Scientific Research Publishing

Journal of Computer and Communications, 2025, 13(4), 324-346
https://www.scirp.org/journal/jcc
ISSN Online: 2327-5227
ISSN Print: 2327-5219

Storytelling Style Speech Generation System: Emotional Voice Conversion Module Based on Cycle-Consistent Generative Adversarial Networks

Guangfeng Deng

Applied Information & Japanese Program, College of Languages, National Taichung University of Science and Technology, Taichung, Taiwan Region
Email: gfdeng@nutc.edu.tw

How to cite this paper: Deng, G.F. (2025) Storytelling Style Speech Generation System: Emotional Voice Conversion Module Based on Cycle-Consistent Generative Adversarial Networks. *Journal of Computer and Communications*, **13**, 324-346.
https://doi.org/10.4236/jcc.2025.134020

Received: February 23, 2024
Accepted: April 27, 2025
Published: April 30, 2025

Abstract

Telling a story requires various emotional ups and downs as well as pauses. Preparing a parallel corpus for emotional voice conversion is often costly and impractical. Developing high-quality non-parallel methods poses a significant challenge. Although non-parallel methods have been shown to enable emotional voice conversion, its capability for Chinese storytelling has not been clarified. Additionally, the storytelling results of emotional voice conversion have not been validated within a 3-12-year-old children. This study proposes a two-stage Chinese Storytelling Style Speech Generation System (SSPGS) composed of a text-to-speech system and an emotional voice conversion module. The SSPGS requires no parallel utterances, transcriptions, or time alignment procedures for speech generator training and requires only several minutes of training examples to generate reasonably realistic sounding speech. A small corpus neutral speech model is constructed on the text-to-speech system in the first stage, which is based on the speech synthesis system using a Hidden Markov Model (HMM). In the second stage, the emotional voice conversion module based on Cycle-Consistent generative adversarial networks (CycleGAN) is built. It enables the neutral speech generated by the text-to-speech system in the first stage to be transformed into the happiness, anger, and sadness necessary for storytelling tone using the timbre (spectrum), pitch (fundamental frequency F0), and rhythm (speech rate) of neutral speech. The validity of SSPGS is verified in two ways. First, a 5-point Mean Opinion Score (MOS) was performed for young children's parents. The results demonstrated that compared with general speech synthesizers, such as Google, the system generated more natural and genuine sound, that was more preferrable to the

target audience. After that, the kids underwent a story immersion evaluation. Analysis of the degree of engagement, liking, and empathy in listening to the story revealed no statistically significant difference between real-person dubbing and emotional speech synthesis dubbing. As a result, it has been initially confirmed that SSPGS might be added to the story robot product in the future.

Keywords

Storytelling Style Speech Generation System, Emotional Voice Conversion Module, Cycle-Consistent Generative Adversarial Networks, Text-to-Speech System, Mean Opinion Score, Immersion Measurement

1. Introduction

Story robot has a strong and rigid demand in the 3-12-year-old children's toy market. Therefore, many electronics manufacturers use cute animal designs to enter the market. Each story robot contains 20,000 stories and stories are continually improving every year. The number of stories made by domestic manufacturers in Taiwan region is only 1% of mainland China's. The cost of real-person dubbing is a significant factor. Using voice actors to record stories cost about 4~5 Taiwan region dollars per word. A short 1500-word story would cost about $6000~7500, so 10,000 short stories would cost more than $70 million for dubbing. Reduced dubbing costs and shorter dubbing times have emerged as problems that the industry must address.

Can a text-to-speech system (TTS) be used to tell the story? Leite *et al.* [1] point out that there are high-tension emotional ups and downs when telling a story. Sarkar *et al.* [2] also point out that when telling stories to children, it's necessary to pause more frequently. As a result, the storyteller will change the voice for the words with prominent emotions to introduce sad or happy in the story to increase the liveliness of the storytelling voice and gather the attention of the audience.

