SJTU Entry in Blizzard Challenge 2019

Bo Chen, Kuan Chen*, Zhijun Liu, Zhihang Xu, Songze Wu,*
*Chenpeng Du, Muyang Li, Sijun Li, Kai Yu**

SpeechLab, Shanghai Jiao Tong University

bobmilk@sjtu.edu.cn, azraelkuan@sjtu.edu.cn, kai.yu@sjtu.edu.cn

Abstract

This paper presents the techniques that were used in sjtu-tts entry in Blizzard Challenge 2019. The main architecture is Tacotron with WaveNet vocoder. The corpus in BC2019 is 8 hours audios from a Chinese male speaker with mixed Mandarin and English speech. The audios and transcriptions are found on the Internet with heavily corruption and noise. To deal with the corpus, our system is divided into 4 parts, data preprocessing, spectrogram model, WaveNet vocoder and speech bandwidth extension. The WaveNet vocoder is more relative to the speech quality and the spectrogram model is more relative to the prosody(pitch and duration). We didn't successfully train a good WaveNet vocoder for the predicted mel-spectrogram. Thus, some useful techniques in other parts have no significant improvement after WaveNet vocoding. These attempts which were not included in the final submission are also analyzed.

Index Terms: speech synthesis, prosody modelling, tacotron, expressive speech synthesis

1. Introduction

In Blizzard Challenge 2019, the speech corpus is a list of Chinese talk-show from a single male speaker[1]. The speech waveform and the transcriptions are found on the Internet. The recording equipment and recording environments are both not stable and not provided. According to the transcription of the corpus, the audios were likely to be recorded by mobile phones in rooms. The spectrogram is heavily corrupted in variety of styles. The audios have very strong prosody including tones, speaking speed, stress and etc.

The participants are asked to build text-to-speech systems using the provided data and any other data except the extra audios from the same speaker. The systems are evaluated by the synthetic speech only, including naturalness, similarity and phone error rate. The text for evaluation are provided by the organizers including talk-show scripts, wikipedia, numbers, mixture of Chinese and English, ancient poetry, and special pronunciation in Mandarin.

The main architecture of the `sjtu-tts` entry is to map the input phoneme sequence into the spectrogram sequence by Tacotron [1, 2], then the spectrogram is vocoded into speech waveform by a WaveNet vocoder [3]. Tacotron with WaveNet is the state-of-art text-to-speech model that can build high quality speech synthesis system on clean corpus, but it is still under investigation on applying them on found corpus. In this paper, most of the techniques are proposed to overcome the problems in modelling the found speech using the typical framework. Because of the fact that: the corpus consist of both Chinese and English words, all the words are analyzed into phoneme combinations to form the input sequence instead of the original character sequence. Same to the typical Tacotron architecture, no more context features are provided to the input sequence. Since the corpus has only 8 hours speech, some clean auxiliary corpus are adopted to train a multi-speaker model. We use the speaker embedding to model the speaker identity and environment embedding to model the recording environment. The clean corpus are all given the same environment id, but each original audio is given a different environment id in `LZY` corpus, under the assumption that: each 1-minute audio is recorded using the same equipment in the same environment. We didn't make further manual labeling or clustering for the different environments.

[1] We call it `LZY` by the first characters of the speaker name.

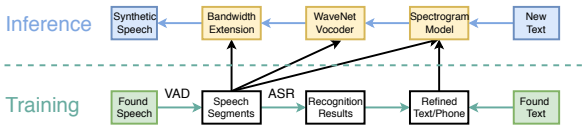


Figure 1: *Architecture of the `sjtu-tts` entry.*

2. System Description

2.1. Overall Architecture

The overall architecture is presented in Fig. 1 with 4 parts: data pre-processing, spectrogram model, WaveNet vocoder and bandwidth extension. In the training stage, the long waveform is first cut into a group of short waveform by a voice activity detection (VAD) module, then the audio segments are recognized with a multi-speaker hidden markov model(HMM) based automatic speech recognition (ASR) module. The recognized result of audio segments are then refined with the transcriptions of the whole sentence to get the correct text. The text are analyzed into phoneme sequence which are trained together with the mel-spectrograms of audios to get the spectrogram model. The WaveNet vocoder is trained with both the provided audios and audios from other speaker. The bandwidth extantion model is trained only with the audios from other speaker with higher sample-rate. In the inference stage, the new text is first analyzed into phoneme sequences. Then the spectrogram model predicts the spectrogram sequence. The spectrogram sequence is vocoded into waveform by the WaveNet vocoder, and finally upsampled to higher sample-rate.

