Data Augmentation Based on Frequency Warping for Recognition of Cleft Palate Speech

Kento Fujiwara*, Ryoichi Takashima*, Chihiro Sugiyama†, Nobukazu Tanaka†,
Kanji Nohara†, Kazunori Nozaki†, Tetsuya Takiguchi*
* Graduate School of System Informatics, Kobe University, Kobe, Japan
E-mail: kfujiwara@stu.kobe-u.ac.jp, rtakashima@port.kobe-u.ac.jp,
takigu@kobe-u.ac.jp
† Graduate School of Dentistry, Osaka University, Osaka, Japan
E-mail: {sugiyama, n-tanaka, nohara, knozaki}@dent.osaka-u.ac.jp

***Abstract*—In this paper, we present an automatic speech recognition (ASR) system for the speech of a person with a cleft lip and palate (CLP). The accuracy of speech recognition for a person with CLP is lower than that of a physically-unimpaired (PU) person because the CLP speech has characteristics that differ from those of a PU person; moreover, the amount of available training data is quite limited. In the field of ASR for PU people, data augmentation and self-supervised learning have been studied to tackle this problem of data scarcity. In this paper, we evaluate the effectiveness of those approaches on CLP speech recognition, and propose a data augmentation technique based on frequency warping. The formant of CLP speech tends to fluctuate compared to that of PU people. In order to compensate for the large variety of formant components, our data augmentation method stretches or contracts the spectrogram through the frequency axis. The experimental results on an ASR task with two CLP subjects showed that both data augmentation and self-supervised learning were effective for CLP speech recognition, and our proposed method further improved the performance of those two approaches based on conventional SpecAugment techniques.**

Index Terms: speech recognition, data augmentation, self-supervised learning, cleft lip and palate, dysarthria

I. Introduction

CLP is a congenital disorder that results in a cleft in the lip or palate. Because disorders of the oral cavity can adversely affect speech production, CLP can also cause dysarthria depending on the case. Therefore, the speech of people suffering from such dysarthria is often unintelligible. Figure 1 shows the spectrograms of speech uttered by a physically-unimpaired (PU) person (top) and a person with CLP (bottom). As shown in this figure, the energy of CLP speech is lower than that of a PU person in the high-frequency range, which is one of the factors worsening the intelligibility.

Recently, speech recognition systems have been widely used in various aspects of daily life, such as cell phones and smart speakers. However, since most speech recognition systems are designed for PU people, it is difficult for such existing systems to recognize CLP speech accurately. Therefore, there is a great need for an accurate ASR system for CLP speech. Such a system is expected to be used for various applications, such as communication support and articulation training.

There has been an attempt to construct an ASR system for CLP speech [1]. However, even when the system uses the CLP speech as training data, the recognition accuracy of CLP speech is generally lower than that of normal speech. The main problem that makes it difficult to build an ASR system for CLP speech is the lack of training data. In order to build an ASR system that is accurate enough for practical use, it is essential to prepare a sufficient amount of training data. There are two ways to collect the training data (speech and text label); 1) recording speech of speakers who read prepared scripts; and 2) recording speakers' spontaneous speech and