For emotional voice conversion technology, most studies aim to build an emotion conversion system using machine learning or deep learning [3]. These systems are generally trained on parallel data, which consists of the same sentences recorded in audio files that capture both neutral and various emotional tones. The systems include the Japanese language family [4]-[7], the Indian language family [8], the Korean language family [9] [10], English language family [11]-[13], and French language family [14]. These researches from different countries have focused on developing emotional conversion models in line with local language and culture to convert neutral speech to emotional speech (happy, angry, sad, surprised, etc.), and to increase the effect of emotional speech in storytelling. Given the challenges in obtaining emotional parallel corpus, the latest research trend is to use the non-parallel methods to establish emotional conversion under the condition of non-parallel training data.

As a result, generative adversarial networks (GAN) have attracted the most at-

tention. Kaneko *et al.* [15] are the first to create Cycle-Consistent GAN (CycleGAN) in the field of speech conversion. The research concentrated on transforming timbre (spectrum/spectrum parameters) eigenvalues with CycleGAN. Subsequent Asakura *et al.* [16] immediately propose to add the pitch parameter (pitch/fundamental frequency F0) to the application of emotional speech conversion.

Zhou *et al.* [17] use CycleGAN to convert the two eigenvalues of timbre and pitch and then merge them in English language. The results revealed that the spectral distortion rate (mel-cepstral distortion) was lower than that of the linear transformation model trained on the parallel corpus. Meanwhile, Rizos *et al.* [18] propose CycleGAN implementation for speech emotion conversion as a data augmentation method in English language. Additionally, He *et al.* [19] refine StarGAN for emotional voice conversion, reducing distortion and improving data augmentation with 2% and 5% F1 score gains in Japanese language. Finally, Moritani *et al.* [20] adapt StarGAN for emotional voice conversion in Japanese language, evaluating neutrality, similarity, and interdependence between emotional domains. Although non-parallel methods have been shown to enable emotional voice conversion, its capability for Chinese storytelling has not been clarified. Additionally, the results of emotional voice conversion have not been validated within a 3-12-year-old children.

This study proposes a Storytelling Style Speech Generation System (SSPGS) to convert rich emotional speech needed for Chinese speech synthesis to be applied to children's stories. Two stages make up the system. The first stage is developing a TTS system in the case of a small corpus recording. The second stage is to build an emotional speech conversion module that, in the absence of a parallel corpus, can change the timbre (spectrum), pitch (fundamental frequency F0), and speech rate necessary for emotional speech. The module allows for the conversion of a neutral voice to happy, angry, and sad. It uses CycleGAN to transform the spectrum (Mel-Generalized Cepstral Coefficients: MGCEP) and fundamental frequency F0 parameters, respectively. A 5-point Mean Opinion Score (MOS) was performed for young children's parents, while the kids themselves underwent a story immersion evaluation.

2. Related Work

2.1. Storytelling System

The idea of using speech synthesis in storytelling systems was first put forth in 2006 [21]. This research recorded three stories in the storytelling and conversation styles and manually examined the pitch contours of the two stories. It compared speech's average, variation, maximum, and minimum pitches developed a conversion model that can be applied at various points in a sentence and carried out an experiment on speech synthesis in the form of a story. The study successfully attracted attention. However, at the time this study was conducted, neither the idea of data-driven conversion models nor the exploration of emotional speech was included.

The follow-up research on speech synthesis for storytelling concentrated on the pause technology of speech synthesis. Using the pause model, sentences are given pauses of varying lengths to create the rhythm of the storytelling style (pause prediction). The rhythms of storytelling styles include Kazakh studies [22], Malay studies [23]-[27], Indian language studies [28]-[30], Chinese language studies [31]-[33] and Japanese studies [34] [35].

