IBC2022

AUTOMATIC COMMENTARY TRAINING TEXT- TO- SPEECH SYSTEM FOR BROADCASTING

Kiyoshi Kurihara and Mayumi Abe

NHK (Japan Broadcasting Corporation), Japan

ABSTRACT

Thanks to the rapid progress in deep learning technology, the text-to-speech (TTS) system we developed has achieved the same quality as human speech, enabling us to launch a fully automatic program production system known as “AI Anchor.” The TTS system needs a large amount of speech and label data, but data production costs are high and TTS speakers cannot be easily added. This paper presents a novel TTS method featuring automatic training from broadcast commentary. It uses an approach that allows for a new semi-supervised learning method using an accentual data recognition method specialized for TTS.

We have automated the entire training process for generating training data and performing label data recognition from broadcast commentary. In this paper, we present practical examples of automated program production such as an automatic weather forecast system for radio, automatic sports commentary system, and slow and easy-to-understand commentary news using our automated TTS training system based on broadcast commentary.

INTRODUCTION

Owing to recent progress in deep learning technology (1- 2), text-to-speech (TTS) technology has come to be widely used in smartphones, smart speakers, social networking service (SNS) videos, etc. We have been researching and developing this technology and have been putting it to practical use in broadcast programs and video distribution such as “AI News Anchor (3)” since 2018. However, TTS requires studio-quality speech training data, but this increases costs. The work of generating training data requires anchors, sound engineers, directors, and studio resources as well as annotation of the phoneme label files, which is the most costly. However, given that a broadcast station airs high-quality speech on a daily basis, studio-quality speech is easy to acquire. With this in mind, we developed a technique that treats broadcast commentary as training data for TTS.

The work presented here focuses on semi-supervised learning (SSL) TTS, which uses state-of-the-art speech recognition and broadcasting commentary data. In related research, the system of Chung (4) considered single-speaker and unpaired noisy data for SSL TTS, but the system did not use speech recognition, multi-speaker learning, and sequential audio stream like broadcasting commentary. Thus, this method could not be used for broadcasting commentary. In addition, the system of Tu (5) considered multi-speaker learning and speaker identification for SSL TTS, but it did not train sequential audio streams like broadcasting commentary, its architecture was outdated, and its evaluation results were worse than those of the latest TTS. The proposed method automatically generates data from audio streams like broadcasting and adopts a state-of-the-art TTS model. Also, many TTS methods do not support sequential audio stream and only support waveforms one sentence long. Moreover, the proposed method is versatile for many of the latest TTS methods.

In this paper, we introduce a method that automatically extracts speech from broadcast commentary through a combination of deep learning methods and automatically generates training data for TTS by using the latest speech-recognition method. We also introduce practical examples of applying TTS for broadcasting.

TEXT-TO-SPEECH SYNTHESIS

The appearance of a waveform generation method called WaveNet (2) in 2016 and that of the sequence-to-sequence with attention method (seq2seq) (6) (7) of TTS in 2017 brought the quality of TTS to a level equivalent to that of human speech. In addition, the training data became simpler and less expensive to produce.

We proposed a novel TTS method for pitch-accent languages (8) in 2018. This method can train speech in pitch-accent languages by using phonemes and prosodic features (PPF). We developed and implemented an interface to this latest TTS system to make it easy for anyone to use. Furthermore, to enable speech contents to be automatically generated, we installed a Web API (Application Programming Interface) that could be linked to other systems so that our TTS system (3) could be used in automatic speech-content generation techniques.

Overview of Text-to-Speech

TTS can be broadly divided into two sections (Figure 1). The first is “text analysis” and the second is “waveform synthesis.” The text analysis section estimates the kind of sequences made from text and is therefore the section that determines pronunciation. It mainly uses natural language processing (NLP) technology for this purpose. The waveform synthesis section, on the other hand, estimates waveforms by inputting PPF and affects sound quality as a result. This section has undergone significant improvements in performance due to contributions from deep learning.