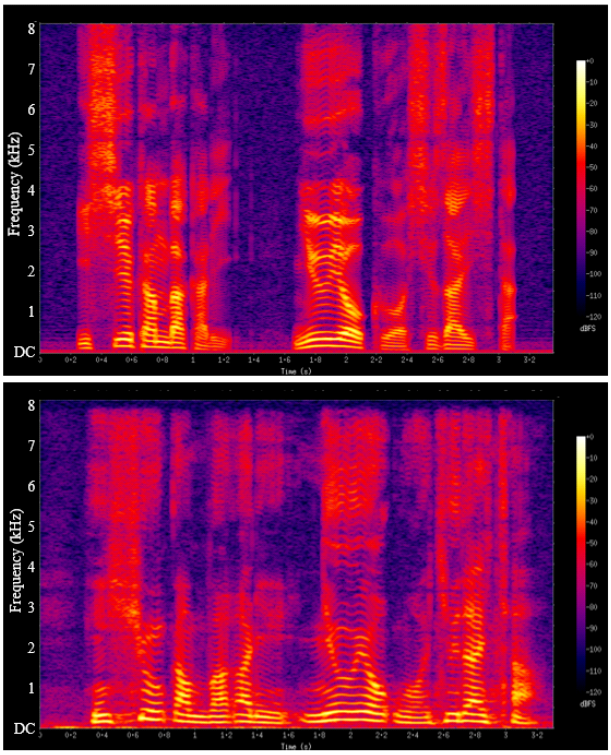

Fig. 1. Example of spectrogram uttered for a Japanese sentence /i q sh u: k a N b a k a r i ny u: y o: k u o sh u z a i sh i t a/ ("I covered the New York news for just one week") of a physically-unimpaired person (top) and a person with CLP (bottom).

transcribing it later. In the former way, the amount of data that can be collected is quite limited because reading scripts is a great burden for CLP people and it has a high cost. On the other hand, although collecting speech is relatively easy in the latter way, the transcription of CLP people's spontaneous speech is extremely difficult because of their low intelligibility. Consequently, the former way is employed in many datasets of speaking disorders [2][3] although the amount of collectable data is limited. Therefore, it is necessary to construct an ASR system using a smaller amount of training data than that collected from PU people.

One of the ways to increase the amount of data is through "data augmentation". Data augmentation is a technique that artificially increases the number of learning patterns by applying signal processing to existing training data [4][5][6][7][8]. Data augmentation is expected to contribute to the improvement of ASR performance, especially when the amount of training data is small. Another way to increase the amount of available data is leveraging unlabeled data through "self-supervised learning". This approach trains a neural-network model through a pseudo-task generated without human annotation. In this way, the model learns the representation of input features, and, therefore, this pre-trained model can be used as a good initial model for the target task [9][10][11][12][13].

Figure 2 depicts our proposed strategy for the training model for a CLP person. In this approach, we use both spontaneous speech and script-reading speech. The spontaneous speech, which is easy to be collected but difficult to be annotated, is used to pre-train the model without labels through self-supervised learning. Then, a small amount of script-reading speech, which does not need human annotation but difficult to be collected, is used to fine-tune the pre-trained model.

Although conventional methods of data augmentation and self-supervised learning have shown promising results in many ASR tasks for PU people, it has still been unclear if they are also effective for dysarthric speech recognition. In this paper, we evaluate the effectiveness of conventional methods of data augmentation and self-supervised learning based on the well-known SpecAugment [14][15] on CLP speech recognition. In addition, because conventional methods are proposed for PU speech, they are not necessarily optimal for CLP speech. Therefore, we propose a new method based on "frequency warping" that takes the nature of CLP speech into consideration. We confirm the effectiveness of the conventional data augmentation and self-supervised learning through experiments carried out on speaker-dependent CLP speech recognition tasks, and we also demonstrate that the performance can be further improved by combining the conventional method with our proposed method.

II. Related Works

We discuss in detail the conventional methods that have been proposed for PU speech recognition tasks and are used as our baseline. Park *et al.* [14] proposed SpecAugment, which performs data augmentation on log-mel spectrograms input to a neural network. SpecAugment uses three types of data

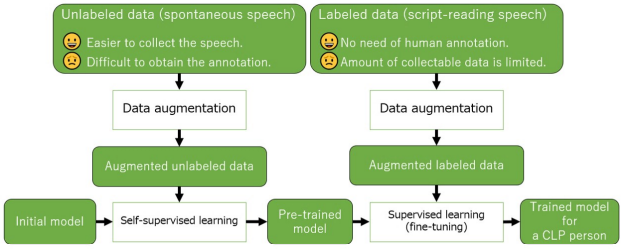

Fig. 2. Proposed training model strategy for a CLP person.