2.2. Data Pre-processing

There are 480 different audios provided in `LZY` corpus. Each audio is a 1 minute talking-show recording. The long speech is first cut into short ones with a voice activity detection (VAD) tool. Then a multi-speaker GMM-HMM ASR system for Chinese is trained with Kaldi [4] to recognize the speech segments and align it with its corresponding transcription. Manual check is applied to correct the bad VAD sentences and missing syllables. After that, we get a processed corpus with short audios

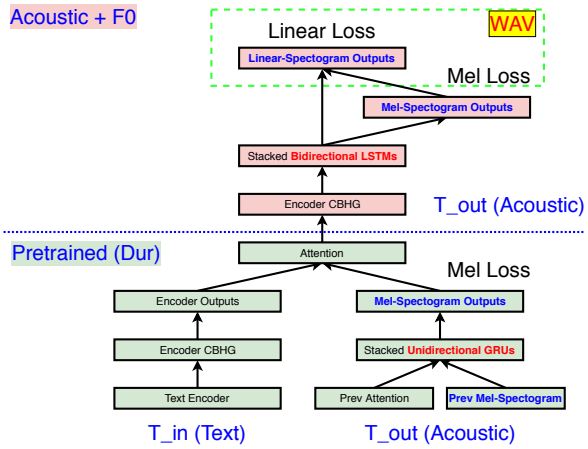


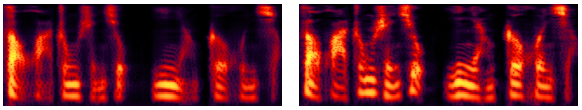
Figure 2: *Architecture of the spectrogram model.*

and transcriptions.

To train the synthesis system, the audio data is first resampled to 16kHz. Then 0.97 pre-emphasis is applied before DTFT. We use 2048-point Fourier transform and 240-band mel-scale spectrogram as the acoustic features. Text is converted into phoneme sequence with text analysis tools. As a feature of Chinese language, every Chinese character in a sentence can be mapped to a Chinese syllable (the mapping maybe varies in different sentences), and every Chinese syllable can be mapped to a certain sequence of phonemes. Therefore, the natural text is directly analyzed into phoneme sequences. For Chinese text, each syllable is split into the phone sequence from a standard mapping table. The same phoneme with different tones are treated as different phonemes. For English text, the phonemes are analyzed by FESTIVAL [5]. Since only a small part of the text is in English (e.g. APP, QQ.), we didn't pay more attention to the English phoneme set. The punctuation is treated as silence phoneme, and the question mark is treated the same with the period (because there is always a explicit "question character" before "question mark" in the LZY corpus).

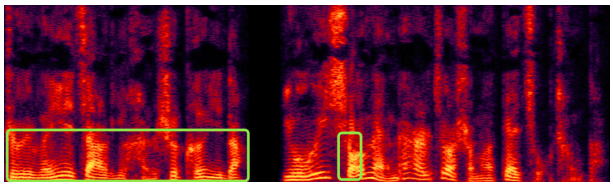
2.3. Spectrogram Model

The acoustic model that maps the phone sequence into the spectrogram sequence is described in Fig. 2. The blue part of architecture is a typical Tacotron, which is a well-known sequence-to-sequence spectrogram model. However, to reduce the error accumulation in inference stage, the decoded mel-spectograms are abandoned. The attentions in the decoding procedure are fed into a statistical parametric speech system (SPSS). The SPSS system follows the Emphasis system [6], which reports that the proposed method works better in modeling the emotional speech. Therefore, the Tacotron is mainly treated as a embedding model that maps the phone-level input sequence into frame-level embedding (attention) sequence. The embedding sequence is then decoded into the spectrogram by stacked bidirectional LSTMs. This architecture takes the advantage of bidirectional recurrent neural networks in predicting the mel-spectrograms. The Tacotron is pre-trained to get the attentions, and the post network is trained separately. It is different from the post net in Tacotron framework that the typical post net usually does not change the prosody. Fig. 3 shows the example of the spectrogram from Tacotron decoding and the spectrogram from the post network.

