Emphatic Speech Synthesis and Control Based on Characteristic Transferring in End-to-End Speech Synthesis

1st Mu Wang
Tsinghua University
Shenzhen, China
wangmu16@mails.tsinghua.edu.cn

2nd Zhiyong Wu
Tsinghua University
The Chinese University of Hong Kong
Shenzhen, China
zywu@se.cuhk.edu.hk

3rd Xixin Wu
The Chinese University of Hong Kong
Hong Kong, China
wuxx@se.cuhk.edu.hk

4th Helen Meng
The Chinese University of Hong Kong
Tsinghua University
Hong Kong, China
hmmeng@se.cuhk.edu.hk

5th Shiyin Kang
Tencent AI Lab
Shenzhen, China
shiyinkang@tencent.com

6th Jia Jia
Tsinghua University
Beijing, China
jjia@tsinghua.edu.cn

7th Lianhong Cai
Tsinghua University
Beijing, China
clh-dcs@tsinghua.edu.cn

***Abstract*—End-to-end text-to-speech (E2E TTS) synthesis has achieved great success. This work investigates the emphatic speech synthesis and control mechanisms in the E2E framework and proposes an E2E-based method for transferring emphasis characteristic between speakers. Characteristic differences between emphatic and neutral speech are learned from a small-scale corpus containing parallel neutral and emphasis speech utterances recorded by one speaker and further transferred to another speaker so that we can generate emphatic speech with latter speakers voice. Emphasis embedding is injected to the encoder of the extended E2E TTS model to capture the aforementioned differences; while the decoder and attention module are used to decode those differences into synthetic neutral / emphatic speech. Speaker codes linked to the decoder and attention module provide the E2E model the ability for characteristic transferring between speakers. To control the emphatic strength, an encoder memory manipulation mechanism is proposed. Experimental results indicate the effectiveness of our proposed model.**

***Index Terms*—end-to-end, expressive speech, multi-speaker speech synthesis, transfer learning, emphatic speech**

I. Introduction

With the progressive development of deep neural networks (DNN), nowadays neutral speech synthesis has achieved great success. Reference [1] proposed Deep Voice, a production-quality text-to-speech (TTS) system constructed entirely from DNN. Tacotron, a complicated end-to-end (E2E) TTS model, was proposed in [2] and obtained superior performance over a productive statistical parametric speech synthesis system in terms of naturalness. However, it still remains a challenge to control the synthesis model to generate speech with desired characteristics such as emotion, emphasis, speaker, etc.

Traditional hidden Markov model (HMM) based speech synthesis is highly controllable [3], [4]. Latest research [5] indicated it is feasible to control DNN-based speech synthesis model using input codes. Reference [6] used implicit emotion codes to synthesize expressive speech. Both studies revealed that the control codes could be determined simultaneously with other parts of the model [5], [6]. For E2E TTS synthesis, Tacotron [2] can achieve better naturalness in synthetic speech than HMM-based methods. However, the flexibility in controlling prosodic and spectral characteristics is compromised. Even though [2] conjectured E2E TTS system more easily allows for rich conditioning on various attributes, such research has not been well studied. Reference [7] performed multi-speaker E2E TTS synthesis by extending Tacotron, exhibiting the feasibility to control E2E TTS model. Reference [7] also found it is necessary to incorporate speaker embeddings into the character encoder, otherwise, the model is incapable of learning its attention mechanism and cannot generate meaningful output. However, no theoretical proof is provided for this finding. Furthermore, instead of global control such as multi-speaker synthesis, effective word- or phoneme-level local control strategies that are required for some of the tasks like emphatic speech synthesis have not been studied for E2E TTS synthesis.

This paper investigates the emphatic speech synthesis and control mechanisms in the E2E TTS framework and proposes an E2E-based method for emphasis characteristic transferring between different speakers. Traditional methods for emphatic speech synthesis usually employ two speech corpora from the same speaker to train the model, one large-scale neutral corpus to ensure the voice quality of synthetic speech and

National Natural Science Foundation of China-Research Grant Council of Hong Kong (NSFC-RGC) joint research fund (61531166002, N_CUHK404/15), National Natural Science Foundation of China (61433018, 61375027), National High Technology Research and Development Program of China (2015AA016305) and National Social Science Foundation of China (13&ZD189). With support from Tsinghua University - Tencent Joint Laboratory.

