SAMSUNG
Best Paper Award 2021

Multi-Speaker Emotional Speech Synthesis with Fine-Grained Prosody Modelling

Chunhui Lu[1], Xue Wen[1], Ruolan Liu[1] and June Sig Sung[2]

(Abstract) As the naturalness of artificially generated speech comes close to real speech, expressiveness becomes the next major research task in speech synthesis. Since labelled expressive speech is expensive and time-consuming to collect in large quantity, data efficiency becomes important when adapting end-to-end techniques to expressive speech. In this paper we present an end-to-end system for multi-speaker emotional speech synthesis. In particular, our system learns emotion classes (angry, happy, etc.) from just two speakers, then generalizes these classes to other speakers from whom no emotional data was seen. We address this highly unbalanced learning task by integrating fine-grained, latent prosodic features with sentence-level emotion embedding. These fine-grained features learn to represent local prosodic variations disentangled from speaker identity, tone and global emotion label, which helps transfer learned emotions between speakers. Compared to systems that model emotions at sentence level only, our method achieves higher ratings on naturalness by ~0.4, and on expressiveness by ~0.6 for angry and happy classes, while retaining comparable speaker similarity. Contributions reported here have been published by ICASSP 2021 [1]. This technique is expected to bring improved expressiveness and controllability to our line of voice-enabled products and services.

1. INTRODUCTION

Emotional speech synthesis (ESS) aims to generate natural expressive speech with prescribed emotion, usually chosen from several predefined classes (happy, angry, etc.). Just like humans modulating voice emotion for effective communication, ESS can potentially improve human-computer voice interface. This is desirable in voice-enabled applications like spoken dialogue systems, voice translation and content creation using text-to-speech (TTS) synthesis.

While recent state-of-the-art speech synthesizers, e.g. [2], [3], take advantage of large multi-speaker datasets, obtaining emotional speech on comparable scale is challenging. One particular difficulty is that consistently producing speech with prescribed emotion needs professional training, and most of our "corpus speakers" have not been trained do it. This also exacerbates the high cost issue in building TTS database. In this paper we consider a more practical setup in which we have a few professional "emotional speakers" who provide emotional speech examples, and many "neutral speakers" from whom we have only emotionally neutral examples. More concretely, we use two emotional speakers (1 male and 1 female), and 8 neutral speakers (4 male and 4 female). We define four emotion classes: angry, happy, sad, and neutral. Our goal is to produce emotional speech in the voices of all 10 speakers.

This data setup carries a bias that associates emotional speech with a few specific speakers, hence poses a risk of the system learning that biased association. This has been observed in our preliminary study using global sentence-level emotion representation for ESS, in which the straight combination of neutral speaker with non-neutral emotion class led to degraded emotion rendering and overall voice quality. In response, we seek a mechanism to dissociate emotion classes from these few emotional speakers.

In this paper, we augment sentence-level emotion tag with fine-grained prosody representation. The latter captures local speech variations not fully characterized by text, speaker ID and global emotion label. Concretely, we associate each spoken phoneme with a 3-dimensional latent code learned within a conditional variational autoencoder (CVAE) framework. This very tight 3-wide information bottleneck encourages learning acoustically important features not already present among the conditioning inputs, therefore detaches the latent code from speaker identity. By transferring this latent code from emotional speakers, we are able to produce emotional speech of neutral speakers. Experimental results show our method achieves consistent improvement in both naturalness and expressiveness. The latent codes themselves are shown to learn meaningful features that are potentially useful to controllable synthesis.

Our contributions in this work are as follows:

(1) We introduce fine-grained latent prosody modelling to multi-speaker ESS under low emotion-resourced setting, based on the information bottleneck principle.
(2) We show that the latent variables can learn disentangled and meaningful acoustic features in unsupervised manner, which provides emotion transfer capability and speech controllability.

We organize the rest of paper as follows. Section 2 briefly reviews recent related work. Section 3 presents our methods in detail. Section 4 describes our evaluation setup and reports results. Section 5 concludes the paper.

[1] Speech Lab, SRC-Beijing, No.12 Taiyanggong Middle Road, Chaoyang District, Beijing, 100028, China
[2] AI R&D Group, Samsung Electronics, Maetan 3-Dong, Yeongtong-Gu, Suwon, Gyeonggi-Do, Korea

2. RELATED WORK

2.1. End-to-end speech synthesis

Neural network based TTS systems have made substantial progress and attained state-of-the-art results in the last five years. Notable examples include Tacotron [4], Tacotron 2 [5], DeepVoice 3 [3], DeepConvTTS [6], and Transformer TTS [7], to name a few. These end-to-end (E2E) synthesizers leverage the encode-attend-decode framework to generate acoustical frames autoregressively (AR) from text. Long-range attention was first designed for machine translation with word reordering in mind [8], and is known to be related to word skipping and repeating issues in TTS when used as is. Autoregressive prediction is hard to parallelize hence incurs a speed penalty on modern neural hardware.

2.1.1. FastSpeech and FastSpeech 2

Non-autoregressive TTS models [9-12] have been designed to address the issues above, among which FastSpeech [9] is a good representative. FastSpeech does not use long-range attention for text-speech alignment, but opts to make phoneme durations (and therefore text-speech alignment) explicit. It uses feed-forward Transformer networks on both phoneme and frame levels. These networks feature a residual backbone which prioritizes 1-to-1 monotonic alignment. A length regulator deterministically expands phoneme-level features to frame level using explicit durations. At synthesis time these durations are predicted for all phonemes. FastSpeech uses no auto-regression, but entrusts temporal correlation modelling to deep self-attention instead.

FastSpeech 2 [12] extends FastSpeech mainly by exposing frame-level pitch and energy as explicit control variables of the system. These are supervised under ancillary objectives to make the synthesizer produce pitch and energy contours similar to training examples. Prosody control is possible by perturbing the contours, but intuitive control via frame-level parameters is a nontrivial problem by itself.