Text Analysis

The text analysis section consists of a technique that converts graphemes of text spelled out into phonemes, which is why it is called a grapheme-to-phoneme (g2p) method (9). If not done well, this estimation of phonemes will negatively affect pronunciation, resulting in strange intonation. If the intonation of synthesized speech is felt to be unnatural, the

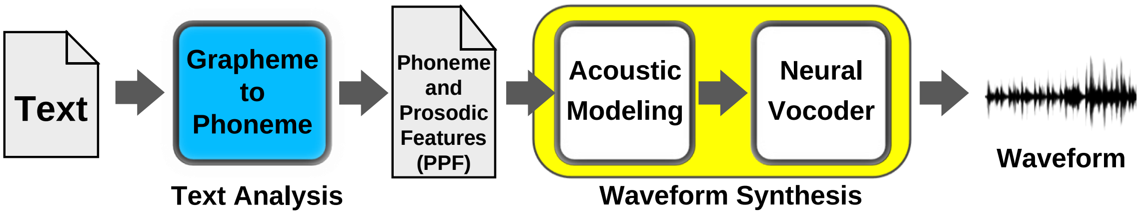

Figure 1 – Overview of TTS system

reason for this is thought to be the quality of g2p. The g2p method differs from one language to another. In the case of English, many words can be straightforwardly converted because many of them are one-to-one conversion in graphemes and phonemes, but because some words have multiple readings, using deep learning methods has been proposed (9). We proposed a novel g2p method for acoustic modeling of pitch-accent language (8).

Waveform Synthesis

The waveform synthesis section takes PPF estimated from the text analysis section and generates appropriate waveforms from those phonemes. This section affects sound quality. The waveform synthesis section is broadly divided into two methods. The first method estimates frequency components of speech using acoustic modeling. Here, it is common to use mel spectrograms—graphs of frequency waveforms—that are used in voiceprint analysis. The second method constitutes a waveform generation section that takes the information on those frequency components to generate a waveform as a signal along the time axis.

Acoustic modeling appeared as a method based on deep learning (10) in 2014, resulting in a significant improvement in the quality of TTS. This method adopted an approach that replaces the conventional hidden Markov model (HMM) TTS (11) with deep learning. Then, in 2017, a method appeared that abolished the HMM legacy and used the seq2seq method in a revised configuration. The appearance of this method significantly improved the naturalness of pronunciation. Additionally, a neural vocoder using a deep learning method called WaveNet (2) (12) appeared in 2016. The appearance of this method significantly improved the sound quality. When WaveNet appeared, the large amount of data generated because of sequential predictions from the beginning of the waveform signal meant that much time was normally needed to generate that waveform signal. More recently, it has become common to estimate the speech waveform entirely from the noise signal at one time by shifting from a method that makes sequential predictions about discrete waveform signals to a deep learning method called a generative adversarial network (GAN) (13).

AUTOMATIC TRANING DATA GENERATION METHOD

We developed a method for automatically generating TTS labels. This system is outlined in Figure 2. The system can learn TTS acoustic modeling using the PPF generated by an automatic label generator and waveform data.

This method consists of preprocessing and a method for estimating PPF. The

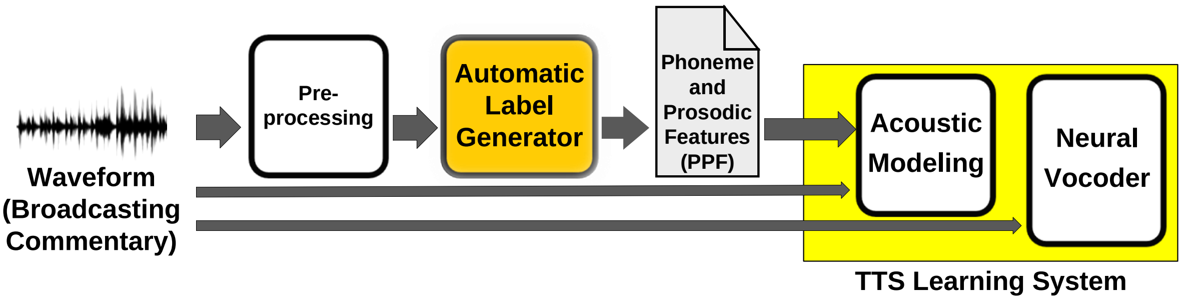
Waveform (Broadcasting Commentary)
Pre-processing
Automatic Label Generator
Phoneme and Prosodic Features (PPF)
Acoustic Modeling
Neural Vocoder
TTS Learning System

Figure 2 – Overview of automatic training data generation method

preprocessing section automatically selects the desired speech from the audio stream and splits the speech into separate sentences. Then, for each sentence of speech prepared in this way, the system automatically estimates labels for TTS using an original speech-recognition method that we developed. This method applies Wav2vec 2.0 (14), a speech-recognition technology based on self-supervised learning (15). Applying self-supervised learning in this way enabled us to successfully estimate labels for speech synthesis purposes with a relatively small amount of data compared with other speech-recognition methods. Using this method to automatically generate data and to learn from that data to synthesize speech makes it possible to construct a system that can automatically train a TTS model from broadcasts.