978-1-5386-5311-1/18/$31.00 ©2018 IEEE

the other small-scale emphatic corpus to fine-tune the model so that it can generate speech with appropriate emphasis characteristics. However, when only neutral corpus of the target speaker is available, how to synthesize emphatic speech of this specific speaker remains a challenging problem. In this paper, a small-scale parallel corpus (with neutral and emphatic speech) of another different speaker is adopted to represent the characteristic differences between emphatic and neutral speech. Such differences are modeled by the encoder and corresponding encoder memory through the joint training of encoder and decoder of the E2E TTS model, which are further transferred to the target speaker by the decoder and attention module with the aid of speaker codes. Furthermore, an encoder memory manipulation mechanism is proposed so that the emphasis strength of the synthetic speech can be easily controlled.

II. Model

A. Model architecture

The architecture of the controllable multi-speaker end-to-end emphatic speech synthesis model is illustrated in Fig. 1.

During training, the model learns to infer mel- and linear-scale spectrogram from character sequence. In the encoder side, the input character / emphasis flag sequence is first mapped to a distributed embedding representation by embedding lookup. The embeddings, after processed by the PreNet, are then fed to the CBHG architecture to generate the encoder memory. The CBHG [2] is composed of a bank of convolutional filters (ConvNet), highway networks (HighwayNet) and the bidirectional gated recurrent units (BiGRU). In the decoder side, the previous time step ground-truth mel-spectrogram, pre-processed by the PreNet, is sent to the AttentionRNN, together with the Attention module, to generate the attention of current time step. Then the AttentionRNN output and the attention context of current time step are fed to the DecoderRNN to generate the mel-spectrogram output of current time step. The PostNet is then used to convert the mel-spectrogram to the linear-scale spectrogram.

During synthesis, what is different from the training phase is that the previous time-step output instead of ground-truth mel-spectrogram is used in the decoder side.

B. Emphasis characteristic transferring between speakers

Different from the basic Tacotron [2] model, we inject the emphasis information into the CBHG architecture so that the encoder can model the differences between the neutral and emphatic inputs. For the decoder, speaker embedding (SpeakerEmb, which is trained by back-propagation) is used and linked to the AttentionRNN and DecoderRNN to capture the acoustic differences of speakers.

Through joint training of emphasis-injected encoder, speaker-dependent decoder and the attention module, the model is able to learn the differences between the emphatic and neutral speech of one speaker and *transfer* such information to another speaker with just *neutral* recordings. More details are elaborated as follows. With the above design, using the training data from the speaker with both emphatic and neutral speech recordings, the acoustic differences between emphatic and neutral speech are modeled by the decoder and attention module, and are conditioned on the emphasis related linguistic differences captured by the encoder memory of the emphasis-injected encoder. Furthermore, using the training data from different speakers with neutral speech recordings, the speaker related acoustic differences are also modeled by the decoder and attention module with the aid of speaker embedding codes. To generate emphatic speech for the speaker with only training data of neutral speech recordings, what we need to do is just set the appropriate speaker embedding code and feed the emphasis character sequence input to the encoder.

C. Different setups of linguistic emphatic encoder

To synthesize emphatic speech, emphasis information must be fed to the model. We designed several encoders, as shown in Fig. 2, to find the most effective way to inject emphasis information into the basic encoder that consists of one PreNet layer and one CBHG module.

1) **Emphatic encoder A (EEA).** The architecture of EEA is the same as the basic encoder, but the input is different. For EEA, the input character sequence is augmented with the emphatic code, "0" for neutral "1" for emphasized, at the end of every word. An example sentence for EEA is "The0 trend1 of0 pretending0 to0 contend1 has0 extended0." where "trend" and "contend" are emphasized.
2) **Emphatic encoder B (EEB).** In EEB, we investigate injecting the emphatic embedding (EmphasisEmb) into ConvNet. The input of EEB consists of two parts, the character sequence and the emphasis flag sequence. For characters in the character sequence, their emphatic codes (1 or 0 indicating if the corresponding character is from emphasized or neutral word) compose the emphasis flag sequence. The EmphasisEmb, a distributed representation of the emphatic codes by embedding lookup, is concatenated with the outputs of the encoder PreNet and fed to the ConvNet.
3) **Emphatic encoder C (EEC).** In EEC, a little different from EEB, EmphasisEmb is concatenated with the outputs of the HighwayNet and fed to the BiGRU in CBHG. For both EEB and EEC, the dimension of EmphasisEmb is set to 32, while a smaller one works as well.

