45] W.-N. Hsu, Y. Zhang, and J. Glass, "Unsupervised learning of disentangled and interpretable representations from sequential data," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 1878–1889.

[46] N. Tits, F. Wang, K. E. Haddad, V. Pagel, and T. Dutoit, "Visualization and interpretation of latent spaces for controlling expressive speech synthesis through audio analysis," in *Proc. Interspeech*, 2019, pp. 4475–4479.

[47] W.-N. Hsu *et al.*, "Hierarchical generative modeling for controllable speech synthesis," in *Proc. Int. Conf. Learn. Representations*, 2019.

[48] E. Battenberg *et al.*, "Effective use of variational embedding capacity in expressive end-to-end speech synthesis," 2019, *arXiv:1906.03402*.

[49] G. Sun, Y. Zhang, R. J. Weiss, Y. Cao, H. Zen, and Y. Wu, "Fully-hierarchical fine-grained prosody modeling for interpretable speech synthesis," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 6264–6268.

[50] Z.-H. Ling *et al.*, "Deep learning for acoustic modeling in parametric speech generation: A systematic review of existing techniques and future trends," *IEEE Signal Process. Mag.*, vol. 32, no. 3, pp. 35–52, May 2015.

[51] J. Lorenzo-Trueba, G. E. Henter, S. Takaki, J. Yamagishi, Y. Morino, and Y. Ochiai, "Investigating different representations for modeling and controlling multiple emotions in DNN-based speech synthesis," *Speech Commun.*, pp. 135–143, 2018.

[52] S. O. Arik *et al.*, "Deep voice 2: Multi-speaker neural text-to-speech," *Adv. Neural Inf. Process. Syst.*, 2017.

[53] J. Shen *et al.*, "Natural TTS synthesis by conditioning wavenet on mel spectrogram predictions," *2018 IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2018, pp. 4779–4783.

[54] Y. Wang *et al.*, "Tacotron: Towards end-to-end speech synthesis," *Interspeech*, 2017, pp. 4006–4010.

[55] X. Wu, S. Kang, L. Sun, Y. Ning, Z. Wu, and H. Meng, "Attention-Based recurrent generator with gaussian tolerance for statistical parametric speech synthesis," *Affect. Social Multimedia Comput.*, 2017.

[56] R. Valle, K. Shih, R. Prenger, and B. Catanzaro, "Flowtron: An autoregressive flow-based generative network for text-to-speech synthesis," in *Proc. Int. Conf. Learn. Representations*, 2020.

[57] V. Klimkov, S. Ronanki, J. Rohnke, and T. Drugman, "Fine-Grained robust prosody transfer for single-speaker neural text-to-speech," in *Proc. Interspeech*, 2019, pp. 4440–4444.

[58] R. Valle, J. Li, R. Prenger, and B. Catanzaro, "Mellotron: Multispeaker expressive voice synthesis by conditioning on rhythm, pitch and global style tokens," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 6189–6193.

[59] T. Raitio, R. Rasipuram, and D. Castellani, "Controllable neural text-to-speech synthesis using intuitive prosodic features," *Proc. Interspeech*, 2020, pp. 4432–4436.

[60] S. R. Livingstone and F. A. Russo, "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*, vol. 13, no. 5, 2018, Art. no. e0196391.

[61] G. E. Hinton, S. Sabour, and N. Frosst, "Matrix capsules with EM routing," in *Proc. Int. Conf. Learn. Representations*, 2018.

[62] X. Wu *et al.*, "Speech emotion recognition using capsule networks," *Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 6695–6699.

[63] A. M. Lamb, A. Goyal, Y. Zhang, S. Zhang, A. C. Courville, and Y. Bengio, "Professor forcing: A new algorithm for training recurrent networks," *Adv. Neural Inf. Process. Syst.*, 2016, pp. 4601–4609.
[64] H. Guo, F. K. Soong, L. He, and L. Xie, "A new GAN-based end-to-end TTS training algorithm," in *Proc. Interspeech*, 2019, pp. 1288–1292.
[65] L. Qin, Z.-H. Ling, Y.-J. Wu, B.-F. Zhang, and R.-H. Wang, "HMM-Based emotional speech synthesis using average emotion model," in *Proc. Int. Symp. Chinese Spoken Lang. Process.*, 2006, pp. 233–240.
[66] C. Busso *et al.*, "IEMOCAP: Interactive emotional dyadic motion capture database," *Lang. Resour. Eval.*, vol. 42, no. 4, pp. 335–359, 2008.
[67] A. Satt, S. Rozenberg, and R. Hoory, "Efficient emotion recognition from speech using deep learning on spectrograms," *Interspeech*, pp. 1089–1093, 2017.
[68] S. King and V. Karaiskos, "The blizzard challenge 2016," in *Proc. Blizzard Challenge Workshop*, 2016.
[69] R. K. Srivastava, K. Greff, and J. Schmidhuber, "Highway networks," *Deep Learn. Workshop, ICML*, 2015.
[70] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
[71] A. van den Oord *et al.*, "Parallel WaveNet: Fast high-fidelity speech synthesis," in *Proc. 35th Int. Conf. Mach. Learn.*, Jul. 2018, pp. 3918–3926.
[72] M. Wang *et al.*, "Speech super-resolution using parallel wavenet," in *Proc. 11th Int. Symp. Chin. Spoken Lang. Process.*, 2018, pp. 260–264.
[73] R. Skerry-Ryan *et al.*, "Towards end-to-end prosody transfer for expressive speech synthesis with tacotron," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 4693–4072.
[74] J. Lee and I. Tashev, "High-level feature representation using recurrent neural network for speech emotion recognition," in *Proc. Interspeech*, 2015, pp. 1537–1540.
[75] P. Li, Y. Song, I. McLoughlin, W. Guo, and L. Dai, "An attention pooling based representation learning method for speech emotion recognition," in *Proc. Interspeech*, 2018, pp. 3087–3091.
[76] Y. Lee and T. Kim, "Robust and fine-grained prosody control of end-to-end speech synthesis," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2019, pp. 5911–5915.

