Real-Time Communication Aid System for Korean Dysarthric Speech

Kwanghyun Park and Jungpyo Hong *

Posted Date: 3 December 2024

doi: 10.20944/preprints202412.0289.v1

Keywords: dysarthria; communication aid system; conformer-based automatic speech recognition; JETS-based text-to-speech; Korean

 doi:10.20944/preprints202412.0289.v1

Article

Real-Time Communication Aid System for Korean Dysarthric Speech

Kwanghyun Park and Jungpyo Hong *

Department of Information and Communication Engineering, Changwon National University, Changwon 51140, Republic of Korea

* Correspondence: hansin@changwon.ac.kr

Abstract: Dysarthria is a speech disorder characterized by difficulties in articulation and vocalization due to impaired control of the articulatory system. Around 30% of individuals with speech disorders have dysarthria, facing significant communication challenges. Existing assistive tools for dysarthria either require additional manipulation or only provide word-level speech support, limiting their ability to support effective communication in real-world situations. Thus, this paper proposes a real-time communication aid system that converts sentence-level Korean dysarthric speech to non-dysarthric normal speech. The proposed system consists of two main parts in cascading form. Specifically, a Korean Automatic Speech Recognition (ASR) model is trained with dysarthric utterances using a conformer-based architecture and the graph transducer network - connectionist temporal classification algorithm, significantly enhancing recognition performance over previous models. Subsequently, a Korean Text-To-Speech (TTS) model based on Jointly Training FastSpeech2 and HiFi-GAN for End-to-End Text-to-Speech (JETS) is pipelined to synthesize high-quality non-dysarthric normal speech. These models are integrated into a single system on an app server, which receives 5-10 seconds of dysarthric speech and converts it to normal speech after 2-3 seconds. This can provide a practical communication aid for people with dysarthria.

Keywords: dysarthria; communication aid system; conformer-based automatic speech recognition; JETS-based text-to-speech; Korean

1. Introduction

Dysarthria is a type of speech disorder caused by abnormalities in the motor neurons, which impair the control of the articulatory organs necessary for speech production [1]. It is classified as a type of language disorder along with language comprehension and expression disorders, as well as voice disorders. Although there is no universally accepted objective classification system for dysarthria, it is estimated to account for approximately 30% of all speech disorders [2,3]. There are many factors that contribute to the development of dysarthria, including brain function problems, speech problems, hearing problems, and laryngeal problems [4], and many people have difficulty communicating socially due to dysarthria [4,5].

Unlike aphasia, which is caused by damage to the language centers of the brain and results in an inability to speak fluently, dysarthria occurs when motor control of the speech organs is impaired due to neurological problems, leading to difficulties in articulation and phonation while retaining the ability to generate speech. In contrast to aphasia, which prevents speech production entirely, dysarthria is caused by neurological disorders and can still produce speech. However, the pathological intelligibility of the speech is significantly reduced [6,7]. Therefore, effective treatment for dysarthria often involves the use of various communication aids along with speech therapy to improve the clarity of speech.

In decade, studies to develop various assistive tools for treating dysarthria have been carried out [8–10]. Among the various assistive tools for treating dysarthria, one of the most prominent assistive tools is Augmentative and Alternative Communication (AAC). AAC is an augmentative and alternative form of communication that helps people with speech or vocalization difficulties to

 doi:10.20944/preprints202412.0289.v1

communicate effectively. It can be implemented through diverse means such as pictures, symbols, boards, or electronic devices, and is widely used for people with different speech disorders, including cerebral palsy, developmental disabilities, and dysarthria [8]. In particular, Voice Output Communication Aids (VOCAs), which convey a user's intended message through speech output, is based on inputs like text entry or the selection of pictures or symbols. VOCAs have been shown to be effective in improving communication for individuals with speech impairments [9]. A notable example of VOCAs research is the Voice-Input Voice-Output Communication Aid (VIVOCA), which takes simple speech inputs from individuals with dysarthria in specific situations, recombines them, and produces speech output through voice synthesis [10].

In addition to AAC, a prominent method for aiding communication in individuals with dysarthria is Automatic Speech Recognition (ASR) technology. If ASR can accurately recognize speech from individuals with dysarthria, it could be used in various ways to support their communication needs [11]. Research trends in ASR technology for dysarthria suggest that ASR studies based on early machine learning methods began in the mid-1990s [12–14]. From the 2010s onward, research focused on developing ASR models using deep learning technologies and exploring ways to augment the limited speech data available for dysarthria [15,16]. Subsequently, to enhance model performance, studies on Audio-Visual Speech Recognition (AVSR) were introduced, which involved learning both the speech sounds of dysarthria and the speaker's lip movements during speech [17–19]. Furthermore, Speak Vision, which interprets speech data visually, have been introduced, and ongoing research continues to improve ASR for individuals with dysarthria [20,21].

Recently, ASR performance for elderly individuals and those with dysarthria have been significantly advanced [22–24]. Hussain Albaqshi and Alaa Sagheer (2020) [22] developed the necessity for a tailored dysarthric ASR system based on a hybrid model combining Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) because ASR systems trained on normal speech are not effective for dysarthric language. In [23], Soleymanpour et al. (2022) used a multi-speaker end-to-end Text-to-Speech (TTS) system, capable of synthesizing high-quality speech with diverse prosodies, to generate realistic dysarthric speech for data augmentation. This allowed the inclusion of speech reflecting varying degrees of dysarthria, thus improving the performance of the ASR system. And Baali et al. (2023) proposed dysarthric ASR for Arabic speech [24]. Despite advancements in ASR, especially for Arabic, it is difficult to obtain dysarthric speech data. In [24], the first ASR model for Arabic dysarthric speech is developed with simulated dysarthric speech artificially modified from healthy Arabic speech and contributes significantly to the Arabic ASR field.

Table 1 summarizes the key research studies mentioned earlier. Most of these studies were conducted using English datasets, i.e., UA-Speech [25] and Torgo [26], and research on ASR models for dysarthria in languages with limited available data is unusual. Moreover, most studies report performance levels below 70%, and the AAC system with the highest performance was evaluated based on the recognition of specific target words and message reconstruction, which makes it difficult to directly compare with other studies that have different evaluation criteria. Furthermore, research conducted in word-level speech recognition environments cannot be applied to sentence-level speech recognition typically used in everyday conversations, making it unsuitable for real-time communication systems in various situations.