D. Speaker-dependent decoder

In this work, we use two speech corpora recorded by different speakers for emphatic speech synthesis, a large scale neutral corpus by one speaker and a small scale emphatic corpus by another speaker. The purpose is to train the model to learn the emphasis characteristics from the emphatic corpus and then transfer them to the neutral corpus speaker. We extend Tacotron [2] to handle multi-speaker speech synthesis. An insight into the functionalities of different modules of the E2E model, the DecoderRNN may act as the traditional acoustic module of a TTS system, the AttentionRNN has the

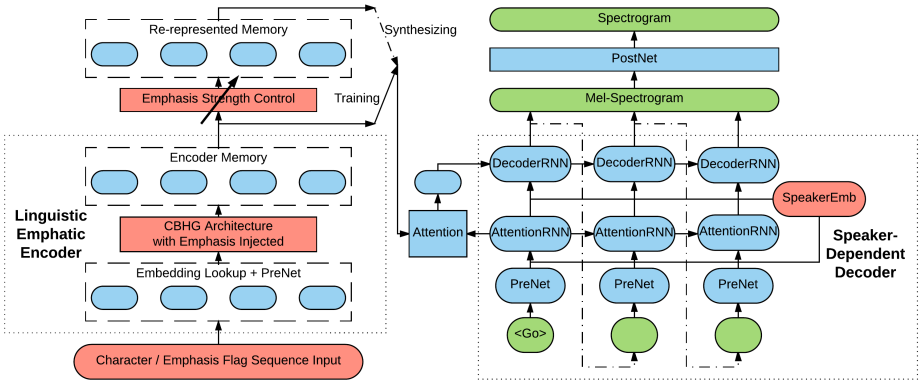

Fig. 1. Architecture of the controllable emphatic multi-speaker end-to-end text-to-speech (TTS) model.

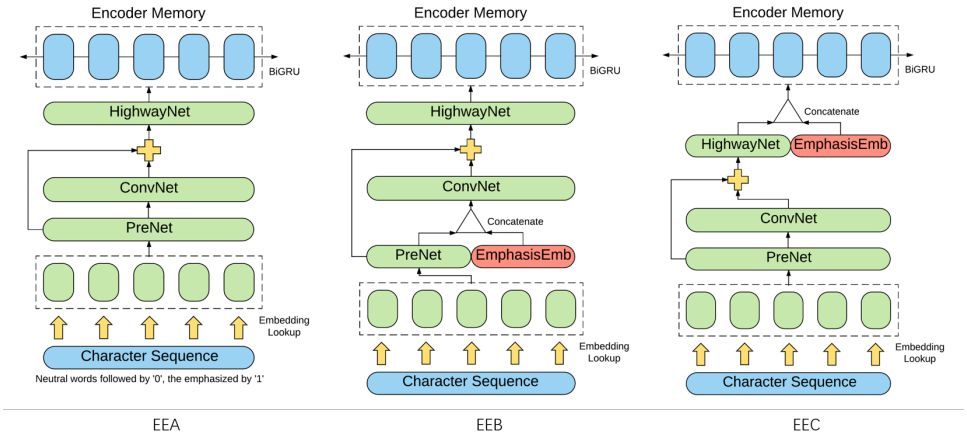

Fig. 2. Different setups of the emphatic encoder.

similar function as the traditional duration module, while the encoder serves as the module for linguistic emphatic feature extraction and representation. In this way, we can implement multi-speaker speech synthesis like [5].

In this work, the speaker information is only considered and injected to the decoder, leading to the speaker-dependent decoder (SDD); while the encoder is speaker independent (i.e. speaker independent encoder, SIE). We concatenate the speaker embedding, the context vector and the AttentionRNN output as the input of the DecoderRNN; and concatenate the SpeakerEmb and the decoder PreNet output as the input of the AttentionRNN. SpeakerEmbs are learned by back-propagation in conjunction with other parts of the model. The dimension of SpeakerEmb is set to 32.