Our work in this paper also extends a FastSpeech-like synthesizer with fine-grained features. Unlike FastSpeech 2, our features are unsupervised latent variables, and they apply at phoneme level.

2.2. Prosody modelling

2.2.1. Sentence-level prosody modelling

Speaking style, or expression, refers to the different ways one can speak the same text in, and is key to expressive TTS. As there is no intuitive way to quantize speaking style at high level, several works explored unsupervised latent style representations [13-17] with the variational autoencoder (VAE) framework. [13] introduced fix-length sentence-level latent code to condition the decoder alongside input text. At synthesis time the code drawn from one reference sentence can be used to transfer its style onto a newly synthesized sentence. In [14] the latent code is constrained to be the linear combination of a several vectors called global style tokens (GSTs), each of which conceptually represents a distinct speaking style. This construct imposes a tightened information bottleneck on the style embedding, and empirically this system was found to disentangle style and text better than [13]. [15-17] work along the same lines, differing mostly by design choices such as observed variables, probability specifications, and annealing of the Kullback-Leibler term in VAE objectives. Some latent dimensions have been reported to associate with intuitive surface features like speed, pitch and noise level, although no direct supervision was used to learn them.

Sentence-level prosody modelling can have difficulties when scaling to long sentences, or when fine-grained control at sub-sentence level is needed.

2.2.2. Fine-grained prosody modelling

Fine-grained, or “variable length”, prosody embeddings are computed at finer granularities, e.g. per phoneme or per frame. [18] directly encoded pitch, energy and duration statistics of forced-aligned phoneme states as control variables to reproduce reference prosody in a synthesized sentence. [19] took a subtler approach to learn phoneme- and frame-level prosody embeddings as hidden features. [20] and [21] treated these fine-grained embeddings as latent variables to be learned with VAEs under different model specifications. Notably, both [19] and [20] learned low dimension embeddings that turned out to explicitly associate with meaningful local acoustic attributes. This is consistent with the general intuition that an information bottleneck prioritizes passing important information that is not available elsewhere, see also [22, 23].

2.3. Emotional speech synthesis

2.3.1. Single speaker ESS

Most ESS systems so far use emotional speech from one single speaker. [24, 25] implemented emotional statistical parametric speech synthesis (SPSS) with emotion codes. [26] introduced an E2E emotional speech synthesizer based on Tacotron by injecting a learned emotion embedding. [27-29] adopted pretrained GSTs to encode different emotions.

Some systems also considered synthesizing emotion at different strength levels [29-32]. [29] controlled emotion strength by directly scaling the emotion embedding, while [30] used population-based interpolation between emotional and neutral embeddings. [31]and [32] calibrated utterance- and phoneme-level emotion strengths against a ranking function trained separately to tell between happy and neutral classes, then exposed these strengths variables as explicit controls.

2.3.2. Multi speaker ESS

For multi-speaker ESS, [33] investigated different combinations of speaker and emotion representations with a convolutional neural network (CNN) synthesizer. They used 10 speakers, each with examples in all emotion classes. For generalizing emotional expressions to new speakers, early

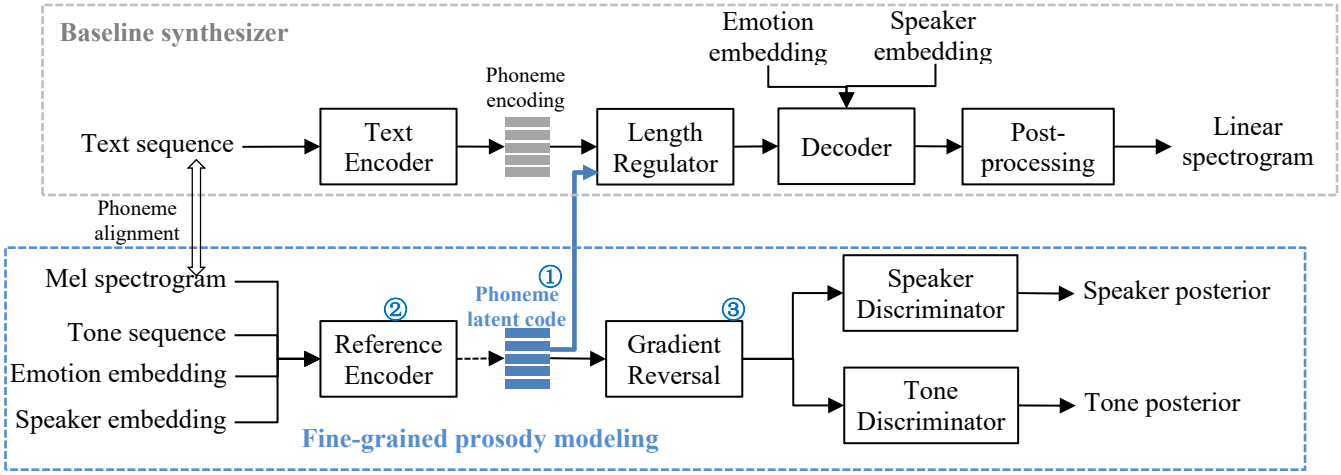


Fig. 1. Proposed architecture (training). Dashed lines denote sampling via reparameterization [37]