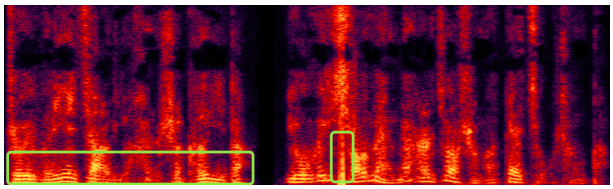


(a) Tacotron decoding. (b) Post network.

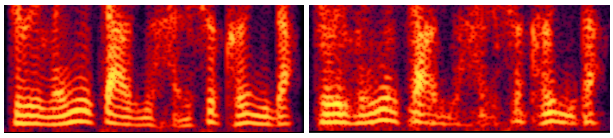
Figure 3: *Comparison between spectrogram from Tacotron decoding and post network (Vocoded by Griffin-Lim).*



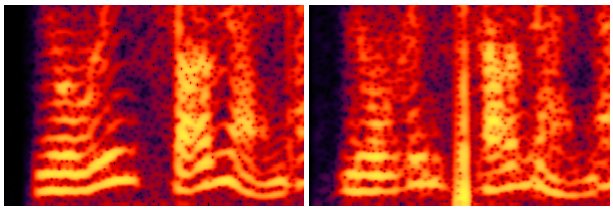
(a) LZY's speaker embedding



(b) The closest speaker embedding of LZY



(c) Expanded Area1



(d) Expanded Area2

Figure 4: *Different Speaker Embedding of WaveNet Vocoder*

2.4. WaveNet Vocoder Adaptation

In our system, we choose a 10-bit autoregressive WaveNet as the vocoder model that maps the acoustic feature to waveform. In order to get the initialization model, a unified WaveNet Vocoder model is first trained using multi-speaker dataset which include several male speakers. Besides the acoustic feature as the condition input, we add a speaker embedding vector which is expected to capture the speaker-related information as the extra condition input. In the adaption step, a new speaker embedding vector is learned using the training data from LZY and also update all the parameters in the speaker-dependent WaveNet Vocoder. As we know, the Wavenet Vocoder model gains most speaker-related information from acoustic feature, so in the inference step, we choose a internal speaker whose speaker embedding is the most closest to LZY to make the sampling more stable shown in Fig. 4.

2.5. Bandwidth Extension

Constrained by unstable recording equipment and environments, the target audio speech of LZY contains more noise in the high-frequency band than in the low-frequency band. Motivated by this, our acoustic model only targets on audio with 16kHz sample-rate though that of natural speech which is 22kHz to avoid the noise in high-frequency band. Then we apply a waveform modeling neural network on the 16K synthesized speech to upsample the 16kHz audio to 32kHz audio, namely speech bandwidth extension(BWE).

Our BWE model follows [7], which is a waveform modeling and generation method based on hierarchical recurrent neural networks(HRNN). The HRNN model represents the distribution of each high-frequency waveform sample conditioned on the input narrowband waveform samples using a neural network composed of hierarchical-structured LSTM and feed-forward layers. Limited to the quality and quantity of LZY speech, we trained a multi-speaker BWE model instead of solely using LZY speech to train a single-speaker model tailored for LZY. To improve the generality of our model, we collect about 2000h 32kHz[2] audio-book speech from 25 speakers without noise or background music from the Internet. To best fit LZY's speech, among the 25 speakers, 24 are male, and 21 are Chinese speakers. In our experiments, we only take approximately 2h speech of each speaker to both reduce the training time and balance the fitness of each speaker.

3. Experiment and Analyze

The identifier of our team is T in the Blizzard Challenge 2019 shown in Fig. 5 and Fig. 6. Our scores have a gap to the best systems. We suspect that the gap to the better teams is mainly caused by our vocoding that the WaveNet vocoded waveform has too much noise than the Griffin-Lim vocoded waveform using the same mel-spectrograms. It requires other techniques to make the WaveNet vocoder performs better on the predicted speech. The prosody is not explicitly measured in the final evaluation, thus we cannot measure the performance of our method to model the prosody.

3.1. Vocoder Selection

Vocoder is the key point to produce high quality synthetic speech. Each vocoder has its corresponding acoustic features (e.g. mel-spectrogram, linear-spectrogram, linear prediction coefficients). Recently, statistical vocoders have shown their great performance in vocoding high quality speech on clean corpus. Since the audio is heavily corrupted and noisy, the performances of statistical vocoders are not certain on LZY corpus. To select the vocoder, we examined several candidates vocoders with their corresponding acoustic features: WaveNet [3], LPCNet [8], Griffin-Lim [9] and WORLD [10].