E. Emphasis strength control

In the E2E model, the emphatic encoder, the decoder and the attention module are jointly trained. The output of encoder module, i.e. the encoder memory, captures information from not only the input character / emphasis flag sequences but also the acoustic characteristics back-propagated from the decoder and attention modules.

During synthesis, for the same character sequence input, we can get either an emphatic encoder memory (M_{emp}) when desired emphasis flag sequence is given or a neutral encoder memory (M_{neu}) when the emphasis codes are all set to 0. Such memories are further fed to attention and decoder modules to generate emphatic or neutral synthetic speech respectively. To control the emphasis strength of synthetic speech, we propose an encoder memory manipulation mechanism by linear interpolation to derive the re-represented memory (M_{re}), as shown in Fig. 3. The emphasis strength can be controlled by the hyperparameter α. The larger value of α, the stronger the emphasis level is.

III. Experiments and Analysis

A. Corpora

Two corpora are used in our work for experiments. The first one is a large-scale neutral corpus from Blizzard Challenge 2011, which contains about 10 hours of speech data uttered by a professional female speaker, Nancy.

The second one is a small-scale emphatic corpus with parallel neutral and emphatic speech recordings. 350 text prompts, each of which contains one or more emphatic words, are

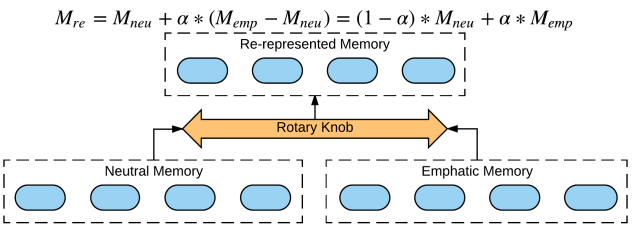

Fig. 3. Emphasis strength control.

designed to cover all pronunciation mechanisms of phonemes. However, since we only have 350 text prompts, apparently, the phonetic contexts are not covered thoroughly. Two contrastive speech utterances, a neutral version and an emphatic one, are recorded for each of the prompts by another professional female speaker in a sound proof studio.

The large-scale Nancy corpus ensures the well training of the E2E model to generate speech with high voice quality, while the small-scale emphatic corpus is used for modeling emphasis characteristics that are further transferred to Nancy speaker to synthesize emphatic speech with Nancys voice.

B. Experimental setup

Since we synthesize speech directly from character sequences, common text analysis routine is not needed. The text prompts are just preprocessed by a simple normalization procedure to convert the digits into spoken form according to the speech utterance, e.g. "1908" is normalized to "nineteen O eight".

The speech waveforms of the two corpora are sampled at 16 kHz. Griffin-Lim [8] is used to reconstruct waveform, only spectrogram is extracted. Before extracting features, all waveforms are pre-emphasized with a coefficient of 0.97 as suggested by [2]. Spectral analysis is conducted with Hann windowing, 50 ms frame length, 12.5 ms frame shift and 1024-point fast Fourier transform (FFT). 513-dimensional spectrogram is set as the target of the post-processing net of the model, and 80-band mel-spectrogram is set as the target of the decoder. The log-magnitude spectrograms of training set are standardized to have zero mean and unit variance. The spectrograms of validation and test set are normalized using the statistical parameters of training set.

We randomly select 80% of the neutral and emphatic corpus as training set, 10% as validation set and the rest as test set. The parallelization property of the emphatic corpus is kept for the training, validation and test set. Since the basic E2E TTS model we adopted is Tacotron [2], which predicts several frames at one decoding time step, we set the reduction rate (r) as 5, i.e. five frames are predicted at each decoding time step. Adam optimizer [9] is used with fixed learning rate 0.001. All our models are trained for 50,000 global steps with a batch size of 32, while longer training time may further improve the quality of synthetic speech.

To evaluate the proposed model, two sets of subjective tests are conducted. 15 sentences were randomly selected from the test set for synthesizing speech. 20 subjects without listening impairment were invited to participate in the tests.