Preprocessing

Automatic PPF recognition requires that speech be converted into one sentence that can be learned in TTS. Since training data for TTS consists of one-sentence units, speech from the broadcast audio stream must be automatically split into sentences. It is also necessary to remove any noise like sound effects from speech and recognize speakers targeted for learning purposes and to then automatically create training data in units of sentences. As shown in Figure 3, preprocessing performs voice active detection (VAD) (16), noisy speech classification, and speaker verification in that order.

Voice active detection

The purpose of this process is to split sequential speech into sentences, so that TTS can train one sentence. VAD is a process that calculates speech power and splits that speech whenever its power falls below a specific threshold value. This process enables continuous speech to be automatically split into sentence-by-sentence files.

Noisy speech classification

Noisy speech classification is a process that classifies broadcasting commentary in which noise is included or not in the audio stream. To prevent degradation in TTS sound quality, this process simply removes any speech that includes noise without using noise removal technology. Specifically, the process identifies a waveform that includes noise by displaying speech in a graph that visualizes frequency characteristics called a mel spectrogram and subjecting that graph to image recognition processing (17) using deep learning. We trained image recognition to learn a mel spectrogram with and without noise and implemented a recognizer that can classify speech that does not contain noise. Applying image recognition can determine whether noise or sound effects are included so that the speech can be classified accordingly.

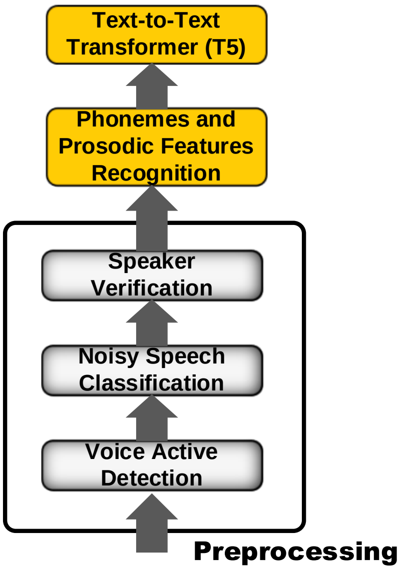
Text-to-Text Transformer (T5)
Phonemes and Prosodic Features Recognition
Speaker Verification
Noisy Speech Classification
Voice Active Detection
Preprocessing

Figure 3 – Procedure of automatic label generation

Speaker verification

The speaker verification system (18) recognizes acoustic features of speech by using a deep learning technique used in image recognition and other areas. In this development, the system learned how to identify the speech of a specific speaker appearing in programs and extracted only the speech of speakers to be used in TTS. If the system were to be set so as to learn the speech of an indefinite number of speakers, there would be cases of learning the speech of unintended speakers, and it is for this reason and speakers' licensing considerations as well that we installed this speaker recognition method.

Phonemes and Prosodic Feature Recognition

PPF recognition is used to recognize phoneme and accentual feature sequences for TTS. Waveform and phonemes or PPF for TTS often form pairs of data peculiar to a language, and when collecting data for use in speech recognition, the amount of data that can be practically collected is often small. For this reason, we focused on the Wav2vec 2.0 (14) method developed by Baevski that can learn even with a small amount of data using self-supervised learning. We found in experiments that this system could recognize pitch-accent by applying it to the recognition of PPF. The amount of training data that can be used for TTS purposes is only over ten hours in general, which falls short of the several hundreds of hours of data needed for PPF recognition. However, this self-supervised learning method could be used to generate a recognition system with just five hours of data, which enabled us to reduce the amount of labor expended in subsequent annotation work. With this method, it also became possible to use the data obtained from actual broadcasts for learning.

With deep learning, a recognition system can be constructed with a small amount of data by using a large-scale learning dataset. This time, to enhance recognition accuracy with even a small amount of data, we used a pre-trained model of Wav2vec 2.0 released as public domain (19). This was trained by using 56,000 hours of speech data covering 53 languages. As a broadcaster, we possess a massive amount of high-quality data excelling in both pronunciation and sound quality, so once we have constructed a system for PPF recognition with a small amount of data, we will be able to produce data for use in automatic learning TTS.