We first trained the WaveNet and LPCNet on the training set of LZY corpus, then evaluated the performance on the evaluation set. All the acoustic features are directly extracted from the natural speech in 10ms frame-shift. The result shows that, both the WaveNet and the LPCNet performs good on the extracted acoustic features, and the performances are similar, which both outperform the signal-processing vocoders. But it is still not certain about which vocoder is more robust on the predicted acoustic features with different kinds of mismatches.

[2]We use 32kHz because the audios on the Internet is usually 44.1kHz in MP3 format, which has 32kHz valid spectrogram.

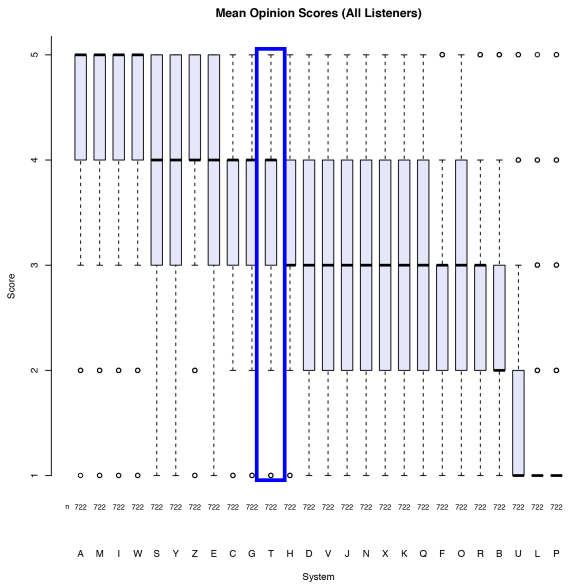


Figure 5: *Mean opinion score of all participants.*

We trained two naive Tacotron models, with mel-spectrogram and LPC as their acoustic features, and then vocoded the predicted acoustic features with the statistical vocoders. We still get the similar performance on these two vocoders. Eventually, we selected the WaveNet as the vocoder of our system since mel-spectrogram is a visible feature that we can directly look into the visiable image to find ways to improve the quality.

3.2. Spectrogram Model

3.2.1. Multi-Speaker Tacotron

In the preliminary experiments, we directly trained the Tacotron model on the LZY corpus with forward attention [11]. But there are too many mispronunciation in the inference stage, especially the mispronunciation in tones. It shows that, the data is not enough for the model to learn the difference between different tones, which indicates that the auxiliary corpus is necessary to fully cover the vowel with more audios in Chinese language. We selected some Chinese male corpus as auxiliary data, which have similar pitch contours compared to LZY (the speaker timbre are not similar). Some of the corpus has mixed Chinese and English text. All the corpus are first trained together to get a multi-speaker model, and then adapted to the LZY corpus. After doing this, the tone mistakes are largely solved.

We adopted the multi-speaker structure from the Deep Voice 2 [12], which concatenates the responding trainable speaker embedding vector with network inputs in different model positions like encoder, decoder, prenet, postnet etc. Furthermore, we compared the effects of different positions to concatenate embeddings and the amount of extra speakers in this system. The result is that speaker embedding always works and has little differences in different model positions.

3.2.2. Acoustic Feature Selection

Since we adopted WaveNet as the vocoder, mel-spectrograms are elected as the acoustic features. But we still seeking for taking the advantage of other acoustic features to help train-