Parakrant *et al.* [28]-[30] applied speech synthesis to the Hindi-speaking storytelling system, which was based on a neutral text-to-speech system. It deployed: 1. Story-specific emotion detection module. 2. Story-specific prosody generation module (SSPG): Through the prosody rule set, each emotion has its corresponding prosodic parameters, which are converted through a simple linear transformation function, including pitch, duration/tempo, intensity, and pause. 3. Story-specific prosody incorporation module.

The operation process is as follows: Two processes are performed after the story corpus is read. First, the neutral voice is synthesized through the neutral voice TTS, and the sentence text is sent to the emotion recognition module for latent semantic analysis. Each sentence is then categorized into emotions. The corresponding acoustic parameters are then generated following the SSPG, and the modified parameters are included in the synthesized neutral speech by the merging module. In the follow-up, the scholar continued to develop the system. An effective three-stage story pause prediction model was created using the classification and regression tree (CART) through the data-driven approach [27]-[29].

For Malay speakers, Ramli *et al.* [25]-[27] discusses several storytelling speech synthesis systems that transform neutral speech into emotionless storytelling speech (emotion-insensitive) for Malay. In his research, experienced storytellers (one male and one female) recorded story parallel corpus. Each sentence was recorded in normal and storytelling style, comprising 116 sentences, 1164 words, and 2732 syllables. 124 stressed syllables were manually tracked using the program tool, and the stressed syllables were subjected to a cluster analysis of pitch contours. By comparing the differences between the global (sentence) and regional (word) conversion models of prosodic parameters, it was able to identify six patterns and solve the issue of inadequate conversion of existing pitch parameters.

Costa *et al.* [36] developed a storytelling robot that can present facial expressions as a storytelling carrier. It contrasted the use of real-person dubbing and voice synthesis for storytelling using empirical data. The results demonstrate significant empathy differences among listeners. However, there were no significant differences in engagement and liking. This study created a crucial benchmark for the advancement of speech synthesis in the narrative. However, Costa only used a single neutral voice TTS to synthesize stories. The neutral pronunciation like Google or Siri, is very different from the speech synthesis needed for storytelling.

2.2. Emotional Voice Conversion

Vuppala *et al.* [37] focus on converting neutral speech to angry speech by using

non-uniform duration modification, specifically altering vowel and pause durations. Subjective listening tests show better emotion perception compared to uniform duration modification. Building on this, Vydana *et al.* [38] propose a rule-based emotion conversion technique that also focuses on vowel-based non-uniform prosody modification. They analyze factors like position and identity to address non-uniformity in prosody and report better performance than existing methods in subjective listening tests.

Chatziagapi *et al.* [39] address data imbalance in Speech Emotion Recognition (SER) by exploring conditioned data augmentation with Generative Adversarial Networks (GANs). They enhance a conditional GAN architecture to generate synthetic spectrograms for underrepresented emotions. Compared to signal-based data augmentation methods, their GAN-based approach demonstrates a 10% performance improvement in IEMOCAP and 5% in FEEL-25k when augmenting minority classes.

Subsequently, Rizos *et al.* [18] introduce an adversarial network for speech emotion conversion that uniquely does not require parallel data. Their method improves classification by 2% and 6% in Micro- and MacroF1 scores, although they note some human confusion exists in complex emotional attributes. In a similar vein, He *et al.* [19] propose an improved StarGAN framework for emotional voice conversion that effectively separates emotional from non-emotional features. Their model not only reduces audio distortion but also enhances data augmentation, improving Micro-F1 by 2% and Macro-F1 by 5%. Lastly, Moritani *et al.* [20] extend the capabilities of StarGAN by adapting it to emotional voice conversion for Japanese phrases. They perform subjective evaluations to assess neutrality and similarity, and also take the opportunity to investigate the inter-dependence between source and target emotional domains for conversion quality.

3. Storytelling Style Speech Generation System

3.1. SSPGS System Architecture

The SSPGS architecture of this study is depicted in **Figure 1** below, which is mainly divided into the upper half (1st Stage), Text-to-speech system (TTS), and the lower half (2nd Stage), emotional voice conversion module. The text-to-speech system enables users to input story text and send out neutral voices. The emotional voice conversion module then takes the neutral voices as input, converts them into expressive voices, and outputs them as needed for the particular story style.