C. Experiments on the effect of speaker code in encoder

As aforementioned, [7] claimed speaker embedding must be sent to the encoder to learn attention and generate meaningful output. We conducted two subjective tests to compare the encoder without speaker embedding (i.e. the SIE proposed in our work) and the encoder with speaker embedding (speaker dependent encoder, SDE). SDE is built by injecting speaker embedding into the HighwayNet of the encoder CBHG, as suggested in [7]. Both methods are used to synthesize neutral speech for large-scale corpus (SDE-L, SIE-L) and small-scale corpus (SDE-S, SIE-S). Corresponding original recordings from the test set are also provided during tests.

1) **Speaker similarity test.** The subjects were asked to choose which of 2 synthetic neutral speech, generated by SDE and SIE, sounds more similar to the reference recording.
2) **Multi-speaker speech naturalness test.** The subjects were asked to rate the naturalness in 5-point scale for the speech files: 2 neutral utterances synthesized by SDE and SIE and 1 natural reference speech recording.

The results are given in Fig. 4 and 5. From Fig. 4, no apparent preference can be observed between SIE and SDE for both speakers. From Fig. 5, there are also no significant differences for MOS between SIE and SDE for two speakers, respectively. So we speculate that it is not necessary to inject speaker information into the encoder, while further experiments are needed to validate this idea in future work.

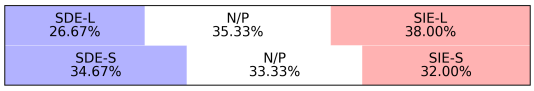

Fig. 4. Results for speaker similarity test.

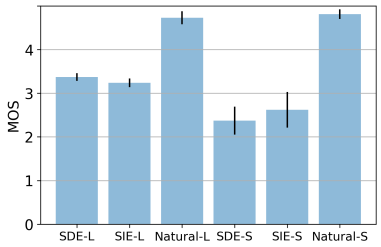

Fig. 5. Results for multi-speaker naturalness test.

D. Experiments for emphasis characteristic transferring on different setups of emphatic encoder

To validate emphasis characteristic transferring and investigate an effective way of injecting emphatic control signal, we further conducted two subjective tests.

1) **Emphatic speech naturalness test on the neutral-corpus speaker.** The first test is for evaluating the naturalness of the emphatic speech synthesized by the models with different setup of emphatic encoder (EEA, EEB or EEC). In this test, only synthetic speech of large-scale corpus speaker (L) is evaluated. The results are shown in Fig. 6, which indicate that EEA and EEB can achieve comparable naturalness of the synthetic emphatic speech, and they both outperform EEC. A possible explanation is that the convolution layers (ConvNet) can better model the complex context information of emphatic words in EEA and EEB.

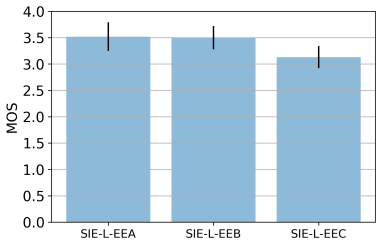

Fig. 6. Results for emphatic speech naturalness test.

2) **Emphasis identification test on the neutral-corpus speaker.** In this test, the subjects are asked to identify all the emphasized words in the synthetic speech. Another 10 sentences were randomly selected from the test set. Due to the similar performance of EEA and EEB, only EEA setup was evaluated. The precision and recall of the perceived emphatic words are 89.47% and 94.44%, which indicates the high performance of the model in generating emphatic speech.

E. Analysis on the encoder memory in capturing and controlling emphasis characteristics

As mentioned in II-E, the encoder memory is expected to capture information from not only the input sequences but also the emphasis characteristics. To validate this, we further performed analysis of the learnt encoder memory. We first calculate the difference vector E_{diff} by subtracting corresponding bits of emphatic encoder memory (M_{emp}) and neutral memory (M_{neu}) of the same character sequence input, at each encoding step. K-means clustering is then performed to cluster all the samples of E_{diff} from the training set into different categories.

We found that the vectors can be clustered in an intuitive way when $K = 7$. The clustering results of the log-scale Euclidean and cosine distance between the emphatic and neutral memory of an example sentence is shown in Fig. 7. As can be seen, the nearer the word is to the emphatic word, the larger the E_{diff} is. This confirms to the phonetic knowledge of acoustic realization of emphasis that the emphasized speech segments tend to have more influence to its adjacent segments and such influence will degrade for segments far away from the emphasized one [10]–[12]. As for the clustering result, it is quite similar to the manually defined 6 emphasis categories as shown in [13]. All these promising findings indicate that our proposed method is really effective in capturing the emphasis characteristics.