Xixin Wu (Member, IEEE) received the B.S. degree from Beihang University, Beijing, China, the M.S. degree from Tsinghua University, Beijing, China, and the Ph.D. degree from The Chinese University of Hong Kong, Hong Kong. He is currently a Research Associate with Machine Intelligence Laboratory, Cambridge University, Cambridge, U.K. His research interests include speech synthesis and recognition, speaker verification, and neural network uncertainty. He is a Member of ISCA.

Yuewen Cao received the B.S. degree in communication engineering from the Huazhong University of Science and Technology, Wuhan, China, in 2017. She is currently working toward the Ph.D. degree with Human-Computer Communications Lab, The Chinese University of Hong Kong, Hong Kong. Her research interests include speech synthesis and voice conversion.

Hui Lu received the B.S. degree in communication engineering from Tongji University, Shanghai, China, in 2017 and the M.S. degree in computer technology from Tsinghua University, Beijing, China, in 2020. He is currently working toward the Ph.D. degree with the Human-Computer Communications Lab, The Chinese University of Hong Kong, Hong Kong. His research interests include speech synthesis and voice conversion.

Songxiang Liu received the B.S. degree in automation from Zhejiang University, Hangzhou, China, in 2016. He is currently working toward the Ph.D. degree with Human-Computer Communications Lab (HCCL), The Chinese University of Hong Kong, Hong Kong. His research interests include voice conversion, accent conversion, audio adversarial attack, and speech synthesis.

Shiyin Kang received the B.S. degree in automation and the M.S. degree in computer science and technology from Tsinghua University, Beijing, China, and the Ph.D. degree in systems engineering and engineering management from The Chinese University of Hong Kong, Hong Kong. From 2016 to 2020, he joined Tencent AI Lab as a Senior Researcher. He is currently a Tech. Lead with Huya Inc. His research interests include speech technology, multimodal speech synthesis, singing synthesis, voice conversion, and applications in machine learning.

Zhiyong Wu (Member, IEEE) received the B.S. and the Ph.D. degrees in computer science and technology from Tsinghua University, Beijing, China, in 1999 and 2005, respectively. From 2005 to 2007, he was a Postdoctoral Fellow with the Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong (CUHK), Hong Kong. He then joined the Graduate School at Shenzhen (now Shenzhen International Graduate School), Tsinghua University, Shenzhen, China, and is currently an Associate Professor. He is also a Coordinator with Tsinghua-CUHK Joint Research Center for Media Sciences, Technologies and Systems. His research interests include intelligent speech interaction, more specially, speech processing, audiovisual bimodal modeling, text-to-audio-visual-speech synthesis, and natural language understanding and generation. He is a Member of International Speech Communication Association and China Computer Federation.

Xunying Liu (Member, IEEE) received the Ph.D. degree in speech recognition and the M.Phil. degree in computer speech and language processing from the University of Cambridge, Cambridge, U.K., prior to his undergraduate study with Shanghai Jiao Tong University, Shanghai, China. He was a Senior Research Associate with Machine Intelligence Laboratory, Cambridge University Engineering Department, University of Cambridge, and since 2016, he has been an Associate Professor with the Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong, Hong Kong. His current research interests include large vocabulary continuous speech recognition, statistical language modelling, audio-visual speech processing, machine learning, language learning, speech synthesis and assistive technology. He and his students were the recipients of a number of best paper awards and nominations, including the Best Paper Award at ISCA Interspeech2010 for the paper titled Language Model Cross Adaptation for LVCSR System Combination and the Best Paper Award at IEEE ICASSP2019 for their paper titled BLHUC: Bayesian Learning of Hidden Unit Contributions for Deep Neural Network Speaker Adaptation. He is a Member of ISCA.

Helen Meng (Fellow, IEEE) received the B.S., M.S., and Ph.D. degrees in electrical engineering from the Massachusetts Institute of Technology, Cambridge, MA, USA. In 1998, she joined the Chinese University of Hong Kong, Hong Kong, where she is currently the Chair Professor with the Department of Systems Engineering & Engineering Management. She was the former Department Chairman and the Associate Dean of Research with the faculty of Engineering. Her research interests include human–computer interaction via multimodal and multilingual spoken language systems, spoken dialog systems, computer-aided pronunciation training, speech processing in assistive technologies, health-related applications, and big data decision analytics. She was the Editor-in-Chief of the IEEE TRANSACTIONS ON AUDIO, SPEECH AND LANGUAGE PROCESSING between 2009 and 2011. She was the recipient of the IEEE Signal Processing Society Leo L. Beranek Meritorious Service Award in 2019. She was also on the Elected Board Member of the International Speech Communication Association (ISCA) and an International Advisory Board Member. She is a ISCA, HKCS, and HKIE.