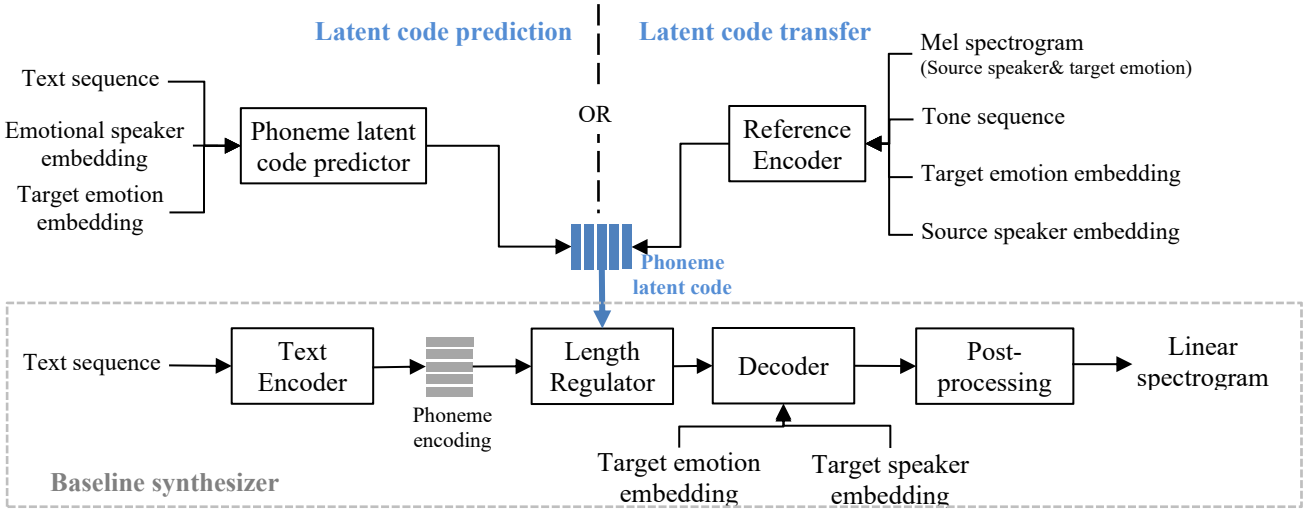


Fig. 2. Proposed architecture (inference). Phoneme latent code can be predicted from input text (top left part) or extracted from reference mel spectrogram (top right part)

SPSS researches [34, 35] tried representing emotions as additive factors. [36] investigated several deep neural network (DNN) architectures for emotion transplantation. In this work they used 3 emotional and 21 neutral speakers.

In this work, we do multi-speaker end-to-end emotional speech synthesis using a fine-grained, latent prosody model with tight information bottleneck, with special focus on knowledge transfer in low emotion-resourced setting.

3. METHODS

3.1. Baseline

Our baseline multi-speaker ESS is adapted from FastSpeech [9], as shown in the top part of Fig. 1. Its core part is an encoder-decoder DNN that converts a sequence of phonemes to a sequence of mel spectral frames, using a length regulator to match their lengths by repeating encoder outputs. Encoder sees only the text (as a phoneme sequence with tones); decoder sees emotion class, speaker ID and length-regulated encoder output. We investigate two approaches [26, 28] to represent different emotion class, either as a free, learnable vector which we denote as BASE-EMB, or as weighted combination of GSTs which we denote as BASE-GST.

We follow [9] to predict phoneme durations in log domain, now conditioned on speaker and emotion label too. Reference durations are obtained by forced alignment. Finally we use a post-net [4] to convert mel spectrogram to linear spectrogram, and use the Griffin-Lim [38] algorithm to construct audio output.

3.2. Fine-grained prosody modelling

We introduce fine-grained latent variables to BASE-EMB to learn disentangled local prosody variations in emotional speech. As shown in Fig. 1, the latent code① joins the main TTS at encoder output and conditions the decoder alongside global speaker and emotion labels. The latent code is learned with a conditional VAE framework. A reference encoder② computes a variational posterior of the latents from phoneme-aligned spectrogram, phoneme tone and global

speaker and emotion IDs. From this we draw a latent code and append it to the phoneme encoding before sending to the length regulator. We apply domain adversarial training③ [39] to further disentangle the latent code from speaker and tone. This is implemented with two discriminators and a gradient reversal unit [40].

The objective function is formulated as the combination of an evidence lower bound (ELBO) of expected reconstruction loss, and a domain adversarial loss term:

$$\mathcal{L} = \mathcal{L}_{ELBO} - \lambda_{adv}\mathcal{L}_{adv} \quad (1)$$

where $\mathcal{L}_{adv}$ is the sum of minimum cross-entropy losses for the two discriminators. $\mathcal{L}_{ELBO}$ is a β-VAE objective [41] under standard Gaussian latent prior:

$$\mathcal{L}_{ELBO} = \mathbb{E}_{q(\mathbf{z}|\mathbf{x},\mathbf{y},\mathbf{s},\mathbf{e})}[\log p(\mathbf{x}|\mathbf{y},\mathbf{z},\mathbf{s},\mathbf{e})] - \lambda_{KL}\sum_{u=1}^{U} D_{KL}(q(\mathbf{z}_u|\mathbf{x},\mathbf{y},\mathbf{s},\mathbf{e})||\mathcal{N}(0,I)) \quad (2)$$

where $\mathbf{z}$ represents the sequence of latent codes and $\mathbf{z}_u$ is the latent code for the u-th phoneme. U is the number of phonemes. $\mathbf{s}$ and $\mathbf{e}$ are speaker and emotion embeddings, respectively. We will use $0 < \lambda_{KL} < 1$, which favours accuracy over latent space exploration.

3.3. Latent code prediction

Once this enhanced TTS model has been trained, we can use the reference encoder to collect mean latent codes of each emotion class. We then train a separate model to predict them from text, speaker and emotion under the mean square error (MSE) loss. For emotional synthesis in the voice of a neutral speaker, we pick one emotional speaker to compute the latents, then use the target speaker for decoding.

Alternatively, if we are tasked with rendering a given emotional utterance in a neutral target speaker's voice, the latent code may also be computed using the reference encoder directly, as shown in the top right part of Fig. 2. A forced alignment module is needed for this type of emotion transfer.