augmentation: 1) time masking, which masks the spectrogram through the time axis; 2) frequency masking, which masks the spectrogram through the frequency axis; and 3) time warping, which expands and contracts the spectrogram so that the reference point moves through the time axis. The size of the mask or time warping are set as parameters according to the dataset. Each augmentation corresponds to partial loss of speech segments, partial loss of frequency information, and deformation in the time direction. Since the generated training data is partially different from the original data, it is reported that SpecAugment can train an ASR model robustly against such information loss and deformation and can improve the recognition accuracy. In this study, we use SpecAugment as the baseline method of data augmentation.

Next, we discuss a method of self-supervised learning for speech recognition, which is used as the baseline method in this research. Wang *et al.* [15] proposed a method using data augmentation similar to SpecAugment. In this method, an encoder-decoder network is trained before training a speech recognition system. The encoder extracts hidden features from the input, and the decoder generates the output, which has the same size as the input, from the hidden features extracted by the encoder. During learning, the original input speech is transformed using the same process as SpecAugment (time warping, time masking, and frequency masking), after which the transformed data is input to the encoder. Then, the mean square error between the output of the decoder and the original input before performed time/frequency masking is evaluated as training a loss function. In other words, the encoder-decoder network is trained such that the network recovers the original input from the masked one. In order to recover the original data accurately, the encoder should extract high-quality representations of speech features from the input. Therefore, a robust encoder can be obtained without using human annotation, and it can be used as a good pre-trained ASR model. It is reported that the performance of speech recognition systems incorporating the encoder trained by this method can be improved. In addition, it is also reported that the performance of speech recognition systems can be further improved by increasing the amount of training data using SpecAugment.

TABLE I
STATISTICS OF F1 FREQUENCY [HZ].

Statistics [Hz]	PU1	PU2	PU3	SPK1	SPK2
Mean	420	417	410	437	450
Max - min	128	140	152	176	185
Standard deviation	31	30	39	44	48

III. METHODS

Since the conventional methods described in section II were proposed for PU people, they do not necessarily consider the characteristics of CLP speech. Therefore, it is expected that if we add a new transformation that reflects the CLP speech's characteristic to the conventional methods, the data augmentation and the self-supervised learning will be further effective for CLP speech recognition. In [16], it is reported that the difference between the dysarthric speech and the normal speech appears in their formant frequencies; that is peaks in the spectrum, and they proposed a method of assessment for dysarthric speech based on the formant frequencies.

With this in mind, we analyzed the first formant (F1) frequencies of the utterances for two subjects with CLP (SPK1 and SPK2) and three PU subjects (PU1, PU2, and PU3) in the ATR Japanese speech database [17]. (For details, see section IV.) Table I shows the statistics of the F1 frequency of each speaker. Comparing the F1 frequency statistics of CLP subjects and PU subjects, we see that the values of "Max - min" and the standard deviation of CLP subjects are higher than those of PU subjects. These results indicate that the CLP speech has a larger formant frequency variety than PU speech. However, since the amount of the collectable training data of CLP speech is limited as described in the Introduction, the amount of training data is not sufficient to train the variation of the formant frequency, and we hypothesize that it is a factor in the decreased accuracy associated with CLP speech recognition.

Therefore, we propose "frequency warping", which transforms the speech data through the frequency axis. Since frequency warping moves the formant frequency of the original speech data, it is expected that the data augmentation and self-supervised learning with frequency warping can help us to train a model to be robust against the variation of the formant frequency, which contributes to the improvement of the CLP speech recognition system.

Figure 3 shows the procedure of our proposed frequency warping. The procedure is similar to the time warping used in SpecAugment and performed on the mel-spectrogram. First, the reference mel-frequency bin and the reference bin shifting size are selected. Then, the reference frequency bin is shifted by expanding and contracting the spectrogram through the frequency axis. This process is performed in a randomly selected time segment in an utterance. The specific implementation is described as follows.