Table 1. Summary of communication aid systems.

Reference	Type	Method	Speech Recognition Level	Dataset	Performance (WRA)
Hawley et al. (2012) [10]	AAC	ASR + Message Reconstruction + TTS	Extremely limited Word Level	Self-collection (English)	96% (Restrict specific words)

 doi:10.20944/preprints202412.0289.v1

Yu et al. (2023) [20]	ASR	Audio-Visual fusion framework (Audio + Motion images of faces)	Word Level	UA-Speech (English)	36% ~ 94% (Extremely Severe ~ Mild)
Shahamiri (2021) [21]	ASR	Spatial-CNN (Using Voicegram)	Word Level	UA-Speech (English)	64%
Almadhor et al. (2023) [22]	ASR	Spatial-CNN + Transformer (Using Voicegram)	Word Level	UA-Speech (English)	65%
Hussain Albaqshi and Alaa Sagheer (2020) [23]	ASR	CRNN (CNN + RNN)	Word, Phrase, Sentence	Torgo (English)	40%
Soleymanpour et al. (2022) [24]	TTS	Multi-Speaker TTS (Data augmentation)	Word, Sentence	Torgo (English)	59% ~ 61%
Baali et al. (2023) [25]	ASR	Wave GAN + Conformer (Data augmentation + ASR)	Word, Sentence	Self-collection (Arabic), Torgo, LJspeech (English)	82%

Individuals with dysarthria often experience unpredictable variations in specific pronunciations, and these speech patterns exhibit distinct characteristics unique to each person. Moreover, there is a higher uncertainty at phoneme boundaries compared to standard speech. These peculiar speech patterns and the increased phoneme uncertainty are more pronounced depending on the language, indicating that dysarthric ASR systems must be custom-trained for each language. However, the majority of ongoing research is conducted using English datasets, with very little research on dysarthric ASR for minority languages, including Korean.

Therefore, this paper develops a custom-designed dysarthric ASR model for Korean dysarthria. This model efficiently learns the phonetic and visual characteristics of dysarthric speech through the Conformer encoding architecture, and its performance is compared with that of traditional transformer-based ASR models. Additionally, by combining the developed ASR model with a Korean TTS model, we aim to create a cutting-edge real-time communication aid system for dysarthria that can assist in social communication situations. To achieve this, the ASR model, using an advanced TTS model with excellent performance, synthesizes the text recognized by the dysarthric ASR model into high-quality Korean speech. Through this process, individuals with dysarthria will be able to freely produce normal speech in various situations, significantly improving their communication abilities.

2. Related Works

In this section, we briefly describe two prior studies that are highly relevant to this research: VIVOCA, proposed in [10], is a study on a communication aid system for dysarthria, which is closely related to this research in that it performs the process of speech output through speech input; and [27] is a study on Korean dysarthric speech recognition, which is highly relevant in that it implements a dysarthric ASR model using the same dataset as the one used in this research.

2.1. AAC:VIVOCA

VIVOCA is one of the representative AAC systems developed for individuals with severe dysarthria. It works by receiving simple spoken commands from the user, which are then

 doi:10.20944/preprints202412.0289.v1

reassembled into messages and output as speech. The traditional input methods used in VOCAs, i.e., switches or keyboards, have been found to be cumbersome and tiring, which hinders natural communication. To address this problem, VIVOCA proposes using speech input instead of switches or keyboards. The system allows the user to read words from a pre-configured list, and it recognizes the spoken word to generate possible following words. By repeating this process, the user's intent is gradually refined, and the final sentence that the user intends to convey is generated, which is then output as speech using TTS. The model for dysarthric speech recognition in VIVOCA is based on a statistical model, the Hidden Markov Model, and the results of validating the word reassemble process showed high recognition accuracy. Therefore, VIVOCA can be highly useful for individuals with severe dysarthria in performing basic communication tasks. However, a limitation of VIVOCA is that users cannot manipulate their speech with fine control, which means it can only be used in specific situations. Subsequently, the system operates based on specific words or phrases, making it insufficient for processing all of the user's speech in real-time.

2.2. Korean Dysarthric Speech Recognition

In [27], an ASR model built for dysarthric speech recognition is provided with a dataset. This model is based on a hybrid Connectionist Temporal Classification (CTC) and attention mechanism (CTC/attention) [28], utilizing a transformer-based encoder and decoder. The model was developed and trained using the ESPnet framework [29], referencing the transformer ASR architecture described in [30] and [31].

The overall construction and training process of the model is outlined in Figure 1. The model first extracts features from the input data through a Visual Geometry Group (VGG) with multiple CNN layers. The extracted feature vectors are then used to perform sequence modelling using a transformer. Through this process, the visual patterns of the speech data used as input can be reflected in the learning. However, this structure has the disadvantage that each process is performed independently, resulting in low computational efficiency and difficulty in handling long sequence inputs. It is also classified as a rather classical model because it consists of separate modules, which can lead to a lack of integration of local patterns obtained from the VGG into global patterns, and scaling problems often occur in the connection between the two modules.

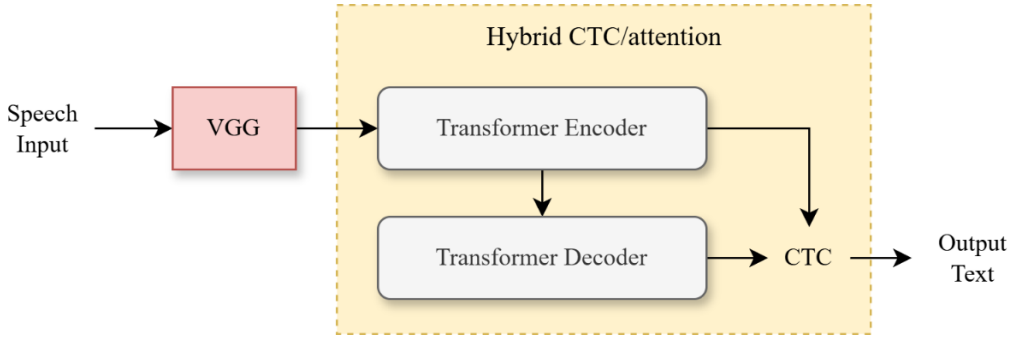


Figure 1. The state-of-the-art Korean dysarthric ASR model structure.