4. EXPERIMENTS

4.1. Setup

4.1.1. Data

We conduct experiments on internal Mandarin Chinese datasets, detailed in Table 1. The two emotional speakers are professional voice actors; other (neutral) speakers are non-professionals. All utterances come with transcriptions in the form of phoneme-tone sequences. Recordings are sampled at 24kHz. 80-dim mel spectrograms and 1025-dim linear spectrograms are computed from 50ms windows at 12.5ms intervals.

Table 1. Training data details

Speaker Type	#speakers		#utterance per speaker			
	F	M	Angry	Happy	Neutral	Sad
Emotional	1	1	500	500	500	500
Neutral	4	4	-	-	1000	-
Total	10		1000	1000	9000	1000

4.1.2. Details

Details of the FastSpeech network is same as [9]. Both speaker and emotion embeddings are set to 64-dim in all systems. We follow [28] for emotion embedding in BASE-GST. The reference encoder closely follows that in [13] and extracts a 3-dim latent code for each phoneme. Both tone and speaker discriminators use a 2-layer feed forward network with hidden layer size 256. We set λ_{KL} and λ_{adv} to 0.01 and 0.02 respectively. Training the TTS engine took about 150k steps at batch size 32 on one P40 GPU.

Systems compared in our experiments are as follows.

- GT: ground-truth speech reconstructed by Griffin-Lim.
- BASE-EMB: baseline method using learnable emotion embedding as emotion representation.
- BASE-GST: baseline method using weighted combination of GSTs to represent different emotions. The reference audio samples of each emotion are selected from the visualization cluster centre and used when synthesizing target emotion.
- OURS-TRANS: proposed method with latent codes transferred from emotional speakers' emotional test speech.
- OURS-SYN: proposed method with latent codes predicted from text.

4.2. Subjective evaluation

4.2.1. Mean opinion scores

We first investigate mean opinion scores (MOS) in naturalness, speaker similarity and clarity of emotional expressiveness. 15 native Mandarin speakers are asked to listen to the synthesized speech and rate them on a scale between 1 (Bad) and 5 (Excellent). For speaker similarity, the target speaker's neutral speech is used as reference. The results are given in Tables 2 and 3. Results for neutral speakers (N) and emotional speakers (E) are calculated separately.

Naturalness: For angry and happy classes we observe little difference among systems for emotional speakers, but significant improvements with our fine-grained approach for neutral speakers. This shows that explicitly modelling phoneme-level variations can help generalize emotions to neutral speakers. Difference between OURS-SYN and OURS-TRANS is small, showing that the latent means are indeed predictable. For the neutral class, emotional speakers is rated 0.3 higher than neutral speakers by ground truth, indicating professional speakers sound more natural than nonprofessionals to the evaluators. The same gap is also observed in all systems tested, and we do not find notable difference between baseline and proposed systems. For the

Table 2. Comparison of MOS with 95% confidence intervals on naturalness and speaker similarity

Type	Method	Naturalness				Speaker similarity			
		Angry	Happy	Neutral	Sad	Angry	Happy	Neutral	Sad
E	GT	4.86±0.15	4.73±0.17	4.61±0.18	4.38±0.25	3.51±0.36	3.63±0.30	4.68±0.17	3.58±0.20
	BASE-EMB	4.41±0.31	4.34±0.28	4.30±0.26	3.63±0.32	3.57±0.34	3.61±0.31	4.43±0.25	3.31±0.16
	BASE-GST	4.37±0.26	4.29±0.19	4.30±0.23	3.64±0.28	3.57±0.20	3.58±0.19	4.41±0.21	3.31±0.21
	OURS-SYN	4.43±0.29	4.43±0.29	4.26±0.30	**3.88±0.23**	3.53±0.27	3.65±0.31	4.43±0.16	**3.41±0.26**
N	GT	-	-	4.29±0.23	-	-	-	4.70±0.15	-
	BASE-EMB	3.80±0.36	3.87±0.34	4.02±0.25	3.48±0.25	3.92±0.30	3.96±0.24	4.42±0.17	3.39±0.36
	BASE-GST	3.78±0.23	3.83±0.23	4.00±0.28	3.90±0.24	3.90±0.29	4.21±0.21	4.48±0.19	3.93±0.28
	OURS-TRANS	4.43±0.36	4.41±0.29	-	3.74±0.26	3.93±0.28	4.01±0.28	-	3.33±0.34
	OURS-SYN	**4.42±0.33**	**4.20±0.21**	3.94±0.28	3.65±0.23	4.02±0.27	4.03±0.26	4.49±0.17	3.43±0.33

Table 3. Comparison of MOS with 95% confidence intervals on emotional expressiveness

Type	Method	Angry	Happy	Sad
E	GT	4.92±0.09	4.89±0.07	4.75±0.09
	BASE-EMB	4.54±0.23	4.48±0.26	4.02±0.27
	BASE-GST	4.53±0.23	4.43±0.18	4.03±0.26
	OURS-SYN	4.52±0.28	4.55±0.24	4.04±0.33
N	BASE-EMB	3.72±0.27	3.75±0.36	3.96±0.29
	BASE-GST	3.70±0.23	3.53±0.20	3.57±0.28
	OURS-TRANS	4.60±0.19	4.55±0.21	4.06±0.22
	OURS-SYN	**4.47±0.28**	**4.31±0.28**	3.96±0.25

sad class, we observe low ratings across all systems (except BASE-GST).We conjecture that slow speaking rate with low F0 can give an unnatural feeling. Comparison between systems remain inconclusive.