Additionally, recognition results generally include errors. To enable character strings with phoneme errors to be corrected, we tried improving recognition accuracy by using Text-to-Text Transformer (T5) (20). Figure 4 shows the architecture of a combination of Wav2vec 2.0 and T5, and this part works as "Automatic Label Generation (Training)" in Figure 5. Specifically, we configured a transformer for correcting phoneme errors by arranging and learning estimated character sequences from Wav2vec 2.0 in the transformer's input section and PPF that have already been manually corrected as ground truth in the output section. Moreover, to extend the T5 training data, a phonetic and accentual error generator was created in a data augmentation process by deleting characters at random, switching consonants. When synthesizing speech, the PPF estimated by g2p is input into the TTS.

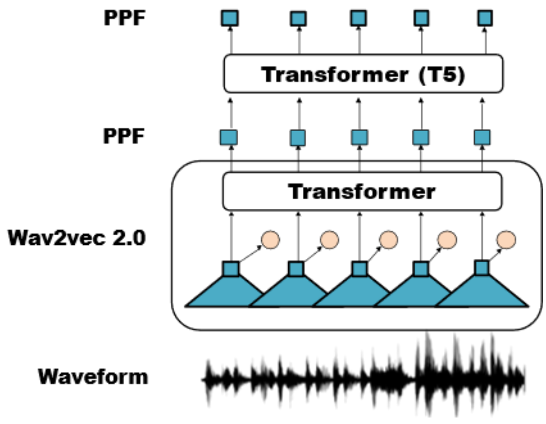

Figure 4 – Architecture of combination of Wav2vec 2.0 and T5

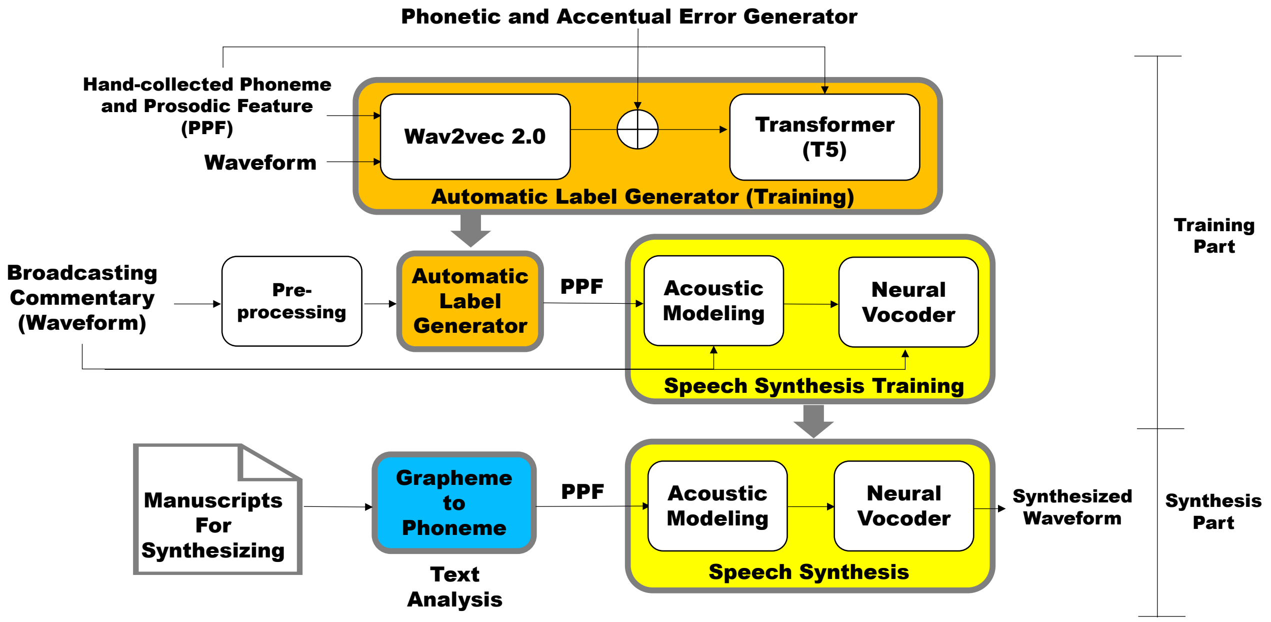
Phonetic and Accentual Error Generator
Hand-collected Phoneme and Prosodic Feature (PPF)
Waveform
Wav2vec 2.0
Transformer (T5)
Automatic Label Generator (Training)
Broadcasting Commentary (Waveform)
Pre-processing
Automatic Label Generator
PPF
Acoustic Modeling
Neural Vocoder
Speech Synthesis Training
Training Part
Manuscripts For Synthesizing
Grapheme to Phoneme
Text Analysis
PPF
Acoustic Modeling
Neural Vocoder
Speech Synthesis
Synthesized Waveform
Synthesis Part

Figure 5 – Hybrid system of PPF recognition and speech synthesis

Automatic Speech-Synthesis Training

Automatic generated labels are used as training data for TTS (Figure 5). We use the pre-trained Wav2vec 2.0 and T5 model as automatic label generation. Broadcasting commentary goes through preprocessing and automatic label generation part. This process automatically produced PPF, which is the training data for TTS. The automatically generated PPF can be used as training data for TTS. These whole processes achieve semi-supervised learning.