3. Proposed Real-Time Communication Aid System for Korean Dysarthric Speech

In order to overcome the shortcomings of [10], which lacks real-time processing of various speech, this study proposes a system that can recognize the user's speech as a whole sentence and process it in real-time. By doing so, we aim to expand the communication range of dysarthria and create an environment where natural dialogue is possible in real time. In addition, to overcome the low computational efficiency of the model structure used in [27] and the processing problem for long sequences, we use the conformer structure. Conformer is optimized for processing temporal sequences such as speech data, and it improves computational efficiency and enhances the learning of local (short-range) patterns in the input speech data. By doing so, we aim to improve the performance of the improved model on the same dataset compared to the existing model.

 doi:10.20944/preprints202412.0289.v1

Figure 2 illustrates the overall architecture of the real-time communication aid system proposed in this study. The core components of the proposed system include:

1. ASR model for Korean dysarthric speech recognition: This model processes speech input from individuals with dysarthria, identifying and converting it into textual information.
2. TTS model for normal speech synthesis: Based on the recognized text, this model generates high-quality non-dysarthric speech, making the output more natural and intelligible for effective communication.

The developed ASR and TTS models are integrated through a web server to ensure seamless operation. The system processes a single speech input and produces a single speech output in real-time, aiming to provide a practical and efficient solution for assisting individuals with dysarthria in various communication scenarios.

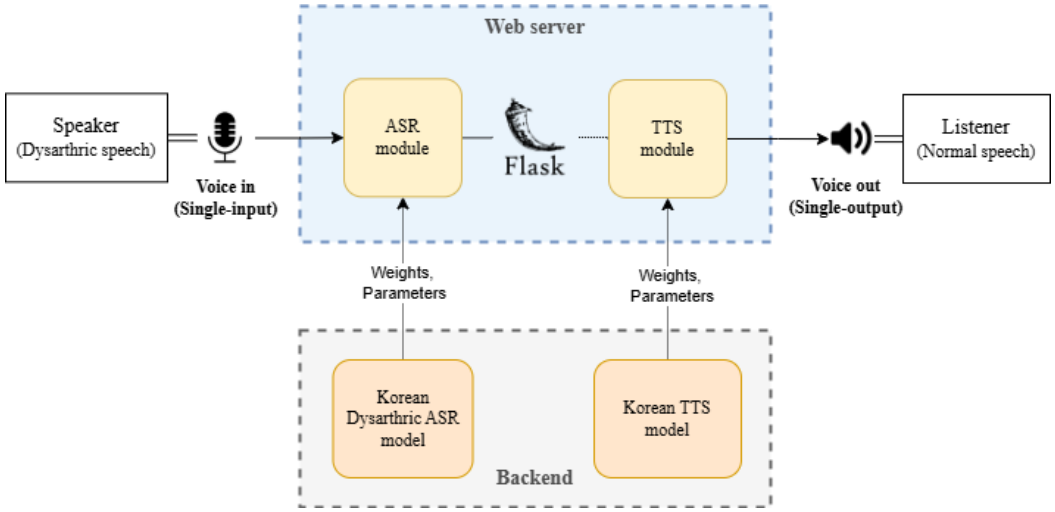


Figure 2. Real-time dysarthria communication aid system structure.

The flow of the entire system starts with the dysarthric speech input captured via a microphone. To process the input speech signal, the ASR and TTS models are modularized by loading the weights and parameters of each model, enabling them to make predictions for a single input on the web server. The ASR and TTS modules are integrated into the web server using Python's Flask framework. The ASR model processes the captured speech to generate a text output, which is then immediately passed as input to the TTS model. Through this integration, the input dysarthric speech is directly transformed into clear non-dysarthric speech on the web server. As a result, users can interact more smoothly with others in social communication situations using the synthesized non-dysarthric normal speech.

3.1. Conformer-Based Korean Dysarthric ASR

The goal of the Korean dysarthric ASR model for the real-time communication aid system is to develop a high-accuracy model to ensure that the system can be correctly applied in real-life social communication situations for individuals with dysarthria.

The ASR model training is based on the hybrid CTC/attention model [28]. It has a similar structure to the model built in [27], but uses a Conformer structure [32], which is a combination of convolution and transformer, rather than transformer, as the hierarchy for modelling the attention mechanism. The conventional transformer structure, which uses VGG to extract feature vectors for the input sequence, suffers from low computational efficiency and difficulty in processing long sequences. In addition, since VGG and transformer operate independently, there is a lack of integration between local and global information at the encoder level. In contrast, the conformer encoder combines self-attention and convolution in parallel, allowing it to capture both global and local patterns within the input sequence. This enables the model to effectively learn the temporal dynamics and visual features of the speech structure in dysarthric speech. Through this dual

 doi:10.20944/preprints202412.0289.v1

functionality, the model maintains a strong understanding of long-range dependencies while focusing on relevant acoustic cues, making it better suited to handle the variability and irregularities present in Korean dysarthric speech.

Figure 3 illustrates the entire training process of the ASR model. The process begins with inputting speech data x(t) in wav format, which is then transformed into a log-Mel spectrogram as described in Equations (1)-(3). In Equation (1), ω(n) represents the window function, τ denotes the time index, and ω corresponds to the frequency index:

$$X(\tau, \omega) = \sum_{n=-\infty}^{\infty} x(n) \cdot \omega(n - \tau) \cdot e^{-j\omega n} \quad (1)$$

The resulting spectrogram is passed through a Mel filterbank H to generate the Mel spectrogram S_{Mel}:

$$S_{Mel} = H \cdot |X(\tau, \omega)|^2 \quad (2)$$

Subsequently, the log transformation is applied to compress the dynamic range, yielding the log-Mel spectrogram S_{LogMel}:

$$S_{LogMel} = \log(S_{Mel} + \epsilon), \quad (3)$$

Here, ϵ is a small positive constant added to prevent numerical instability and ensure that the logarithm operation does not encounter undefined values when S_{Mel} approaches zero. This log-Mel spectrogram serves as the input to the conformer encoder. Both the encoder and decoder perform positional embedding at the early stages of the model, which vectorizes the word order and the order of acoustic features to use as positional information. In particular, during the positional embedding process in the conformer encoder, a CNN is incorporated to more finely learn the local patterns present in the positional information of the dysarthric speech sequence. The CNN generates feature maps reflecting local patterns, which better capture local information and structural relationships at each position.