Speaker similarity: Speaker similarity is rated higher for the neutral class than other emotions across all systems and speakers, but the gap is particularly large for emotional speakers. This may be related to the wide prosody variations in emotional speech, and we do notice professional speakers modulate their voice traits to deliver diverse emotions. Within each speaker type (E or N), similarity ratings are similar across systems (again except BASE-GST+sad). Scores of the sad class are again notably lower. We observe that F0 differences between sad and neutral emotions are bigger than other emotions, which may partially explain the observation.

Emotional expressiveness: For angry and happy classes of neutral speakers, we observe significant improvement in emotional expressiveness of our models over the baselines, with ratings on par with those for emotional speakers. This shows fine-grained modelling adds to the model's capability to deliver these emotions. For the sad class the ratings are similar across systems, maybe because global traits like long duration and low F0 are already sufficient to express sadness so local variations add little to them.

4.2.2. AB preference tests

We run A-B preference test to probe the effect of emotional phoneme duration models. In each pair of synthesized examples, one has duration input predicted with the target emotion, the other predicted with neutral emotion. Evaluators are asked to give their preference in terms of the naturalness. The results are given in Fig. 3. It shows that emotional duration modelling is more effective for sad examples, which are much slower than neutral and other emotions. In the angry and happy classes the effect is still significant but not as dominating, for other prosodic traits have become more relevant in expressing these emotions.

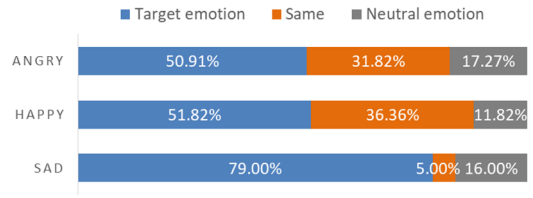
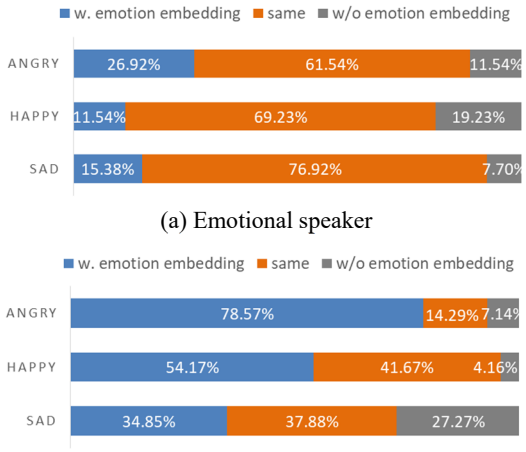


Fig. 3. AB preference results on naturalness of duration input predicted with different emotions

We run another preference test to check the effect of having sentence-level emotion embedding in the presence of phoneme-level embeddings. In this test we compare our system against one trained without sentence-level emotion embedding (but still using sentence-level emotion input to compute phoneme-level embeddings). Evaluators are asked to choose the one with higher emotion clarity. The results are given in Fig. 4. We find that the sentence-level emotion embedding is essential for neutral speakers but not for emotional speakers. This indicates remnant entanglement between speaker and fine-grained latents, and may be related to the model not seeing emotional fine-grained latents for neutral speakers during training.



(a) Emotional speaker

(b) Neutral speaker

Fig. 4. AB preference result on the clarity of emotional expressiveness of systems with and without global emotion embedding

4.3. Objective evaluation

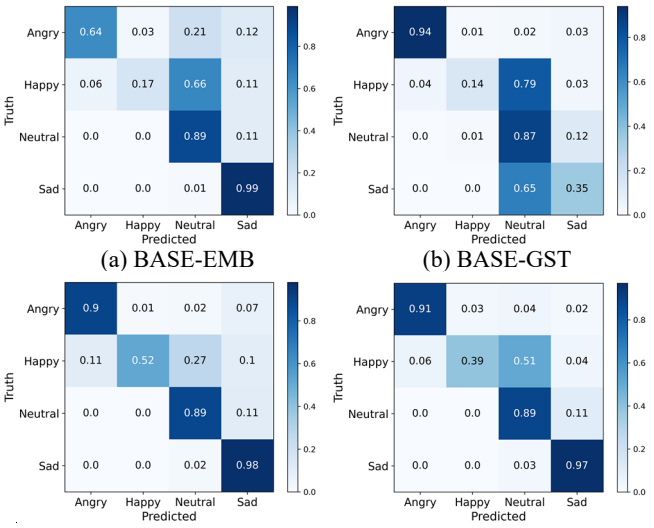
We train an emotion classifier to assess emotion clarity of synthesized mel spectrograms. This has similar structure to the reference encoder, except that the depth of the convolutional network is reduced to 3 layers. It was trained using 4000 speech examples (1000 per emotion) from the TTS training data. For each speaker, we generate 80 test sentences (20 per emotion) using each system, then classify them using the trained classifier. Results are listed in Table 4, calculated separately for emotional and neutral speakers.

Table 4. Accuracy of emotion classification

Method	Emotional	Neutral
BASE-EMB	92.50%	67.25%
BASE-GST	91.88%	57.50%
OURS-TRANS	96.88%	82.25%
OURS-SYN	94.38%	79.00%

For both emotional and neutral speakers OURS-TRANS achieves best classification, ahead of OURS-SYN in 2nd place by 2~3% absolute, which quantifies the gap between predicted and extracted phoneme-level latent codes. OURS-SYN shows marginal emotion clarity improvement on emotional speakers but a big one on neutral speakers, showing the effectiveness of our fine-grained approach in generalizing learned emotions to non-emotional speakers.

Fig. 5 shows confusion matrices computed from neutral speakers. These results are generally consistent with the MOS results on emotional expressiveness, but we do observe some variances. For example, BASE-GST gets highest classification accuracy on angry emotion but not highest MOS in Table 3. This again shows a discrepancy between a mechanic, sentence-level emotion model (in this case the classifier) and subjective judgement. The happy class has the highest error rates (false negatives) across all systems. The neutral class attracts the most false positives, again across all systems.