In general, deep learning improves the quality of a model by re-training the model using the initial weights of the previously trained model. The model used as the initial weights is called a pre-trained model, and re-training is called fine-tuning. Training data for pre-training purposes has the property that overall quality is improved by using as much training data as possible from a variety of speakers as the amount of data increases. On the other hand, fine-tuning can improve quality with only a single speaker. In PPF recognition, a certain amount of errors occurs, but using this data in pre-training data will prevent such errors from affecting generated data. It is also known that quality improves even with a small amount of fine-tuning data as the amount of training data for pre-training increases. An example of fine-tuning is shown in Figure 6. When generating a TTS model for Speaker E, a high-quality model can be obtained by learning the data automatically generated with other speakers as a pre-training model and then training Speaker E using that weight as an initial value.

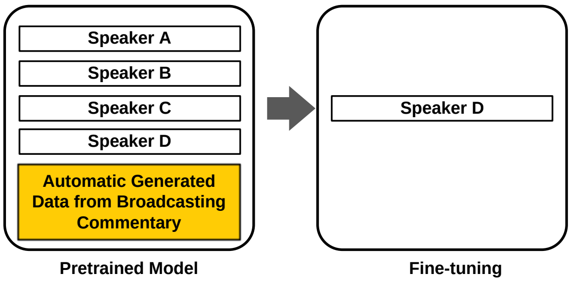

Figure 6 – Specific speaker modelling method which using fine-tuning

EVALUATION EXPERIMENT

To evaluate the effectiveness of the proposed method, we performed an evaluation experiment on the accuracy of recognizing PPF. We used the recorded speech of in-house anchors in a studio booth and manually corrected PPF in speech datasets for fine-tuning the pre-trained model of Wav2vec 2.0.

Experimental Conditions for Phoneme and Prosodic Feature Recognition

We prepared a dataset of four males and a dataset of 3 females. The manuscript of dataset consisted of news, weather forecast, and lifestyle information. The sampling frequency was 16 kHz, and the bit rate was 16 bits. In addition, the text data for pre-training of T5 was used for automatic generated PPF, which our proposed g2p (8) generated from 631,014 sentences of news script obtained from NHK NEWS WEB (21) in the period from April 2018 to April 2021. For a Wav2vec 2.0 pre-trained model, we used XLSR-Wav2Vec2 (19), which contains approximately 56,000 hours of speech data for 53 languages as data for pre-training. We performed fine-tuning against this pre-trained model using speech and manually corrected Japanese kana characters and prosodic symbols (Japanese PPF) and then performed model training. Additionally, we trained the T5 used for correcting PPF errors by automatically creating characters for the 631,014 sentences of news copy using our proposed g2p (8) tool. The following data augmentation of T5 processing was performed against the above data to create training data.

- PPF was deleted at a rate of 5% or less.
- Consonants of PPF were substituted at a rate of 10% or less

We fine-tuned the pre-trained model generated with data used for pre-training as described above by using a training set consisting of 23,024 sentences of manually corrected PPF. In conformance with the properties of PPF, prosodic symbols consisted of initial rising, accent nucleus, accentual phrase boundary, pause, and end of sentence, which related accentual and pause information. PPF estimation by Wav2vec 2.0 was taken to be proposed method 1 (Prop. 1) and correcting the PPF estimated by Wav2vec 2.0 by T5 for correcting phoneme errors was taken to be proposed method 2 (Prop. 2). Additionally, since the amount of training data in the dataset for TTS was insufficient to use the speech-recognition technique for comparison purposes, we decided for this experiment to convert speech into PPF by using a pre-trained model for Japanese speech recognition released by Espnet ASR (22) that uses seq2seq speech-recognition method. We also used the process of automatically converting that speech into PPF using our proposed g2p (8) as a conventional method (Conv.) for comparison.