Furthermore, the method to control emphasis levels as proposed in II-E also utilizes the above findings. By tuning hyperparameter α of linear interpolation, the method changes the value of E_{diff} vector that will further lead to the change of acoustic realization of emphasis degree. To demonstrate the effectiveness of our emphatic strength control method, we analyzed the pitch under different α for a sentence. As shown in Fig. 8, the larger α is, the higher the pitch of the emphasized word is. Furthermore, the acoustic characteristics of the words around the emphasized are influenced too. Besides pitch, duration is also an important acoustic correlates of emphasis. Fig. 9 shows the word duration under different α. The larger α is, the longer the duration of emphasized word is. All these observations confirm to the phonetic knowledge of acoustic realization of emphasis, indicating the effectiveness of our proposed method.

IV. Acknowledgement

This work is supported by National Natural Science Foundation of China-Research Grant Council of Hong Kong (NSFC-RGC) joint research fund (61531166002, N_CUHK404/15), National Natural Science Foundation of China (61433018, 61375027), National High Technology Research and Development Program of China (2015AA016305) and National Social Science Foundation of China (13&ZD189). We would also like to thank Tsinghua University - Tencent Joint Laboratory for the support.

V. Conclusion

In this paper, we tried to transfer the emphasis characteristics from the small-scale corpus to the larger one. To achieve this purpose, we designed a multi-speaker end-to-end TTS model with SIE, found to be comparable to SDE. We also investigated the different ways of injecting emphatic control information; results showed that EEA and EEB achieved the best performance. In addition, we found that the emphatic strength of synthesized speech could be controlled in a simple way. We did some objective analysis and found that the model could learn the influence of emphatic words on their neighbors. For future work, we plan to transfer other character of a small-scale corpus to a larger one.

References

[1] Sercan . Ark, Mike Chrzanowski, Adam Coates, Gregory Diamos, Andrew Gibiansky, Yongguo Kang, Xian Li, John Miller, Andrew Ng, Jonathan Raiman, Shubho Sengupta, and Mohammad Shoeybi. "Deep Voice: Real-time Neural Text-to-Speech." In Proc.International Conference on Machine Learning, pp. 195-204, 2017.

[2] Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J. Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, Quoc Le, Yannis Agiomyrgiannakis, Rob Clark, and Rif A. Saurous. "Tacotron: A fully end-to-end text-to-speech synthesis model." CoRR abs/1703.10135, 2017.

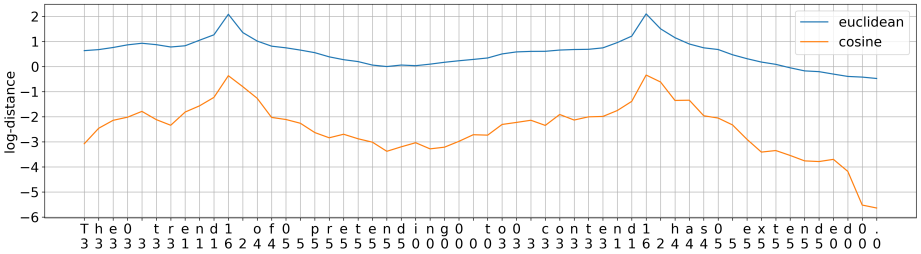

Fig. 7. The clustering result of the log-scale Euclidean and cosine distance of the emphatic memory and its neutral counterpart. The digits in the bottom line stand for cluster labels.

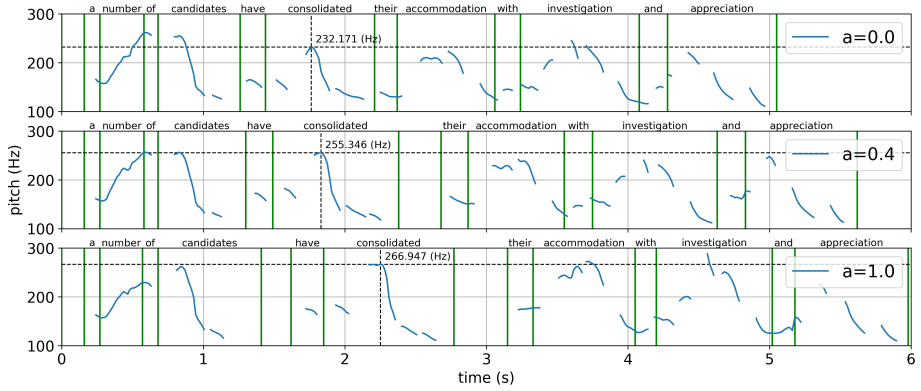

Fig. 8. Pitch contour for different emphatic strength by changing α. The sentence is "A number of candidates have consolidated their accommodation with investigation and appreciation." The emphasized words are "candidates" and "consolidated".