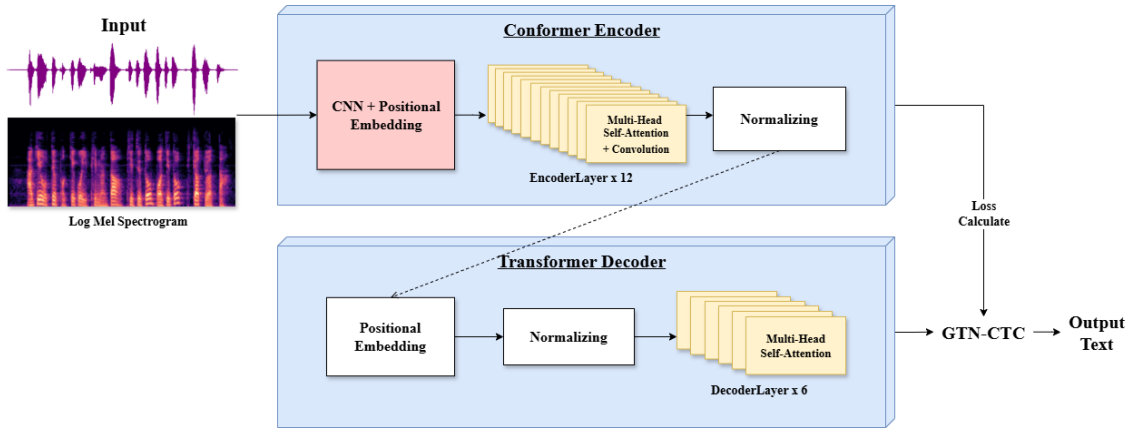


Figure 3. The full learning process for the Korean dysarthric ASR model.

The vectors converted through embedding are then passed through the encoder layer, which consists of 12 layers in total. The learning process for each layer block is shown in Figure 4.

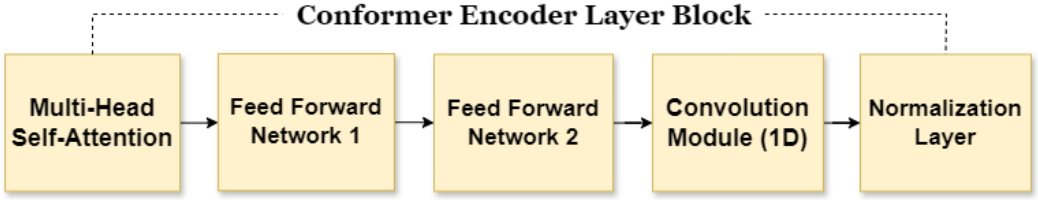


Figure 4. Learning process for conformer encoder layer block.

 doi:10.20944/preprints202412.0289.v1

In the first step of the layer block, Multi-Head Self-Attention (MHSA) [33] is applied. For the given input S_{LogMel}, Query (Q), Key (K), and Value (V) vectors are generated for each word in the input vector sequence. These vectors are then processed using the attention mechanism defined in Equation (4):

$$\text{Attention}(\text{Q}, \text{K}, \text{V}) = softmax(\frac{QK^T}{\sqrt{d_k}})V \quad (4)$$

Here, d_k represents the dimension of the Key vector. The attention mechanism computes a weighted sum of the value vectors, where the weights are determined by the similarity between the query and key vectors. This enables the model to selectively emphasize relevant features in the input sequence.

Following MHSA, the output is passed through two Feed-Forward Network (FFN) layers [34], which apply nonlinear transformations to further refine the extracted features. The computation in each FFN layer, FFN_1 and FFN_2, is described by Equation (5):

$$FFN_{1,2} = Swish(XW_1 + b_1)W_2 + b_2 \quad (5)$$

Here, X represents the input, W_1 and W_2 are the weight matrices of the linear layers, and b_1 and b_2 are the bias vectors. The activation function used in this model is Swish [35], which has been shown to improve model performance in various deep learning tasks.

The next component, the convolution module, is designed to capture local patterns within the input sequence. This module uses a pointwise convolution and a depthwise convolution for each kernel. These operations are followed by batch normalization and activation functions to stabilize and enhance learning. The resulting features are passed through a normalization layer.

The normalization layer integrates multiple operations, including layer normalization, residual connections, and dropout, to produce the final output of the block. The overall process is summarized in Equation (6):

$$\text{Y} = Dropout(LayerNorm(X + Sublayer(X))) \quad (6)$$

Here, Sublayer(X) encompasses all preceding operations, such as MHSA, FFN, and the convolution module. The residual connection ensures stable gradient flow, while layer normalization mitigates internal covariate shift, and dropout reduces overfitting.

Based on the output obtained through the conformer encoder, the decoder generates the output text sequence. Using the MHSA mechanism, the decoder selectively emphasizes important information from the speech sequence during text generation. In this process, CTC learns the alignment between the output of the encoder and the text sequence, ultimately generating the final output text.

For the CTC algorithm, we utilize the Graph Transformer Network-based CTC (GTN-CTC) [36] algorithm. GTN-CTC is effective in handling long input sequences and large datasets compared to traditional CTC approaches. By leveraging a graph structure, it can manage complex dependencies, resulting in faster computation speeds and improved memory efficiency. This allows the model to align and process speech patterns of varying lengths, which are commonly found in dysarthric speech, without significant accuracy loss.