Text-to-speech module is a text-to-speech acoustic model based on the Hidden Markov Model based speech synthesis system (HTS). It enables users to romanize Chinese sentences using the Chinese phoneme model and then uses the context-dependent label format to map each phoneme to the acoustic parameters (pitch, spectrum, duration, and intensity). First, the corresponding HMM is selected through the phoneme decision tree model, and then the HMM is merged. The sound can be played after the HTS module provides the MFCC/F0 parameters by first converting it to a raw or wav file through WORLD.

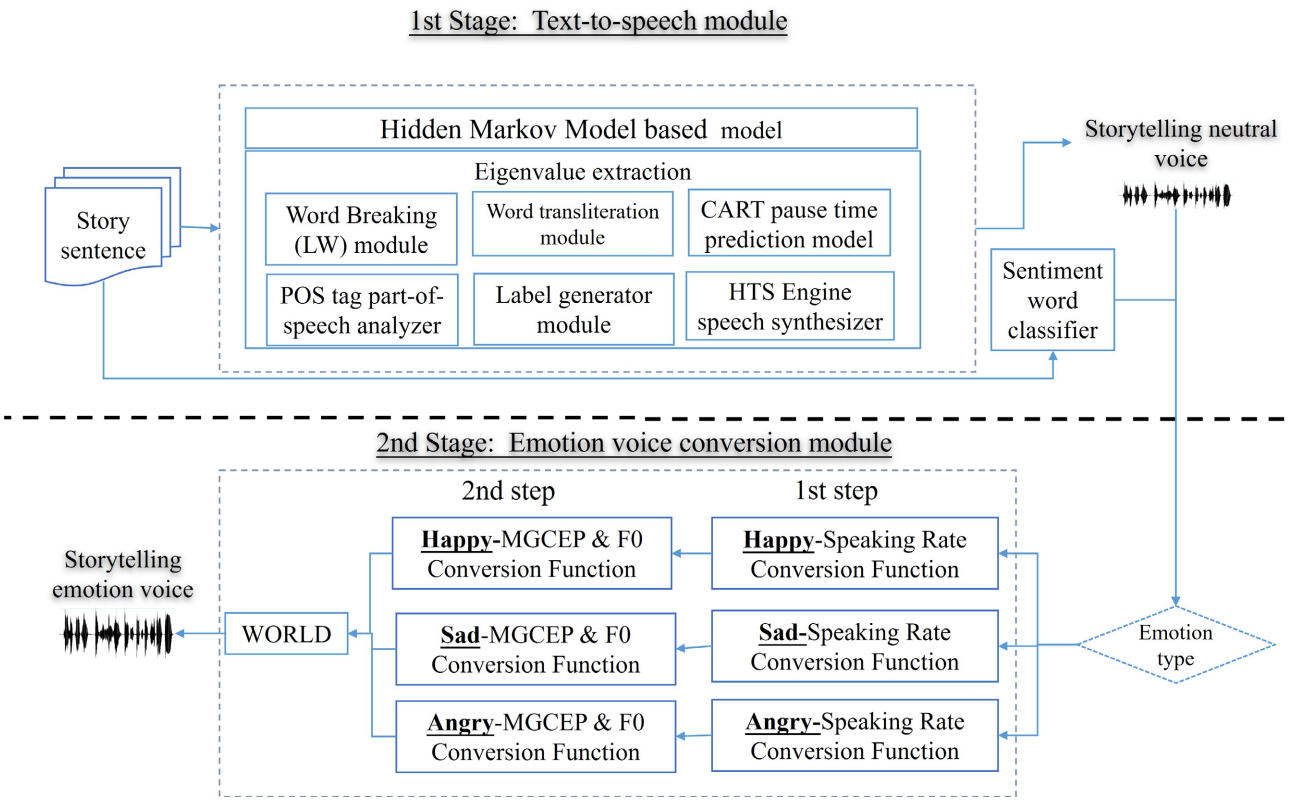

Figure 1. SSPGS system architecture.