The input is a log-mel spectrogram with τ frames and ν frequency bins. First, the reference frequency bin f is selected randomly from the range of [W_{max}, ν], and the size w of shifting f is also selected randomly from the range of [W_{min}, W_{max}]. The W_{min} and W_{max} are set as hyperparameters. Next, we select the beginning frame t of the segment, in which the frequency warping is performed, randomly from the range of [0, $\tau - T_{max}$], and we select the length T of the segment randomly from the range of [T_{min}, T_{max}]. $t + T$ is the end frame of the segment. T_{min} and T_{max} are set as hyperparameters. Then, the selected segment is divided into the low-frequency region (size of $T \times f$, yellow window in Figure 3) and the high-frequency region (size of $T \times (\nu - f)$, green window in Figure 3). By contracting the low-frequency region and expanding the high-frequency region, the sizes of the low- and high-frequency regions are changed into $T \times (f - w)$ and $T \times (\nu - f + w)$, respectively. In this way, the reference frequency bin is shifted. The contracting and expanding processes are carried out using a PyTorch function "*Image.resize*", which handles the spectrogram as image data.

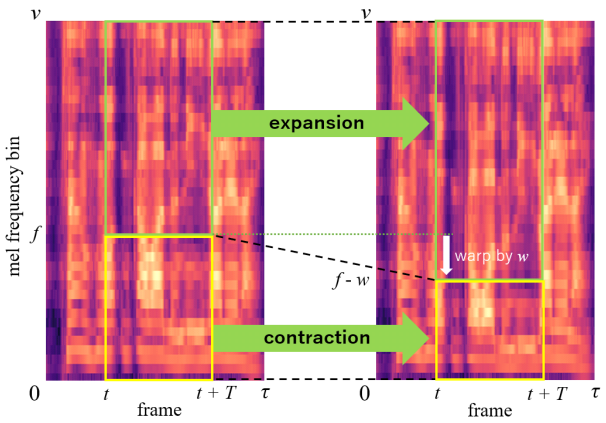

Fig. 3. Procedure of the frequency warping.

A. Frequency warping in self-supervised learning

In our proposed self-supervised learning, the frequency warping is added to the conventional feature transformations described in section II. Figure 4 shows the overview of the proposed self-supervised learning procedure. At first, time warping is applied to the original input in order to increase the training pattern. Next, frequency warping, time masking, and frequency masking are performed sequentially. Then, the mean absolute error (MAE) between the output of the decoder and the input before performing frequency warping and time/frequency masking is calculated as the loss function to train the encoder-decoder network. After proceeding with the self-supervised learning, the trained encoder is used as a feature extractor for inputting to the speech recognition model. Since the encoder is trained to reconstruct the original data from the data that has been subjected to feature transformation, including frequency warping, it is expected that the trained encoder can extract features robustly in spite of the variation in the formant frequencies.

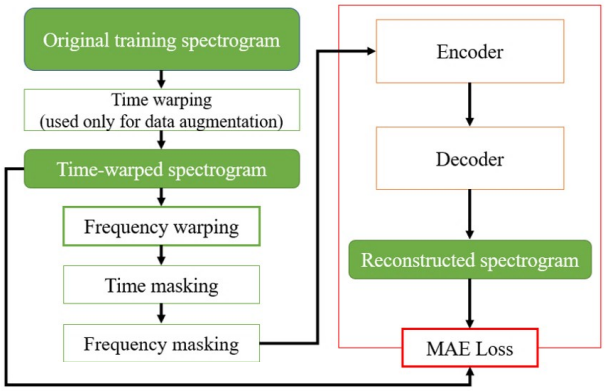

Fig. 4. Procedure of training an encoder-decoder network using the proposed self-supervised learning.