(a) BASE-EMB (b) BASE-GST

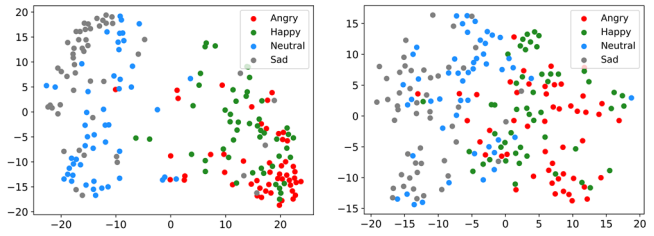
(c) OURS-TRANS (d) OURS-SYN

Fig. 5. Confusion matrices of emotion classification results on neutral speakers

4.4. A visualization: latent codes

We randomly select 200 sentences (50 per emotion) from test set and extract their phoneme-level latent codes from their ground truth mel spectrograms. The mean code of vowels in each utterance is visualized using t-SNE [42], shown in Fig. 6 (a). The result of predicted latents is given in Fig. 6 (b).

Fine-grained latents from ground truth sentences are clustered by emotion, showing that the latent codes do learn emotion related features. Angry and happy are high-arousal classes and typically higher pitch range and dynamics, while sad emotion is the opposite. This is in accord with the emotion distribution in Fig. 6 (a). Similar result is observed on predicted latents, except that the divergence between emotion classes are smaller.



(a) Ground truth (b) Predicted

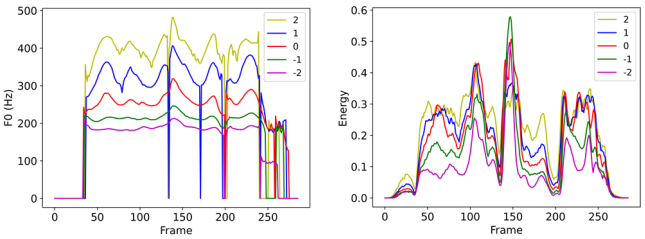
Fig. 6. T-SNE visualization of latent codes on different emotions. The mean code of vowels in each utterance is visualized

4.5. Controllability

4.5.1. Disentangled factors

To find out if our system learns meaningful latent codes, we set the value of each latent dimension to one of {-2, -1, 0, 1, 2} while keeping the other two at 0. Sharing this latent code across all phonemes, we re-run the TTS decoder and

find that the 1st dimension controls F0 of synthesized speech while the 3rd dimension controls the energy, as shown in Fig. 7. This shows our system has learned disentangled and acoustically meaningful latent features without direct supervision. More interestingly, these fine-grained latents can work as adjustable handles to such features, thereby open a way to controllable ESS.

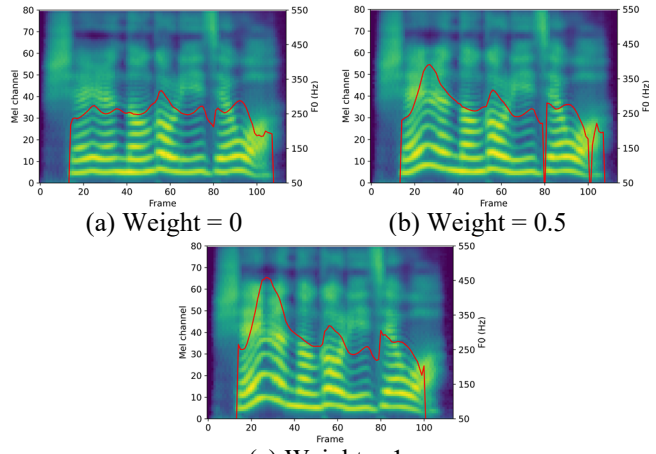


(a) 1st dimension – F0 (b) 3rd dimension - energy

Fig. 7. Controllability via latent codes

4.5.2. Emotion strength

We test a scaling based method for controlling emotion strength similar to [29]. Concretely, we synthesize a sentence with the angry tag, but multiplies the predicted latent codes by a constant weight. The mel spectrograms and pitch contours predicted with different emotion weights are shown in Fig. 8. As expected, the standard strength (weight=1) has higher pitch value and variation compared to moderate (weight=0.5) and weak (weight=0) levels, which indicates the controllability of emotion strength.