Emotional voice conversion module is a two-step emotional voice conversion framework. It converts each type of emotion (Neutral-->Happy, Neutral-->Angry, Neutral-->Sad) to construct its corresponding emotion converter. The first step is to convert the neutral speech into the corresponding duration of the emotional speech (speech rate/Tempo conversion). Then, in the second step, the conversion of spectrum MGCEP and fundamental frequency F0 (pitch) is performed for the emotion, wherein MGCEP adopts 24-dimensional spectral parameters, F0 adopts 10-dimensional frequency F0, and adopts CycleGAN for training and conversion. Finally, the audio files with emotions are combined and produced. The objective indicators Average Mel-cepstral Distortion (MCD) and Root Mean Square Error (RMSE-F0) are used in conjunction with the parallel corpus method and the natural emotional voice to assess the similarity error.

3.2. Emotion Voice Conversion Module

The emotion voice conversion module is depicted in Figure 2 below. The upper part is the training phase, and the lower is the conversion phase. The input source for the training section uses a non-parallel corpus of neutral and emotional voice files that were both recorded by the same speaker, with one sentence serving as the units. For instance, 500 neutral tone files and 500 happy tone files (without the exact sentence text) can be used for input training to train a neutral->happy emotion conversion model. The input audio file is extracted through a speech encoder (such as WORLD/PML) to extract 24-dimensional Mel-generalized cepstral

coefficients (MGCEP), V/UV duration parameters, and one-dimensional frequency F0.

In the preprocessing of the parametric fundamental frequency F0, this research uses Continuous Wavelet Transform (CWT) to carry out high-dimensional extraction of the fundamental frequency F0. To create F0 in various time scales, it employs multiple multi-scale modeling to decompose the eigenvalues of 10 dimensions. The operation mode is primarily to give a mother wavelet $\psi(t)$. We can replace the sound frame t with t/a by 1. Scaling the mother wavelet $\psi(t)$, and 2. Translating the mother wavelet $\psi(t)$. We can replace t with t-b or t+b to obtain the rest of the sub-wavelet functions as the basis of wavelet transformation. The sub-wavelet functions are expressed as $\psi_{(a,b)}(t) = 1/\sqrt{a}\psi\left((t-b)/a\right)$. These wavelet functions must be orthonormal and have a finite amount of energy. Additionally, one can find out how to get Discrete Wavelet Transform through $\psi_{m,n}(t) = a_0^{m/2}\psi\left(a_0^{-m}t - nb_0\right)\ \ m,n \in Z$. The parameters a and b are expressed as follows: $\alpha = a_0^{-m}, b = nb_0a_0^{-m}$. The wavelet set can be obtained from them since m and n are integers. If $a_0 = 2$, $b_0 = 1$, the coefficient obtained by the wavelet function, and the inner product of the signal $f(t)$ is the wavelet coefficient, then it is CWT, $\omega_{a,b} = \langle\psi_{a,b}, f(t)\rangle = \int_{-\infty}^{\infty}\psi_{a,b}(t)f(t)\mathrm{d}t$, and the discrete wavelet coefficient can be obtained through: $\omega_{m,n} = \langle\psi_{m,n}(t), f(t)\rangle = a_0^{m/2}\int\psi\left(a_0^m(t) - nb_0\right)f(t)\mathrm{d}t$. Furthermore, the function $f(t)$ can be rebuilt from the wavelet coefficients back through $f(t) = \sum_m\sum_n\omega_{m,n}\psi_{m,n}(t)$.