3.2. Korean TTS Model

Based on the recognition results from the ASR model, we construct a TTS model to generate Korean non-dysarthric speech. For the TTS model architecture, we use Jointly Training FastSpeech2 and HiFi-GAN for End-to-End Text-to-Speech (JETS) [37]. JETS combines the popular TTS architecture FastSpeech2 [38] and the GAN-based speech synthesis model HiFi-GAN [39] for joint training. This approach simplifies the traditional two-stage TTS training process into a single unified training process, allowing for high-quality speech synthesis results. The entire process is shown in Figure 5.

 doi:10.20944/preprints202412.0289.v1

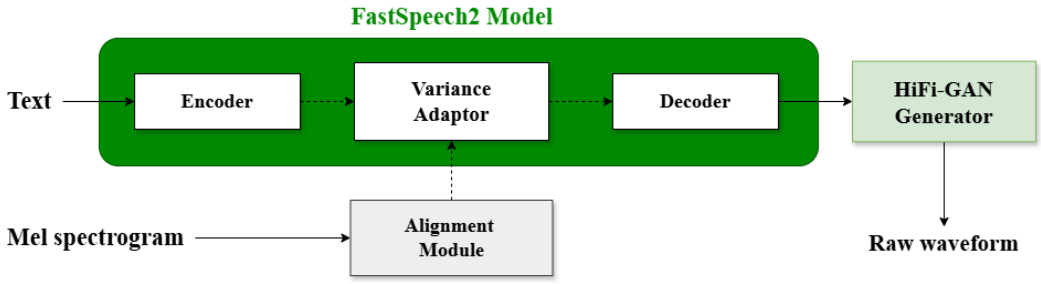


Figure 5. Learning process in JETS.

The components of JETS can be broadly divided into three parts: FastSpeech2, HiFi-GAN, and the Alignment Module. In the first stage, FastSpeech2 takes text input and learns the duration d, pitch p, and energy e for each text token through a variance adaptor. The corresponding loss function, L_{var}, is defined in Equation (7) and is based on the predicted values of these feature sequences $\hat{d}$, $\hat{p}$, and $\hat{e}$:

$$L_{var} = \left\| d - \hat{d} \right\|_2 + \|p - \hat{p}\|_2 + \|e - \hat{e}\|_2 \quad (7)$$

FastSpeech2 combines the text embeddings and the variance features predicted by the variance adaptor to generate the Mel spectrogram M.

Then, in the alignment module, the alignment between the Mel spectrogram M and the text embedding E is learned. The objective function for alignment learning can be simplified as shown in equation (8), and it is computed as shown in equation (9):

$$L_{align} = -\log P(M|E) \quad (8)$$

$$P(M|E) = \prod_{i=1}^{N} P(M_i|E) \quad (9)$$

Here, N represents the number of frames in the Mel spectrogram. This alignment process ensures that the temporal structure of the Mel spectrogram matches the sequence of text embeddings.

Finally, HiFi-GAN takes the output of the decoder as input and synthesizes the raw waveform. During this process, the adversarial learning between the generator and the discriminator results in the GAN loss function, denoted as L_{GAN}. The overall loss function L for JETS integrates the contributions from the variance adaptor, alignment module, and HiFi-GAN, as defined in Equation (10):

$$L = \lambda_{GAN} L_{GAN} + \lambda_{GAN} L_{var} + \lambda_{GAN} L_{align}, \quad (10)$$

Here, λ_{GAN}, λ_{var}, and λ_{align} represents the weight of each loss function. Therefore, based on the overall loss function L defined as in equation (10), JETS converts text to Mel spectrograms through FastSpeech2, learns the alignment between text and Mel spectrograms through the Alignment Module, and finally generates high-quality raw audio through HiFi-GAN, implementing an efficient and integrated TTS model.

4. Experiments

4.1. Experimental Setup

For the development of a Korean ASR model for dysarthria, we use the dysarthric speech recognition data provided by AI-hub [27], which was collected from hospitals in Korea. This dataset contains over 5,000 hours of speech data collected from approximately 1,200 dysarthric patients, covering various ages, regions, genders, and diseases. The data was collected in a continuous format with about 10-second utterances from dysarthric patients followed by approximately 30 s of silence. To remove this silence and segment the data into individual sentences or words, Voice Activity Detection was used. After preprocessing, 17,205 short sentences with fewer than 3 words, 35,681

 doi:10.20944/preprints202412.0289.v1

medium-length sentences containing 4 to 9 words, and 6,511 long sentences with more than 10 words were generated. Training was performed using approximately 100 hours of pure dysarthric speech data.

For the development of the Korean TTS model, we use the KSS Dataset [40]. This dataset consists of approximately 12 hours of clean speech data from a female announcer and is commonly used for Korean TTS model development. It contains a total of 12,853 medium-length sentences, all of which were used for training without any additional preprocessing.

The hardware environment for the experiments was identical across all tests and consisted of a server computer with an NVIDIA RTX A6000 GPU, a 64-core 2.7 GHz CPU, and 400 GB of RAM. Additionally, both the TTS and ASR model training were conducted using the ESPnet toolkit [29].

4.2. Experimental Results

We conduct training results and performance evaluation for both the ASR model and the TTS model. Additionally, we visualize the input and output speech data of the entire system and present comparisons.

4.2.1. ASR Model

The Korean dysarthric ASR model was trained for a total of 35 epochs, with a batch size of 20, and 23,450 training steps. It was performed based on 2,309 byte-pair encodings. The training took approximately 20 hours, and the results presented an accuracy of 98.1% for the training data and 94.4% for the validation data.

To measure the performance of the proposed model, we conduct a performance evaluation by comparing it with the state-of-the-art Korean dysarthric ASR model based on the hybrid CTC/attention transformer encoder-decoder structure in [27]. The hybrid CTC/attention learning method, which incorporates both CTC and attention losses, is the same as the proposed model. However, in this existing model, VGG for local pattern learning of input data is performed independently from the Transformer, and the CTC algorithm follows the conventional approach. For evaluation, a dataset of approximately 2,500 sentences not used in training was employed. The evaluation metrics include the Character Error Rate (CER), Word Error Rate (WER), and the Error Reduction Rate (ERR), which indicates the performance improvement in comparison with the previous model for each error rate.

Table 2 shows the performance evaluation results. The results show that the performance improvement is about 42% for CER and about 38% for WER compared to the existing model. Consequently, this indicates the superiority of the proposed conformer-based GTN-CTC application model for Korean dysarthric speech over the existing ASR model.

Table 2. Performance evaluation results of the proposed ASR model.