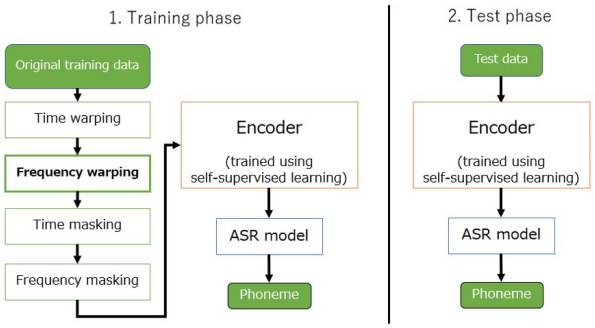

Fig. 5. Training and testing procedures of the ASR model.

B. Frequency warping for data augmentation to train the ASR model

Figure 5 shows the training and test procedures of the ASR system with the encoder trained using the self-supervised learning. During the training of the ASR model, frequency warping is also used with conventional SpecAugment for data augmentation. Here, frequency warping is expected to increase the variety of the formant frequency of the training data, which can train a robust ASR model against the variation of the formant frequency.

IV. Experiments

We evaluated the data augmentation and self-supervised learning through experiments on a speaker-dependent phoneme recognition task with two CLP males (SPK1/SPK2) who uttered 495 and 503 phonemically-balanced sentences from the ATR Japanese speech database [17], respectively. In the scenario of the experiments, it is assumed that only a portion of the training data has human annotation (i.e., a text label), and the remaining data does not have the label and is used for the self-supervised learning. For each subject, 200 sentences having text labels were used for evaluating the ASR model. Among these sentences, we used 100 sentences as training data for the ASR model, 50 sentences as development data, and 50 sentences as test data. The remaining data (295 or 303 sentences) that did not have text labels was used for self-supervised learning of the encoder.

TABLE II
Setting of the hyperparameters.

Parameter	Self-supervised learning	Training ASR model
T_w	0∼200	0∼200
F_w	0∼20	0∼20
T_d	−150∼150	−50∼10
W_{min}	0	0
W_{max}	10	2
T_{min}	τ	50
T_{max}	τ	100

The sampling frequency was 16 kHz. We used 40-dimensional log-mel filterbank features extracted with a frame shift of 10 ms and a window size of 25 ms. The ASR model consisted of two layers of bidirectional gated recurrent units [18] trained on the CTC loss function [19]. The output layer had 40 dimensions, consisting of 38 phonemes, the unknown symbol, and the blank symbol of CTC. An Adam optimizer [20] with an initial learning rate of 0.001 was used for optimization.

For the self-supervised learning, the encoder network consisted of four bidirectional LSTM layers [21], and the decoder network consisted of two linear layers with ReLU function. The self-supervised learning was also optimized using the Adam optimizer with the initial learning rate of 0.001.

Table II shows the settings of the hyperparameters for the self-supervised learning (described in section III-A) and the training of the ASR model (described in section III-B). T_w, F_w, and T_d are the hyperparameters of the SpecAugment. T_w and F_w correspond to the sizes of the time masking and frequency masking, respectively. T_d corresponds to the size of shifting the reference time for the time warping. The values of T_w, F_w and T_d were randomly selected from the range shown in the Table II. The values of the hyperparameters were experimentally confirmed to be the best accuracy.

A. Results without self-supervised learning

Table III shows the phoneme error rates (PERs) without self-supervised learning. In this experiments, the log-mel filterbank features were directly input into the ASR model without passing the encoder network. As a baseline, we show the results of training without any data augmentation. In the case using data augmentation, we compared the results when each augmentation was used alone, when three conventional methods were combined (Comb. of 3 types, that is equal to the original SpecAugment), and when the conventional methods and frequency warping were combined (SpecAugment + Frequency warping).

When data augmentation was used alone, all of the data augmentations that included frequency warping showed better

TABLE III
PERs [%] OF THE SPEAKER-DEPENDENT CLP MODEL WITHOUT SELF-SUPERVISED LEARNING.