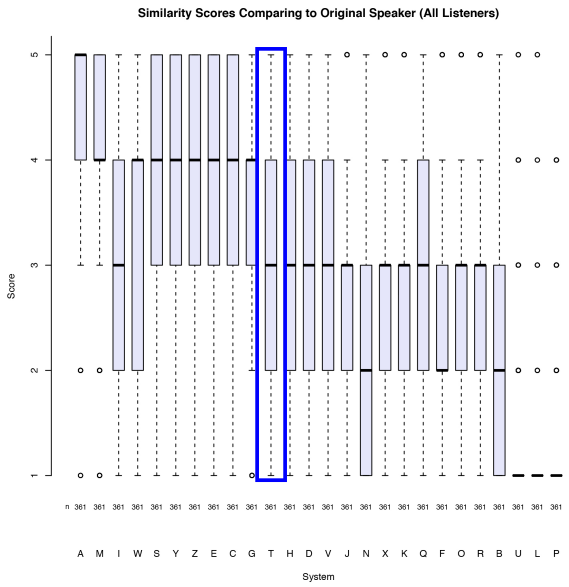


Figure 6: *Similarity score of all participants.*

ing the spectrogram model. So, the mel-spectrograms are extracted 10ms per frame in our system in order to train multi-task model with other acoustic features. The multiple task Tacotron model's structure is almost the same with the typical structure. Its' decoder is fed with a concatenated multiple acoustic features and predict different acoustic features in one decoding step, and compute the prediction loss separately.[3] We tried the different combination method of these features and compared the final speech quality. The result is that adding these additional tasks do introduce some variation in rhythm but it also hurts the intelligibility like mispronouncing the tone in several phonemes. So we simply use the mel-spectrogram as the acoustic features.

3.2.3. Linguistic Features Selection

A typical Tacotron model only uses text as its input sequence. We try to find that whether applying other linguistic features could help control the rhythm of synthesized speech. Widely used linguistic features in natural language processing field include postag, name entity recognition, syntax dependencies. We do word segmentation first and extract the features mentioned above of each segmented word with an open-sourced Chinese text analysis tool LTP[13]. Postag set is 863 standard Chinese postag set containing 28 different types. The provided model of LTP for postag prediction reached 98.35% accuracy on its People's Daily dataset. Name entity recognition model of LTP predicts the name of person, institution and location. It achieves 94.17% F-scrore on People's Daily dataset. We use syntax dependency analysis model whose UAS and LAS score, two criterion for syntax dependency analysis, are 84.00 and 81.14 respectively on Chinese Dependency Treebank dataset. All these linguistic features are embeded as vectors of their corresponding word and concatenated to either phoneme sequence embeddings or encoder outputs. However, our experiments show that almost no perceptible rhythm change is obtained after these features are used as parts of input.

[3] Actually we also tried to train multihead attentions with the multiple acoustic feature, but we found that mel-spectrum is much easier to converge than LPC and gradually, the model just ignored the LPC head and only focused on mel-spectrum, which is degraded like the original Tacotron system.

3.2.4. Paragraph Modelling

To synthesize a paragraph with natural prosody using Tacotron, we can either concatenate sentences with manually add pause between them, or pass entire paragraph as input to the model. Concatenation typically leads to unnatural change of speech volume and intonation at sentence boundaries. We mainly explored the latter probability.

Tacotron is typically trained and tested on short sentences, since RNNs are trained using BPTT(back propagation through time). This limits the maximum length of input sequence during training. We find Tacotron model trained with batched BPTT fail to generalize to longer input, a paragraph for example. A context-sensitive-chunk BPTT(CSC-BPTT) approach to training RNN is adopted to better approximate training on minute long recordings. CSC-BPTT [14] is also used to prevent the model to learn from the absolute position in the input sequence, and to reduce influence of the initial RNN state. CSC-BPTT originally uses a fixed chunk length and context size, which requires force-alignment and chunking training data into fixed lengths. It is relatively complex to implement, as the encoder and the decoder need to be padded differently. In Tacotron model, encoder mainly contains a group of CNN and a bidirectional RNN, and the decoder contains layers of unidirectional RNN. We devised a method to approximate CSC-BPTT, which is easy to implement, by simply do not compute loss on the first and last few frames of decoder output. This is equivalent to treating first few steps of decoder RNN as context. Since the attention is approximately monotonic, the left and right few phoneme embeddings are also treated as context. When using forward attention, this is true even during the first thousands steps during training. In addition, we increase the length of input audio clips to twice as long. In our experiments, masking the first and last 20 frames is enough for the model to generalize to minute long speech continuously.

3.2.5. Environment Modelling

The recordings provided contains one minute long recordings with various noise conditions, including noise introduced by recording devices and environments. Vanilla multi-speaker Tacotron can generate random noise condition in generated speech. When synthesizing long paragraph tens of seconds long, drift of noise condition and gradual deterioration of speech quality becomes noticeable. The accumulation of error during auto-regressive synthesis makes it difficult for decoder RNNs to maintain noise condition throughout long sentences. This is reasonable, as ground-truth information used during training is not available during inference.