Experimental Results

Experiment 1

We used the speech in the dataset of one male and that if one female (2541 sentences, 5.69 hours) to fine-tune Wav2vec 2.0 in proposed methods Prop.1 and Prop. 2. We also used character strings of manually corrected PPF (23,024 sentences) to fine-tune T5 for correcting phoneme errors in proposed method Prop. 2. As for the test dataset, we used

	Method	CER %
Conv.	Espnet ASR + Open JTalk	22.6
Prop. 1	Wav2vec 2.0	8.5
Prop. 2	Wav2vec 2.0 + Transformer (T5)	4.7

Table 1 – Evaluation results of Wav2vec 2.0

Hour	CER %
1.0	9.3
2.5	8.3
5.0	7.5
10.0	7.7
20.0	7.8

Table 2 – Evaluation results of increased data volume

the dataset of two males and one female (1558 sentences, 3.73 hours). In the experiment, we calculated the Character Error Rate (CER) between the estimated labels obtained by each method and ground truth and compared results. Experimental results are listed in Table 1. Proposed methods Prop. 1 and Prop. 2 had lower values of CER than the conventional method. Moreover, comparing proposed methods Prop.1 and Prop. 2, the value for proposed method Prop. 2 that incorporated T5 for correcting phoneme errors was lower, which demonstrated the effectiveness of using T5.

Experiment 2

In this experiment, we observed a change in the amount of training data. A corpus of one male and one female was used as data for fine-tuning in Wav2vec 2.0. Transformer for error correction was not used here. Table 2 shows the evaluation results with varying amounts of data. From the experimental results, the highest performance was found with 5 hours of data.

Discussion

We demonstrated the effectiveness of the proposed method in Experiment 1. Through this experiment, we gained knowledge on how prosodic symbols could be estimated by using "Wav2vec 2.0 + Transformer (T5)" with high accuracy from only speech. Conventional methods cannot estimate PPF that reflects acoustic features. We found that the proposed method Proc. 2 could estimate high-quality PPF. According to Experiment 2, the evaluation value of the training data is constant at over 5 hours. We proved that a small amount of training data is sufficient to obtain good label estimation results.

DEVELOPED SYSTEM

We have begun to implement and operate a method for performing TTS training automatically from a broadcast audio stream. We implemented an "automatic learning management system" to control each learning part and completed the system to automatically learn from the broadcast. A block diagram of the system is shown in Figure 7. The system consists of a TV tuner for outputting speech resources and a server that performs overall speech processing. Speech output from the tuner was recorded via an audio interface. The speech recorded here was automatically labeled by using the method described in the section "Automatic Training Data Generation Method" and used for training TTS using the method described in the section "Waveform Synthesis."

Speech was recorded here using an audio interface, but since our plan was to record only specific programs, we made it possible to search and record only intended programs by referring to a program listing on the Internet and linking to specific audio channels on the

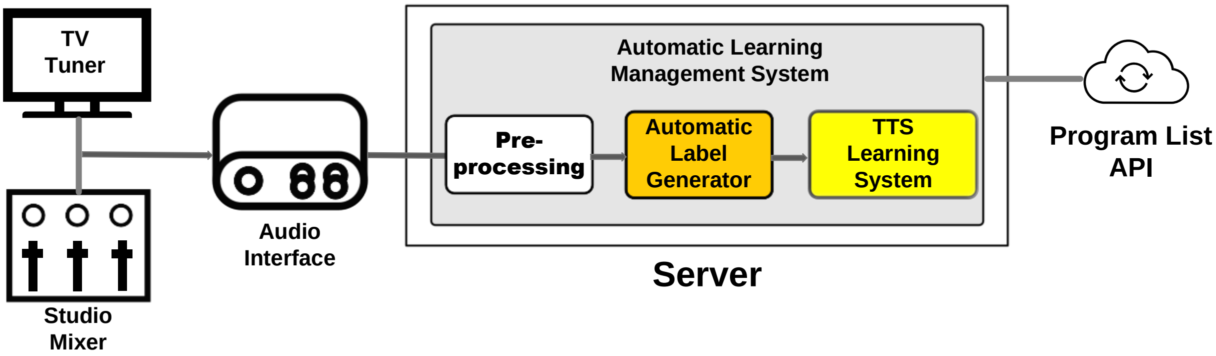

Figure 7 – The system of automatic training to broadcasting

audio interface. The program listing provides program start time and end time and the program name. For this system, we used “NHK Program List API (23)” as the program listing open Web API.