Model	CER (%)	WER (%)	ERR (%)
Hybrid CTC/attention transformer model [27]	15.6	18.8	41.7
Proposed conformer-based GTN-CTC model	9.1	11.6	38.3

4.2.2. TTS Model

The TTS model was trained by using the KSS dataset for a total of 150 epochs over the course of 7 days. To evaluate the performance of JETS to be used in the proposed system, we used the Tacotron2 and Transformer TTS models provided by ESPnet-TTS to measure and compare their performance. The evaluation metric used for performance measurement was the subjective Mean Opinion Score (MOS). For MOS measurement, 10 participants from schools in Korea were asked to evaluate 20

 doi:10.20944/preprints202412.0289.v1

different audio samples generated by each model. The evaluation was based on a scale of 1 (poor), 2 (fair), 3 (average), 4 (good), and 5 (excellent), and was assessed based on how natural the synthesized speech was compared to the original, unprocessed Korean speech.

Table 3 shows the MOS results for the three TTS models. The evaluation results indicate that JETS achieved the highest score of 4.64, while Tacotron2 scored 3.03, and Transformer scored the lowest at 2.51. The overall tendency in the evaluation of the synthesized speech is that the speech generated by JETS was considered natural both in terms of intonation and voice quality, whereas the speech synthesized by Tacotron2 and Transformer TTS has natural intonation but was dominated by mechanical-sounding voices.

Table 3. Evaluate MOS comparisons for different TTS models.

TTS Model	MOS (1-5)
JETS [37]	4.64
Tacotron 2 [42]	3.03
Transformer TTS [43]	2.51

4.3. Integrated Communication Aid Systems

Finally, the trained ASR and TTS models are combined to build the entire communication system. The entire system operates as a single program, and during execution, it receives speech input from the user via the user interface. As the length of the input increases, the time to obtain the final output also increases. However, if the user provides speech input lasting 5 to 10 s, the final output is obtained approximately 2 to 3 s after input. Given that typical speech in daily life lasts about 3 to 5 s, this demonstrates that the system is sufficient for real-time application in practice.

Figure 6 shows a visualization of the Korean dysarthric speech used as input and the synthetic utterance generated by the entire system. In (a), corresponding to the dysarthric speech, we can see that there are frequent silences in the middle of the utterance. This is due to the nature of dysarthria, which is a speech blockage caused by dysregulation of the articulatory system during speech, which causes great confusion for the listener in understanding the speech. When such dysarthric speech is fed into a communication aid system, all the silent intervals caused by speech blockages are removed and resynthesized into a natural rhyming utterance, (b).

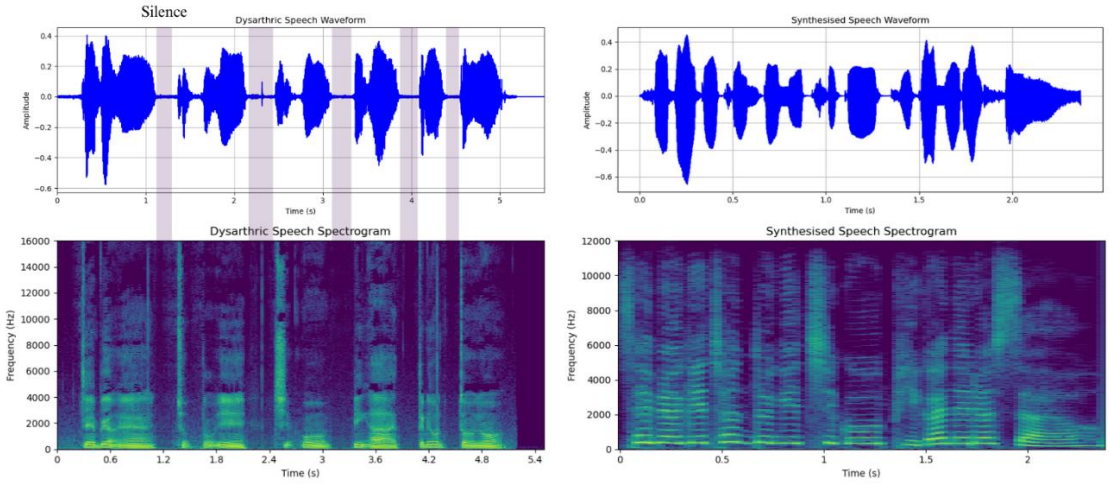
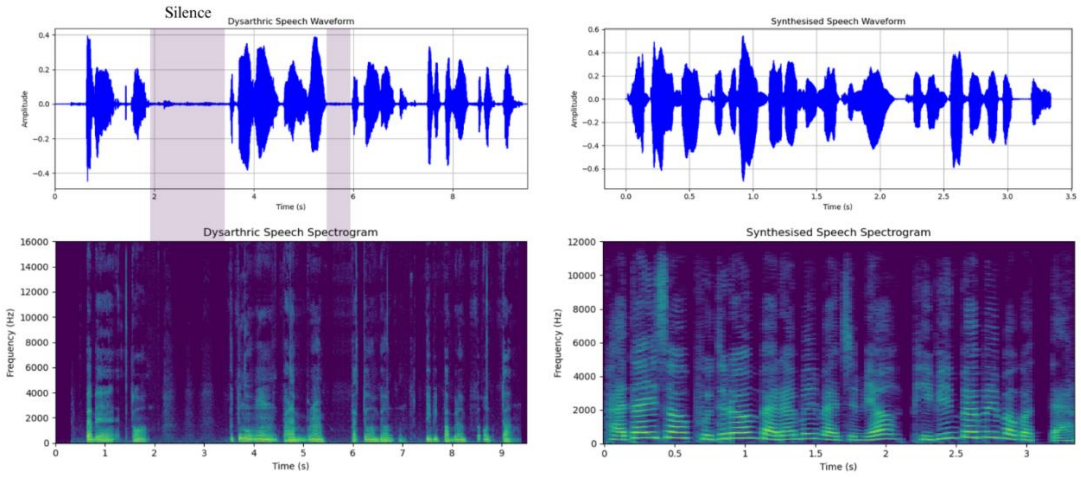


Figure 6. Visualization of the system output for the Korean dysarthric speech input: (Meaning: “At dawn, it thundered and rained.”) and (English phonetic transcription: "saebyeoge cheondungi chigo biga naeryeosseoyo.").

 doi:10.20944/preprints202412.0289.v1

Figure 7 shows the input and output results for a similar case to Figure 6, but for a slightly longer speech: a dysarthric speech of about 9.5 seconds has been significantly shortened to about 3.5, and the long pauses in the middle of the speech have been eliminated. This process of speech recognition and synthesis makes the relatively long speech shorter and clearer, allowing the user to produce natural speech.