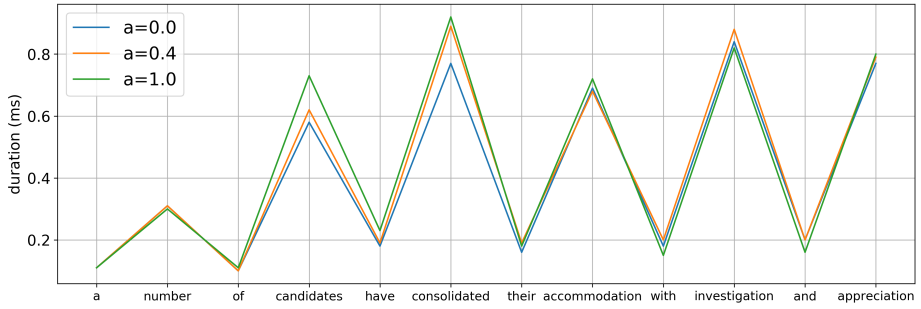

Fig. 9. Word duration for different emphatic strength by changing α. In this sentence, "candidates" and "consolidated" are emphasized.

[3] Junichi Yamagishi, and Takao Kobayashi. "Average-voice-based speech synthesis using HSMM-based speaker adaptation and adaptive training." IEICE Transactions on Information and Systems, vol. 90, no. 2, pp. 533-543, 2007.

[4] Takashi Nose, Junichi Yamagishi, Takashi Masuko, and Takao Kobayashi. "A style control technique for HMM-based expressive speech synthesis." IEICE Transactions on Information and Systems, vol. 90, no. 9, pp. 1406-1413, 2007.

[5] Hieu-Thi Luong, Shinji Takaki, Gustav Eje Henter, and Junichi Yamagishi. "Adapting and controlling DNN-based speech synthesis using input codes." In Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4905-4909, 2017.

[6] Gustav Eje Henter, Jaime Lorenzo-Trueba, Xin Wang, and Junichi Yamagishi. "Principles for learning controllable TTS from annotated and latent variation." In Proc. Annual Conference of the International Speech Communication Association (Interspeech), pp. 3956-3960, 2017.

[7] Sercan Arik, Gregory Diamos, Andrew Gibiansky, John Miller, Kainan Peng, Wei Ping, Jonathan Raiman, and Yanqi Zhou. "Deep Voice 2: Multi-Speaker Neural Text-to-Speech." InProc. Advances in Neural Information Processing Systems, pp. 2966-2974, 2017.

[8] Daniel Griffin, and Jae Lim. "Signal estimation from modified short-time Fourier transform." IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 32, no. 2 pp. 236-243, 1984.

[9] Diederik Kingma, and Jimmy Ba. "Adam: A method for stochastic optimization." arXiv preprint arXiv:1412.6980, 2014.

[10] Yunjia Wang, Min Chu, and Lin He. "An experimental study on the distribution of the focus-related and semantic accent in Chinese." Chinese Teaching in the World, vol. 2, pp. 12, 2006.

[11] Francisco Costa. "Intrinsic prosodic properties of stressed vowels in European Portuguese." In Proc. International Conference Speech Prosody, 2004.

[12] Szu-wei Chen, Bei Wang, and Yi Xu. "Closely related languages, different ways of realizing focus." In Proc. Annual Conference of the International Speech Communication Association (Interspeech), 2009.

[13] Yishuang Ning, Zhiyong Wu, Jia Jia, Fanbo Meng, Helen Meng, and Lianhong Cai. "HMM-based emphatic speech synthesis for corrective feedback in computer-aided pronunciation training." In Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4934-4938, 2015.