In this system, we recorded via a tuner, but depending on the type of operation, a microphone could also be directly connected to the studio mixer bus for recording even cleaner commentary. However, most broadcast stations have multiple studios that would make it difficult to capture clean commentary from many programs, so methods such as capturing audio signal from the master control or from a tuner as in our system should also be effective.

Practical Examples of Speech Synthesis for Broadcasting

We have made TTS practical as an automated system and have made it easy to use within a program in combination with a computer-graphics (CG) character named “AI News Anchor.” We have also implemented an automatic sports commentary system (24- 25) and automatic weather forecast system for radio (26) as examples of automatic audio program production. In the automatic sports commentary system, we developed a method that automatically generates a script from game data obtained from Olympic Broadcasting Services and deployed the method during the Olympic Games in Pyeongchang and Tokyo. We also achieved a method that automatically generates audio programs within a regional radio program by using the automatic weather forecast system for radio. We also launched manually operated services for Japanese-language learners on the website of “Weekly News in Simple Japanese” (3) (27), which uses slow and easy-to-understand commentary speech. These techniques have helped to make program production even easier.

In addition, a user interface that we developed has enabled TTS to be used frequently within general programs such as morning TV news shows, lifestyle information programs, and radio announcements. Speech of our TTS has also found use in YouTube videos (28), in-house information presentation systems, and other applications.

FUTURE PROJECTS

We have constructed original speech-synthesis and speech-recognition systems using PPF and have successfully connected the two. In the past, preparing data for TTS was costly due to the need for studio recordings, manual annotation work, etc., which prevented the construction of a large-scale learning model as achieved by speech-recognition technology. Now, however, a large-scale learning model based on TTS can be easily constructed by using a PPF recognition method, an automatic speech-synthesis training method, and material from past broadcasts. A broadcasting station is always producing speech material, so using this material for constructing a large-scale learning model for TTS makes sense. Large-scale learning models as in GPT-3 (29) in the field of NLP have come to exhibit behavior that seems to indicate that they can comprehend complex context. It has also become possible to construct a speech-recognition system by simply preparing a small amount of training data in Wav2vec 2.0. Although a large-scale learning model does not yet exist for TTS, constructing one for TTS as has been done in other systems as described above holds the possibility of achieving new functions. Going forward, we seek to construct a large-scale TTS learning model on the basis of broadcast commentary.

CONCLUSIONS

We proposed a semi-supervised learning method that automatically collects data from broadcasts and that automatically learns on the basis of a small amount of training data. The broadcast station generates a large amount of high-quality speech data every day. It would be beneficial to return this speech data to programs by recreating it through text-to-speech (TTS). The quality of TTS has improved dramatically in just the last few years to the point that the quality of synthesized speech has reached a human level. However, just because the quality of synthesized speech is close to that of humans does not mean that TTS will soon become a common fixture in broadcasting. TTS cannot be used if other factors such as the speaking style of speech or the character itself does not fit the target program. Furthermore, as of 2022, training data for minor languages has been difficult to collect, and the amount of data needed to achieve TTS in those languages may never be collected. In such a situation, obtaining training data for TTS from broadcast commentary would be beneficial and would no doubt help to broaden the application range of TTS in a variety of local languages.

Our proposed method achieves TTS through semi-supervised learning and can construct a TTS system from a large amount of unlabeled speech such as broadcast materials. For example, training data even for the case of radio broadcasts can be automatically collected on an ongoing basis. Considering that the modern broadcast station is expected to disseminate information instantly, TTS in its capacity to reduce labor and increase the means of transmitting speech looks set to become an essential technology. It would give us great pleasure if this method could contribute to the expanded use of TTS in broadcasting and to the effective use of broadcast materials.

REFERENCES

1. Wang, Y., *et al.* 2017. Tacotron: Towards end-to-end speech synthesis. Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH). pp. 4006 to 4010.

2. Oord, A. v. d., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., & Kavukcuoglu, K. 2016. WaveNet: A generative model for raw audio. arXiv preprint arXiv:1609.03499.

3. Kurihara, K., Seiyama, N., Kumano, T., Fukaya, T., Saito, K., & Suzuki, S. 2021. "AI news anchor" with deep learning-based speech synthesis. SMPTE Motion Imaging Journal, 130 (3). pp. 19 to 27.

4. Chung, Y. A., Wang, Y., Hsu, W. N., Zhang, Y., & Skerry-Ryan, R. J. 2019. Semi-supervised training for improving data efficiency in end-to-end speech synthesis. Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 6940 to 6944.