(c) Dysarthric speech input (d) Synthetic speech output generated from (c)

Figure 7. Visualization of the system output for the Korean dysarthric speech input: (Meaning: “I ate bibimbap with a gentle breeze from the sea”) and (English phonetic transcription: "badaeseo budeureoun barameul majeumyeo bibimbabeul meogeossda.").

5. Conclusions

This paper proposes a Real-Time Communication Aid System for Korean dysarthria using ASR and TTS models. The dysarthric ASR model, the core component of the system, is designed using a conformer-based encoder and GTN-CTC, and is targeted at Korean, a language that has rarely been applied to previous dysarthric ASR research. The proposed dysarthric ASR model achieves a performance improvement of about 41% in terms of CER and 38% in terms of WER compared to the conventional transformer-based model. The TTS model for Korean speech synthesis was developed based on JETS, resulting in high-quality Korean speech synthesis. It also showed the highest MOS score of 4.64, outperforming existing Tacotron2 and transformer-based TTS models. The generated ASR and TTS models are integrated via a web server to immediately convert the input dysarthric speech into clear non-dysarthric speech. This enables individuals with dysarthria to communicate in real-time using non-dysarthric speech in everyday life. Furthermore, the web-based design of this system can be integrated with small wearable devices, making it applicable as a versatile assistive device for various types of dysarthria. Future research may explore expanding the scope of the system by integrating it with hardware to support a wider range of communication needs.

Author Contributions: Conceptualization, K.P. and J.P.; methodology, K.P. and J.P.; software, K.P.; validation, J.P.; formal analysis, K.P.; investigation, K.P.; resources, J.P.; data curation, K.P.; writing—original draft preparation, K.P.; writing—review and editing, K.P. and J.P.; visualization, K.P.; supervision, J.P.; project administration, K.P. and J.P.; funding acquisition, K.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the ‘Student-Initiated Creative Research Project’ at Changwon National University in 2024.

Data Availability Statement: This research (paper) used datasets from ' Dysarthric speech recognition data (AI-Hub, S. Korea)'. All data information can be accessed through 'AI-Hub (www.aihub.or.kr)'.

Conflicts of Interest: The authors declare no conflicts of interest.

 doi:10.20944/preprints202412.0289.v1

References

1. Enderby, P. Disorders of communication: dysarthria. *Handbook of Clinical Neurology* **2013**, 110, 273–281.
2. Kim, S.-J.; Lee, H.-J.; Park, K.-Y.; Lee, J.-S. A survey of speech sound disorders in clinical settings. *Communication Sciences & Disorders* **2015**, 20(2), 133–144.
3. Preston, J.L.; Hull, M.; Edwards, M.L. Preschool speech error patterns predict articulation and phonological awareness outcomes in children with histories of speech sound disorders. 2013.
4. Rampello, L.; D'Anna, G.; Rifici, C.; Scalisi, L.; Vecchio, I.; Bruno, E.; Tortorella, G. When the word doesn't come out: A synthetic overview of dysarthria. *Journal of the Neurological Sciences* **2016**, 369, 354–360.
5. Page, A.D.; Yorkston, K.M. Communicative participation in dysarthria: Perspectives for management. *Brain Sciences* **2022**, 12(4), 420.
6. Jordan, L.C.; Hillis, A.E. Disorders of speech and language: aphasia, apraxia and dysarthria. *Current Opinion in Neurology* **2006**, 19(6), 580–585.
7. Selouani, S.-A.; Sidi Yakoub, M.; O'Shaughnessy, D. Alternative speech communication system for persons with severe speech disorders. *EURASIP Journal on Advances in Signal Processing* **2009**, 2009, 1–12.
8. Light, J.; McNaughton, D. Supporting the communication, language, and literacy development of children with complex communication needs: State of the science and future research priorities. *Assistive Technology* **2012**, 24(1), 34–44.
9. Judge, S. The Design of Voice Output Communication Aids; Ph.D. Thesis, University of Sheffield, Sheffield, UK, 2023.
10. Hawley, M.S.; Cunningham, S.P.; Green, P.D.; Enderby, P.; Palmer, R.; Sehgal, S.; O'Neill, P. A voice-input voice-output communication aid for people with severe speech impairment. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* **2012**, 21(1), 23–31.
11. Tu, M.; Singh, A.; Tang, H. The relationship between perceptual disturbances in dysarthric speech and ASR performance. *The Journal of the Acoustical Society of America* **2016**, 140(5), EL416–EL422.
12. Qian, Z.; Xiao, K. A survey of automatic speech recognition for dysarthric speech. *Electronics* **2023**, 12(20), 4278.
13. Bharti, K.; Das, P.K. A survey on ASR systems for dysarthric speech. In Proceedings of the 2022 4th International Conference on Artificial Intelligence and Speech Technology (AIST), Bhubaneswar, India, 9–11 December 2022; IEEE, 2022; pp. 1–6.
14. Jayaram, G.; Abdelhamied, K. Experiments in dysarthric speech recognition using artificial neural networks. *Journal of Rehabilitation Research and Development* **1995**, 32, 162–162.
15. Vachhani, B.; Bhat, C.; Kopparapu, S.K. Data augmentation using healthy speech for dysarthric speech recognition. In Proceedings of Interspeech 2018, Hyderabad, India, 2–6 September 2018; pp. 471–475.
16. Harvill, J.; Bejleri, A.; Arfath, M.; Das, D.; Thomas, S. Synthesis of new words for improved dysarthric speech recognition on an expanded vocabulary. In Proceedings of the ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 6–11 June 2021; IEEE, 2021; pp. 6428–6432.
17. Liu, S.; Liu, J.; Zhao, J.; Zhang, H.; Li, P.; Meng, H. Exploiting visual features using Bayesian gated neural networks for disordered speech recognition. In Proceedings of Interspeech 2019, Graz, Austria, 15–19 September 2019; pp. 4120–4124.
18. Zhang, S.; Du, J.; Dai, L.; Lee, C.-H. Robust audio-visual speech recognition using bimodal DFSMN with multi-condition training and dropout regularization. In Proceedings of ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; IEEE, 2019; pp. 6570–6574.
19. Yu, C.; Su, X.; Qian, Z. Multi-stage audio-visual fusion for dysarthric speech recognition with pre-trained models. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* **2023**, 31, 1912–1921.
20. Shahamiri, S.R. Speech vision: An end-to-end deep learning-based dysarthric ASR system. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* **2021**, 29, 852–861.
21. Almadhor, A.; Qaddoumi, A.; Al-Obaidi, S.; Al-Ali, K. E2E-DASR: End-to-end deep learning-based dysarthric ASR. *Expert Systems with Applications* **2023**, 222, 119797.
22. Albaqshi, H.; Sagheer, A. Dysarthric speech recognition using convolutional recurrent neural networks. International Journal of Intelligent Engineering & Systems 2020, 13(6), 1–10.
23. Soleymanpour, M.; Karuppiah, S.; Narayanasamy, T. Synthesizing dysarthric speech using multi-speaker TTS for dysarthric speech recognition. In Proceedings of ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 22–27 May 2022; IEEE, 2022; pp. 7382–7386.
24. Baali, M.; Alami, M.; Amar, M.; Merzouk, R. Arabic dysarthric speech recognition using adversarial and signal-based augmentation. *arXiv Preprint* **2023**, arXiv:2306.04368.
25. Kim, H.; Chang, J.; Yun, S.; Kim, Y. Dysarthric speech database for universal access research. In Proceedings of Interspeech 2008, Brisbane, Australia, 22–26 September 2008; pp. 1741–1744.