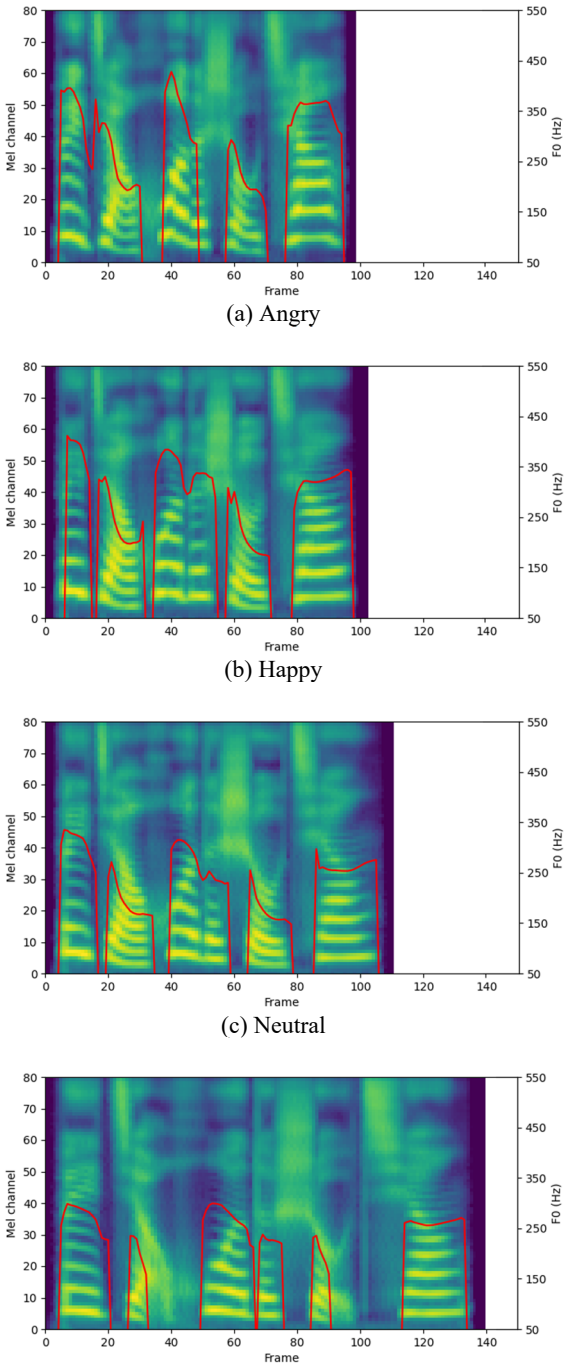
(a) Weight = 0 (b) Weight = 0.5

(c) Weight = 1

Fig. 8. Mel spectrogram and pitch contours for an angry utterance with different emotion weights

4.6. Qualitative inspections

Different emotions are known to carry different patterns in prosody. Fig. 9 shows mel spectrograms and pitch contours of synthesized examples in 4 emotion classes with the same text. We clearly see their differences in speaking rate and pitch modulation. The sad emotion shows slow speaking rate with narrow pitch variation (Fig. 9 (d)). In contrast, the angry emotion (Fig. 9 (a)) shows fast speaking rate and wide pitch variation.



(a) Angry

(b) Happy

(c) Neutral

(d) Sad

Fig. 9. Mel spectrograms and pitch contours of synthesized examples in 4 emotion classes with the same text

5. CONCLUSION

We introduce fine-grained latent prosody modelling to multi-speaker ESS under low emotion-resourced setting, based on the information bottleneck principle. We show that the latent variables can learn disentangled and meaningful acoustic features in unsupervised manner, which in turn improves transferability of emotions learned from emotional speakers to non-emotional speakers. We have demonstrated controlling emotion strength using these latent codes.

For future work we consider applying our design principles to a wider range of expressive synthesis tasks like emphatic expressions, Lombard/whispered speech, and voice acting. We also plan exploring real voice controllability use cases with the company's smart devices and voice services.

REFERENCES

[1] Lu, C., Wen, X., Liu, R., & Chen, X. Multi-speaker emotional speech synthesis with fine-grained prosody modeling. *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5729-5733 (2021)

[2] Gibiansky, A., Arik, S., Diamos, G., Miller, J., Peng, K., Ping, W., Raiman, J. & Zhou, Y. Deep voice 2: Multi-speaker neural text-to-speech. *Advances in neural information processing systems (NeurIPS)*, 2962-2970 (2017)

[3] Ping, W., Peng, K., Gibiansky, A., Arik, S. O., Kannan, A., Narang, S., Raiman, J. & Miller, J. Deep voice 3: Scaling text-to-speech with convolutional sequence learning. *International Conference on Learning Representations (ICLR)*, (2018)

[4] Wang, Y., Skerry-Ryan, R. J., Stanton, D., Wu, Y., Weiss, R. J. et al. Tacotron: Towards end-to-end speech synthesis. *Annual Conference of the International Speech Communication Association* (*INTERSPEECH)*, 4006-4010 (2017)

[5] Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N. et al. Natural TTS synthesis by conditioning wavenet on MEL spectrogram predictions. *International Conference on Acoustics, Speech, and signal Processing (ICASSP),* 4779-4783 (2018)

[6] Tachibana, H., Uenoyama, K., & Aihara, S. Efficiently trainable text-to-speech system based on deep convolutional networks with guided attention. *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4784-4788 (2018)

[7] Li, N., Liu, S., Liu, Y., Zhao, S., & Liu, M. Neural speech synthesis with transformer network. *Proceedings of the AAAI Conference on Artificial Intelligence* **33**(01), 6706-6713 (2019)

[8] Bahdanau, D. et al. Neural machine translation by jointly learning to align and translate. *3rd International Conference on Learning Representations (ICLR)*, San Diego (2015)

[9] Ren, Y., Ruan, Y., Tan, X., Qin, T., Zhao, S., Zhao, Z., & Liu, T. Y. Fastspeech: Fast, robust and controllable text to speech. *Advances in neural information processing systems (NeurIPS)*, 3165–3174 (2019)

[10] Peng, K., Ping, W., Song, Z., & Zhao, K. Non-autoregressive neural text-to-speech. *International Conference on Machine Learning (ICML)*, 7586–7598 (2020)

[11] Kim, J., Kim, S., Kong, J., & Yoon, S. Glow-tts: A generative flow for text-to-speech via monotonic alignment search. *Advances in Neural Information Processing Systems (NeurIPS)*, 33 (2020)

[12] Ren, Y., Hu, C., Tan, X., Qin, T., Zhao, S., Zhao, Z., & Liu, T. Y. Fastspeech 2: Fast and high-quality end-to-end text to speech. *International Conference on Learning Representations (ICLR)*, (2021)

[13] Skerry-Ryan, R. J., Battenberg, E., Xiao, Y., Wang, Y., Stanton, D. et al. Towards end-to-end prosody transfer for expressive speech synthesis with tacotron. *International Conference on Machine Learning (ICML)*, 4700-4709 (2018)

[14] Wang, Y., Stanton, D., Zhang, Y., Ryan, R. S., Battenberg, E. et al. Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis. *International Conference on Machine Learning (ICML)*, 5167-5176 (2018)