5. Tu, T., Chen, Y.-J., Liu, A. H., & Lee, H.-y. 2020. Semi-supervised learning for multi-speaker text-to-speech synthesis using discrete speech representation. arXiv preprint arXiv:2005.08024.

6. Shen, J., *et al.* 2018. Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions. Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 4779 to 4783.

7. Sotelo, J., Mehri, S., Kumar, K., Santos, J. F., Kastner, K., Courville, A., & Bengio, Y. 2017. Char2wav: End-to-end speech synthesis. Proceedings of International Conference on Learning Representations (ICLR).

8. Kurihara, K., Seiyama, N., & Kumano, T. 2021. Prosodic features control by symbols as input of sequence-to-sequence acoustic modeling for neural TTS. IEICE Transactions on Information and Systems, E104-D (2). pp. 302 to 311.

9. Yolchuyeva, S., Németh, G., & Gyires-Tóth, B. 2019. Transformer based grapheme-to-phoneme conversion. Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH). pp. 2095 to 2099.

10. Zen, H., & Senior, A. 2014. Deep mixture density networks for acoustic modeling in statistical parametric speech synthesis. Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 3844 to 3848.

11. Zen, H., Tokuda, K., & Black, A. W. 2009. Statistical parametric speech synthesis. Speech Communication, 51 (11). pp. 1039 to 1064.

12. Hayashi, T., Tamamori, A., Kobayashi, K., Takeda, K., & Toda, T. 2017. An investigation of multi-speaker training for WaveNet vocoder. Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). pp. 712 to 718.

13. Yamamoto, R., Song, E., & Kim, J. M. 2020. Parallel wavegan: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram. Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 6199 to 6203.

14. Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. 2020. Wav2vec 2.0: A framework for self-supervised learning of speech representations. Advances in Neural Information Processing Systems, 33. pp. 12449 to 12460.

15. Zhai, X., Oliver, A., Kolesnikov, A., & Beyer, L. 2019. S4l: Self-supervised semi-supervised learning. Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1476 to 1485.

16. Shin, J. W., Chang, J.-H., & Kim, N. S. 2010. Voice activity detection based on statistical models and machine learning approaches. Computer Speech & Language, 24 (3). pp. 515 to 530.

17. Tan, M., & Le, Q. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. Proceedings of International Conference on Machine Learning (ICML). pp. 6105 to 6114.

18. Desplanques, B., Thienpondt, J., & Demuynck, K. 2020. ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification. Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH). pp. 3830 to 3834.

19. Xu, Q., Baevski, A., & Auli, M. 2021. Simple and effective zero-shot cross-lingual phoneme recognition. arXiv preprint arXiv:2109.11680.

20. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research, 21 (140). pp. 1 to 67.

21. NHK. 2022. News Web (in Japanese). https://www3.nhk.or.jp/news.

22. Watanabe, S., Hori, T., Karita, S., Hayashi, T., Nishitoba, J., Unno, Y., Soplin, N. E. Y., Heymann, J., Wiesner, M., & Chen, N. 2018. Espnet: End-to-end speech processing toolkit. https://github.com/espnet/espnet.

23. NHK. 2014. NHK program list API (in Japanese). https://api-portal.nhk.or.jp.

24. Kurihara, K., *et al.* 2019. Automatic generation of audio descriptions for sports programs. SMPTE Motion Imaging Journal, 128 (1). pp. 41 to 47.

25. Kumano, T., Ichiki, M., Kurihara, K., Kaneko, H., Komori, T., Shimizu, T., Seiyama, N., Imai, A., Sumiyoshi, H., & Takagi, T. 2019. Generation of automated sports commentary from live sports data. Proceedings of IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB). pp. 1 to 4.

26. NHK. 2019. Automated production of weather information radio programs. NHK STRL Open House. https://www.nhk.or.jp/strl/open2019/tenji/pdf/17_e.pdf.

27. NHK. 2022. Weekly news in simple Japanese. https://www3.nhk.or.jp/nhkworld/en/learnjapanese/audionews.

28. NHK. 2021. "Mewe" 2- minute instant SDGs (in Japanese). https://www.youtube.com/watch?v=KoZjtC7Ud0A&list=PLcynJ47QaWNuhl-BXjAUoyF7G7GGqQ_d7&index=3.

29. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., & Askell, A. 2020. Language models are few-shot learners. Advances in Neural Information Processing Systems, 33. pp. 1877 to 1901.

ACKNOWLEDGEMENTS

The authors are grateful to N. Seiyama at NHK Engineering System, Inc. for suggesting the topic discussed in this paper.