A simple technique of adding environment embedding is devised to counteract the drift effect. Following the idea of [15], conditioning the model on noise condition can control the auto-regressive generation process. Since the dataset consists of one minute recordings with a variety of recording environments, there is little correlation between speaker identity and noise conditions. We condition the decoder on environment and speaker embedding. All segments from the same recording shares the same speaker ID and environment ID. The t-SNE visualization [16] of the latent encoding is given in Fig. 7. Latent encodings close to each other tend to generate same noise con-

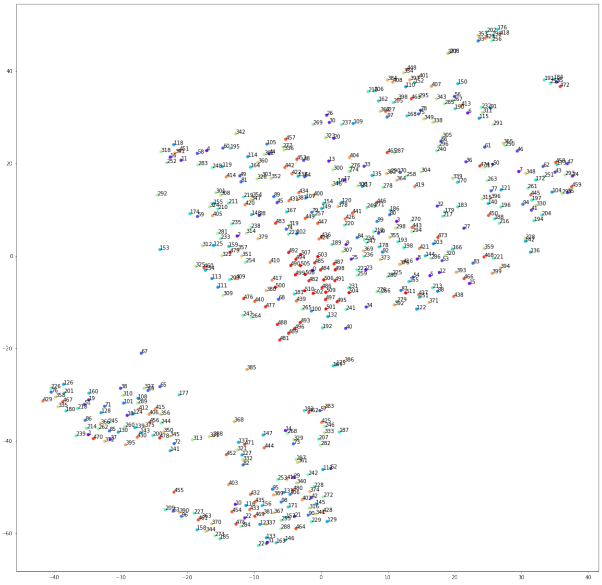


Figure 7: *t-SNE visualization of learnt environment embedding on 480 paragraphs*

ditions. This method stabilizes synthesizing even for thousands of decoder steps using CSC-BPTT trained Tacotron. In the submission, the environment embedding is set to "clean speech".

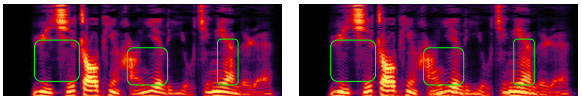
Futher more, an additional VAE (in particular a VAE with Gaussian mixture prior distribution [17]) can be used to model prosody without supervision. When environment embedding is not used, most of the latent space capacity is observed to model recording environments. When environment embedding is used, the latent variable almost have no influence on the noise condition and only influences prosody. A hierarchy of conditioning variables can help modeling different aspects of speech.

3.2.6. Discriminator on Mel-spectrogram

To bridge the gap between predicted acoustic features and real acoustic features, we attempted to apply generative adversarial networks (GAN) to our spectrogram model. We tried directly adding a discriminator to the spectrogram model or an explicit GAN based post-filter. When directly adding a discriminator, it caused a disaster that the WaveNet vocoder got worse after being adapted by the teaching-forced prediction on the training set. This is caused by the fact that: the discriminator does make the predicted mel-spectrograms more natural, but for the same audio, the "predicted natural mel-spectrograms" varies too much from the "extracted natural mel-spectrogram". This makes a large mismatch between the teaching-forced spectrogram and natural waveform. Therefore it is not proper to adapt the WaveNet vocoder with the predicted mel-spectrogram under the supervision of a discriminator. To reduce the influence of the discriminator, we attempted to add an GAN-based postfilter network. As shown in Fig. 8, the post-filter makes the blurred spectrograms more clear. However, this benifit vanished after vocoding by the WaveNet vocoder. Therefore, we removed the GAN module from the system.

3.3. Bandwidth Extension

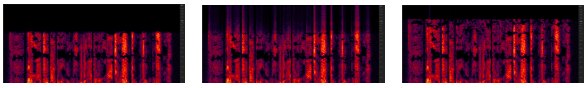
The bandwidth extension model is trained on multi-speaker dataset without speaker embedding, and directly applied to the



(a) Teaching force (b) Postfilter

Figure 8: *Comparison between spectrogram with and without the postfilter*

audio of `LZY`. Fig. 9 shows the spectrograms of the 16kHz natural speech, 22kHz natural speech and bandwidth extended 32kHz speech. We listened to the 32kHz audio and found that: the extended bandwidth has positive effect to speech quality. The effect retains on the synthetic speech from our system. We observed that the narrow-band of the predicted speech is not always as good as the original speech. So we only keep the high-frequency band, and concatenates it with the input narrow-band spectrogram. We also explored the effects of different μ-law bits in the BWE model, but it doesn't make a big difference. So we take 8-bit μ-law quantization in our final submission.