[15] Akuzawa, K., Iwasawa, Y., & Matsuo, Y. Expressive speech synthesis via modeling expressions with variational autoencoder. *Annual Conference of the International Speech Communication Association* (*INTERSPEECH)*, 3067-3071 (2018)

[16] Hsu, W. N., Zhang, Y., Weiss, R. J., Zen, H., Wu, Y. et al. Hierarchical generative modeling for controllable speech synthesis. *International Conference on Learning Representations (ICLR)*, (2018)

[17] Zhang, Y. J., Pan, S., He, L., & Ling, Z. H. Learning latent representations for style control and transfer in end-to-end speech synthesis. *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6945-6949 (2019)

[18] Klimkov, V., Ronanki, S., Rohnke, J., & Drugman, T. Fine-grained robust prosody transfer for single-speaker neural text-to-speech. *Annual Conference of the International Speech Communication Association* (*INTERSPEECH)*, 4440-4444 (2019)

[19] Lee, Y & Kim, T. Robust and fine-grained prosody control of end-to-end speech synthesis. *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5911-5915 (2019)

[20] Sun, G., Zhang, Y., Weiss, R. J., Cao, Y., Zen, H., & Wu, Y. Fully-hierarchical fine-grained prosody modeling for interpretable speech synthesis. *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6264-6268 (2020)

[21] Sun, G., Zhang, Y., Weiss, R. J., Cao, Y., Zen, H. et al. Generating diverse and natural text-to-speech samples using a quantized fine-grained VAE and autoregressive prosody prior. *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6699–6703 (2020)

[22] Qian, K., Zhang, Y., Chang, S., Yang, X. & Hasegawa-Johnson, M. AutoVC: Zero-Shot Voice Style Transfer with Only Autoencoder Loss. *International Conference on Machine Learning (ICML)*, 5210-5219 (2019)

[23] Qian, K., Zhang, Y., Chang, S., Hasegawa-Johnson, M. & Cox, D. Unsupervised Speech Decomposition via Triple Information Bottleneck. *International Conference on Machine Learning (ICML)*, 7836-7846 (2020)

[24] An, S., Ling, Z., & Dai, L. Emotional statistical parametric speech synthesis using LSTM-RNNs. *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 1613-1616 (2017)

[25] Lorenzo-Trueba, J., Henter, G. E., Takaki, S., Yamagishi, J., Morino, Y., & Ochiai, Y. Investigating different representations for modeling and controlling multiple emotions in DNN-based speech synthesis. *Speech Communication* **99**, 135-143 (2018)

[26] Lee, Y., Rabiee, A., & Lee, S. Y. Emotional end-to-end neural speech synthesizer. *Machine Learning for Audio Signal Processing (NeurIPS workshop)*, (2017)

[27] Wu, P., Ling, Z., Liu, L., Jiang, Y., Wu, H., & Dai, L. End-to-end emotional speech synthesis using style tokens and semi-supervised training. *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 623-627 (2019)

[28] Kwon, O., Jang, I., Ahn, C., & Kang, H. G. An effective style token weight control technique for end-to-end emotional speech synthesis. IEEE Signal Processing Letters **26**(9), 1383-1387 (2019)

[29] Li, T., Yang, S., Xue, L., & Xie, L. Controllable emotion transfer for end-to-end speech synthesis. *International Symposium on Chinese Spoken Language Processing (ISCSLP)*, 1-5 (2021)

[30] Um, S. Y., Oh, S., Byun, K., Jang, I., Ahn, C., & Kang, H. G. Emotional speech synthesis with rich and granularized control. *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 7254-7258 (2020)

[31] Zhu, X., Yang, S., Yang, G., & Xie, L. Controlling emotion strength with relative attribute for end-to-end speech synthesis. *Automatic Speech Recognition and Understanding (ASRU)*, 192-199 (2019)

[32] Lei, Y., Yang, S., & Xie, L. Fine-grained emotion strength transfer, control and prediction for emotional speech synthesis. *Spoken Language Technology Workshop (SLT)*, 423-430 (2021)

[33] Choi, H., Park, S., Park, J., & Hahn, M. Multi-speaker emotional acoustic modeling for CNN-based speech synthesis. *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6950-6954 (2019)
[34] Chen, L., Braunschweiler, N., & Gales, M. J. Speaker and expression factorization for audiobook data: Expressiveness and transplantation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* **23**(4), 605-618 (2015)
[35] Ohtani, Y., Nasu, Y., Morita, M., & Akamine, M. Emotional transplant in statistical speech synthesisbased on emotion additive model. *Annual Conference of the International Speech Communication Association* (*INTERSPEECH)*, 274-278 (2015)
[36] Inoue, K., Hara, S., Abe, M., Hojo, N., & Ijima, Y. An investigation to transplant emotional expressions in DNN-based TTS synthesis. *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 1253-1258 (2017)
[37] Kingma, D. P. & Welling, M. Auto-encoding variational bayes. *International Conference on Learning Representations (ICLR)*, (2014)
[38] Griffin, D.W. & Lim, J. S. Signal estimation from modified short-time fourier transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing* **32**(2), 236-243 (1984)
[39] Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H. et al. Domain-adversarial training of neural networks. *Journal of Machine Learning Research* **17**(1), 2096-2030 (2017)
[40] Zhang, Y., Weiss, R. J., Zen, H., Wu, Y., Chen, Z. et al. Learning to speak fluently in a foreign language: Multilingual speech synthesis and cross-language voice cloning. *Annual Conference of the International Speech Communication Association* (*INTERSPEECH)*, 2080-2084 (2019)
[41] Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X. et al. Beta-VAE: Learning basic visual concepts with a constrained variational framework. *International Conference on Learning Representations (ICLR)*, (2017)
[42] Van der Maaten, L., & Hinton, G. Visualizing data using t-SNE. *Journal of machine learning research* **9**(11), (2008)