Augmentation	SPK1	SPK2
Conventional method		
No augmentation	23.72	24.02
Time masking	22.83	23.13
Frequency masking	22.60	22.97
Time warping	20.78	21.85
Comb. of 3 types (SpecAugment)	19.85	19.95
Proposed method		
Frequency warping	22.37	22.78
SpecAugment + Frequency warping	19.43	19.60

performance than the conventional method without augmentation. This result indicates that frequency warping is an effective data augmentation method for CLP speech recognition. When three or four types of data augmentation were combined, the performance was further improved. These results indicate that the original SpecAugment and the proposed frequency warping are complementary to each other on the CLP speech recognition. One of the reasons for the high performance of time warping might be that the speech speed of CLP subjects fluctuates, along with formant frequency, and therefore, the time warping could train the ASR model to be robust against such fluctuations in the speech speed.

B. Results with self-supervised learning

Table IV shows PERs using the encoder network trained on the self-supervised learning. The self-supervised learning was performed in two different ways. One was the conventional method described in section II without frequency warping (Baseline), and the other the proposed method using frequency warping (Ours). Each method was evaluated on three different conditions of the ASR model training: using no data augmentation (None), using SpecAugment, and using SpecAugment with the frequency warping. The settings of the hyperparameters for the data augmentations are the same as in the experiments of Table III.

Comparing Baseline and Ours, Ours showed better performance. The difference of PERs between Baseline and Ours in Table IV is larger than the difference of PERs between SpecAugment and SpecAugment + Frequency warping in Table III. This suggests that frequency warping is particularly effective in self-supervised learning. It was also found that adding frequency warping to both data augmentation and self-supervised learning showed the best performance (17.96 for SPK1 and 17.85 for SPK2). When we use frequency warping for data augmentation, the proposed method can generate the data having formants that are not observed in the original training data. When we use frequency warping for the self-supervised learning, it can be considered that the proposed method can train an encoder that extracts hidden features and is invariant to the fluctuation of the formant frequency. Since the expected effects of frequency warping on the data augmentation and on the self-supervised learning are different, they might be complementary to each other.

TABLE IV
PER S[%] OF THE SPEAKER-DEPENDENT CLP MODEL WITH SELF-SUPERVISED LEARNING.

	SPK1		SPK2	
Augmentation	**Baseline**	**Ours**	**Baseline**	**Ours**
None	23.10	22.41	22.93	21.89
SpecAugment	19.39	18.65	19.71	19.05
SpecAugment + Frequency warping	19.16	17.96	19.44	17.85

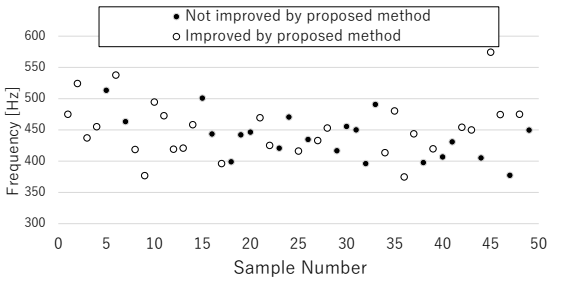

Fig. 6. F1 frequency of each test sample

Finally, Figure 6 shows the mean value of F1 frequency of each test sample of SPK1. The white and black dots show samples in which PER were improved / not improved by using the proposed frequency warping, respectively. As shown in this figure, the proposed method could improve the PER of samples having very high or low F1 frequency compared to the mean of the distribution. This indicates that the proposed method could train an ASR model that is robust against the variation of the formant frequency.

V. CONCLUSIONS

In this paper, we investigated methods of data augmentation and self-supervised learning for CLP speech recognition. We proposed frequency warping as a method that takes into account the characteristics of CLP speech; namely, speech having a large variation of formant frequency. The experimental results showed that, although the conventional method based on SpecAugment was effective for CLP speech recognition, frequency warping further improved the performance. However, the accuracy of speech recognition in this paper is still lower than that of PU people; therefore, more improvement is needed. In the future, we plan to increase the number of subjects and analyze other characteristics that can be used for the data augmentation and self-supervised learning.