 doi:10.20944/preprints202412.0289.v1

26. Rudzicz, F.; Namasivayam, A.K.; Wolff, T. The TORGO database of acoustic and articulatory speech from speakers with dysarthria. *Language Resources and Evaluation* **2012**, 46, 523–541.
27. Dysarthric speech recognition data. Available online: https://aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&dataSetSn=608 (accessed on 30 November 2024).
28. Xiao, Z.; Liu, W.; Huang, L. Hybrid CTC-attention based end-to-end speech recognition using subword units. In Proceedings of the 2018 11th International Symposium on Chinese Spoken Language Processing (ISCSLP), Taipei, Taiwan, 26–29 November 2018; IEEE, 2018; pp. 146–150.
29. Watanabe, S.; Hori, T.; Karita, S.; Hayashi, T.; Nishitoba, S.; Unno, Y.; Morita, K.; Kawakami, K.; Baskar, M.A.; Fujita, Y. ESPnet: End-to-end speech processing toolkit. *arXiv Preprint* **2018**, arXiv:1804.00015.
30. Bang, J.-U.; Kim, J.-S.; Yoon, J.-H.; Joo, S.-Y.; Lee, S.-M.; Cho, S.-K. Ksponspeech: Korean spontaneous speech corpus for automatic speech recognition. *Applied Sciences* **2020**, 10(19), 6936.
31. Karita, S.; Watanabe, S.; Chen, Z.; Hayashi, T.; Hori, T.; Inaguma, H.; Jiang, Z.; Someki, M.; Soplin, N.E.Y.; Yamamoto, R. A comparative study on transformer vs. RNN in speech applications. In Proceedings of the 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Sentosa, Singapore, 14–18 December 2019; IEEE, 2019; pp. 449–456.
32. Guo, P.; Watanabe, S.; Kawahara, T.; Takeda, K. Recent developments on ESPnet toolkit boosted by conformer. In Proceedings of ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 6–11 June 2021; IEEE, 2021; pp. 5874–5878.
33. Voita, E.; Talbot, D.; Moiseev, F.; Sennrich, R.; Titov, I. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. *arXiv Preprint* **2019**, arXiv:1905.09418.
34. Bebis, G.; Georgiopoulos, M. Feed-forward neural networks. *IEEE Potentials* **1994**, 13(4), 27–31.
35. Ramachandran, P.; Zoph, B.; Le, Q.V. Searching for activation functions. *arXiv Preprint* **2017**, arXiv:1710.05941.
36. Hannun, A.; Case, C.; Casper, J.; Catanzaro, B.; Cong, Y.; Fougner, C.; Han, T.; Kang, Y.; Krishnan, P.; Prenger, R.; Sengupta, S. Differentiable weighted finite-state transducers. *arXiv Preprint* **2020**, arXiv:2010.01003.
37. Lim, D.; Jung, S.; Kim, E. JETS: Jointly training FastSpeech2 and HiFi-GAN for end-to-end text-to-speech. *arXiv Preprint* **2022**, arXiv:2203.16852.
38. Ren, Y.; Hu, C.; Tan, X.; Qin, T.; Zhao, S.; Zhao, Z.; Liu, T.-Y. FastSpeech 2: Fast and high-quality end-to-end text-to-speech. *arXiv Preprint* **2020**, arXiv:2006.04558.
39. Kong, J.; Kim, J.; Bae, J. HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis. *Advances in Neural Information Processing Systems* **2020**, 33, 17022–17033.
40. KSS Dataset: Korean Single Speaker Speech Dataset. Available online: https://www.kaggle.com/dataset/speech-recognition (accessed on 30 November 2024).
41. Hayashi, T.; Yasuda, K.; Watanabe, S.; Higuchi, Y.; Takeda, K.; Kawahara, T. ESPnet-TTS: Unified, reproducible, and integratable open source end-to-end text-to-speech toolkit. In Proceedings of ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; IEEE, 2020; pp. 7654–7658.
42. Shen, J.; Pang, R.; Weiss, R.J.; Schuster, M.; Jaitly, N.; Yang, Z.; Chen, Z.; Zhang, Y.; Wang, Y.; Skerrv-Ryan, R. Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; IEEE, 2018; pp. 4779–4783.
43. Li, N.; Liu, S.; Liu, Y.; Zhao, M.; Liu, M.; Zhou, L. Neural speech synthesis with transformer network. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; pp. 6706–6713.