(a) Input (16KHz) (b) BWE (32KHz) (c) Target (22KHz)

Figure 9: *Example of multi-speaker Bandwidth Extension.*

3.4. Performance analysis

Since we didn't make a good WaveNet vocoder for the predicted spectrogram, we need to make a choice on what vocoder to use in our final submission. We examined the performance of Griffin-Lim and the WaveNet and found that: the audio from GL has less noise but sound like machine, the audio from WaveNet has more noise but sound like human. To take the advantage of both of them, we attempted to mix the spectrogram from GL and WaveNet that: the lower bands come from GL but the higher bands come from WaveNet (We call it mixed vocoder). To evaluate the performance[4], we presented the audios from WaveNet vocoder and mixed vocoder to about 40 native listeners. All of them observed that the audios are significantly different, and most of them had a very strong preference. Unfortunately, the preference are not same. Most listeners older than 30 (e,g, teachers, bosses) preferred the WaveNet vocoder, but most listeners younger than 30 (e.g. students) preferred the mixed vocoder. This observation maybe not valuable if the high quality vocoder is available, but it suggests that the people in different ages pay more attention to different part of the same speech.

4. Conclusion

The vocoder is the key point to produce high quality synthetic speech. The vocoder should be robust to deal with the mismatch between natural and predicted acoustic features. The proposed spectrogram model can produce better prosody than the typical Tacotron model on the provided corpus. Bandwidth extension is an useful technique that the corrupted higher-band are not necessarily to be modeled.

[4]This is not a formal questionnaire, we directly asked the people face-to-face, therefore we cannot provide an evaluation figure.

5. References

[1] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio *et al.*, "Tacotron: Towards end-to-end speech synthesis," *arXiv preprint arXiv:1703.10135*, 2017.

[2] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan *et al.*, "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4779–4783.

[3] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[4] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlcek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The kaldi speech recognition toolkit," *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*, 2011.

[5] P. Taylor, A. W. Black, and R. Caley, "The architecture of the festival speech synthesis system," in *The Third ESCA/COCOSDA Workshop (ETRW) on Speech Synthesis*, 1998.

[6] H. Li, Y. Kang, and Z. Wang, "Emphasis: An emotional phoneme-based acoustic model for speech synthesis system," *arXiv preprint arXiv:1806.09276*, 2018.

[7] Z.-H. Ling, Y. Ai, Y. Gu, and L.-R. Dai, "Waveform modeling and generation using hierarchical recurrent neural networks for speech bandwidth extension," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 5, pp. 883–894, 2018.

[8] J.-M. Valin and J. Skoglund, "Lpcnet: Improving neural speech synthesis through linear prediction," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5891–5895.

[9] D. Griffin and J. 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.

[10] M. Morise, F. Yokomori, and K. Ozawa, "World: a vocoder-based high-quality speech synthesis system for real-time applications," *IEICE TRANSACTIONS on Information and Systems*, vol. 99, no. 7, pp. 1877–1884, 2016.

[11] J.-X. Zhang, Z.-H. Ling, and L.-R. Dai, "Forward attention in sequence-to-sequence acoustic modeling for speech synthesis," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4789–4793.

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

[13] W. Che, Z. Li, and T. Liu, "Ltp: A chinese language technology platform." *In Proceedings of the Coling 2010:Demonstrations.*, pp. 13–16, 2010.

[14] K. Chen, Z.-J. Yan, and Q. Huo, "A context-sensitive-chunk bptt approach to training deep lstm/blstm recurrent neural networks for offline handwriting recognition," *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, pp. 411–415, 2015.

[15] W.-N. Hsu, Y. Zhang, R. J. Weiss, Y.-A. Chung, Y. Wang, Y. Wu, and J. R. Glass, "Disentangling correlated speaker and noise for speech synthesis via data augmentation and adversarial factorization," *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5901–5905, 2019.

[16] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of machine learning research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[17] W.-N. Hsu, Y. L. Zhang, R. J. Weiss, H. Zen, Y. Wu, Y. Wang, Y. Cao, Y. Jia, Z. Chen, J. Shen, P. Nguyen, and R. Pang, "Hierarchical generative modeling for controllable speech synthesis," *ArXiv*, vol. abs/1810.07217, 2018.