ACKNOWLEDGMENTS

This work was supported in part by JSPS KAKENHI (Grant No. JP21H00906 and Grant No. 20K19862).

REFERENCES

[1] S. Maria, M. Andreas, H. Tino, N. Emeka, W. Ulrike, R. Frank, E. Ulrich, and N. Elmar, "Evaluation of speech intelligibility for children with cleft lip and palate by means of automatic speech recognition," *International Journal of Pediatric Otorhinolaryngology*, vol. 70, no. 10, pp. 1741–1747, 2006.

[2] H. Kim, M. Hasegawa-Johnson, A. Perlman, J. Gunderson, T. S. Huang, K. Watkin, and S. Frame, "Dysarthric speech database for universal access research," *Proc. of Interspeech*, pp. 1741–1744, 2008.
[3] F. Rudzicz, A. K. Namasivayam, and T. Wolff, "The TORGO database of acoustic and articulatory speech from speakers with dysarthria," *Language Resources and Evaluation*, pp. 1–19, 2011.
[4] N. Kanda, R. Takeda, and Y. Obuchi, "Elastic spectral distortion for low resource speech recognition with deep neural networks," *Proc. of IEEE Workshop on Automatic Speech Recognition and Understanding*, pp. 309–314, 2013.
[5] M. Fadaee, B. Arianna, and M. Christof, "Data augmentation for low-resource neural machine translation," *Proc. of the 55th Annual Meeting of the Association for Computational Linguistics*, vol. 2, pp. 567–573, 2017.
[6] N. Jaitly and G. Hinton, "Vocal Tract Length Perturbation (VTLP) improves speech recognition," *Proc. of ICML Workshop on Deep Learning for Audio, Speech and Language Processing*, 2013.
[7] K. Tom, P. Vijayaditya, P. Daniel, S. M. L., and K. Sanjeev, "A study on data augmentation of reverberant speech for robust speech recognition," *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 5220–5224, 2017.
[8] C. Xiaodong, G. Vaibhava, and K. Brian, "Data augmentation for deep neural network acoustic modeling," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 9, pp. 1469–1477, 2015.
[9] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," *Proc. of North American Chapter of the Association for Computational Linguistics*, pp. 2227–2237, 2018.
[10] K. Lagler, M. Schindelegger, J. Böhm, H. Krásná, and T. Nilsson, "GPT2: Empirical slant delay model for radio space geodetic techniques," *Geophysical Research Letters*, vol. 40, no. 6, pp. 1069–1073, 2013.
[11] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *arXiv*, 2019.
[12] A. Baevski, M. Auli, and A. Mohamed, "Effectiveness of self-supervised pre-training for speech recognition," in *arXiv*, 2019.
[13] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Proc. of NeurIPS*, pp. 12 449–12 460, 2020.
[14] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A simple data augmentation method for automatic speech recognition," *Proc. of Interspeech*, pp. 2613–2617, 2019.
[15] W. Wang, Q. Tang, and K. Livescu, "Unsupervised pre-training of bidirectional speech encoders via masked reconstruction," *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 6889–6893, 2020.
[16] S. Sapir, L. Ramig, J. Spielman, and C. Fox, "Formant centralization ratio: A proposal for a new acoustic measure of dysarthric speech," *Journal of Speech Language Hearing Research*, vol. 53, pp. 114–125, 2010.
[17] A. Kurematsu, "ATR japanese speech database as a tool of speech recognition and synthesis," *Speech Communication*, vol. 9, no. 4, pp. 357–363, 1990.
[18] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," *Proc. of Empirical Methods in Natural Language Processing*, pp. 1724–1734, 2014.
[19] A. Graves, S. Fernandez, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," *Proc. of the 23rd International Conference on Machine Learning*, pp. 369–376, 2006.
[20] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *Proc. of International Conerence on Learning Reprisentations*, 2017.
[21] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.