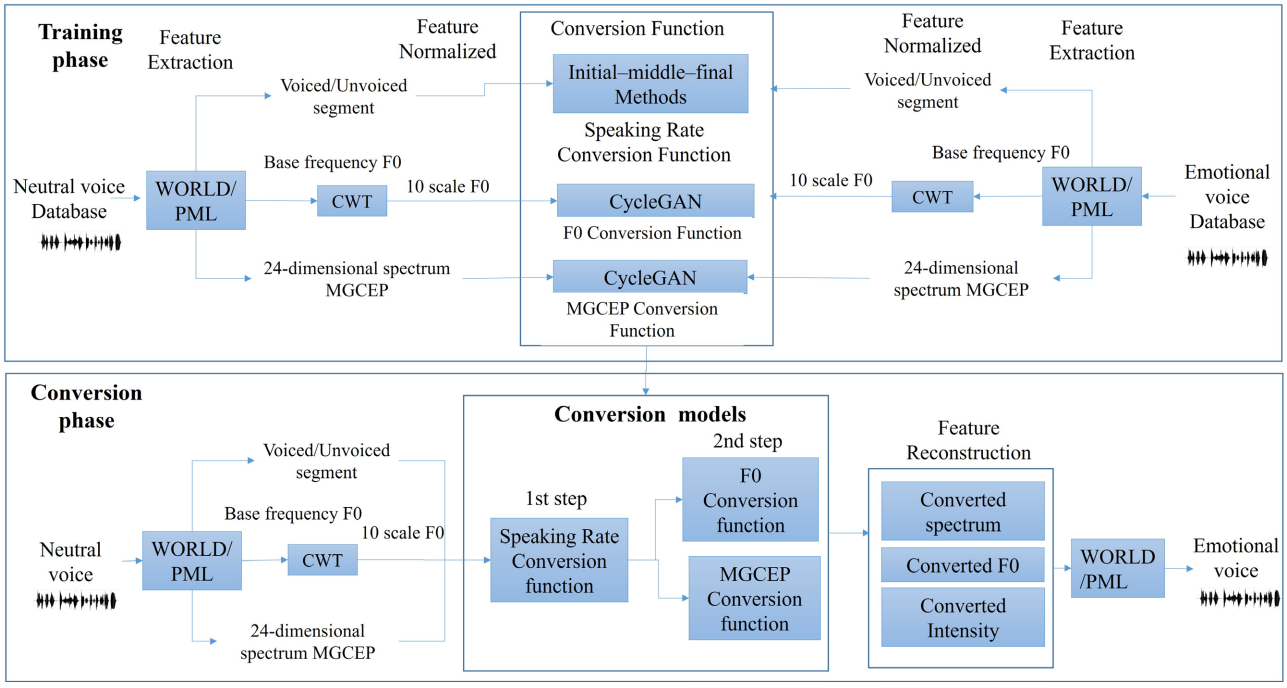

Figure 2. The structure of the emotion voice conversion module.

Three conversion models will be trained during training, including 1. Use ini-

tial–middle–final methods to develop a speech rate conversion function. First, use the eigen-values parameter to extract the V/UV duration made up of voiced and unvoiced segments. After that, compare the average voiced/unvoiced durations in a sentence's front, middle, and back for neutral and emotional speech. Finally, establish a linear mapping function of the front, middle, and back as a speech rate conversion model. 2. Use CycleGAN to establish a 10-dimensional frequency F0 conversion function (F0-CycleGAN Conversion Function). 3. 24-dimensional MGCEP conversion function can be created using CycleGAN (MGCEP-CycleGAN conversion function). We input the neutral voice to start conversion. First, the WORLD/PML encoder converts the signal into V/UV duration parameters, 1-dimensional frequency F0, and 24-dimensional MGCEP. To get a 10-dimensional frequency F0 from one of them, a 1-dimensional frequency F0 can be disassembled using CWT. First, we feed it into the speech rate conversion function. Then, we translate the number of sound frames with neutral speech duration into sound frames with emotional speech duration to produce the converted 24-dimensional MGCEP and 10-dimensional frequency F0 through F0-CycleGAN conversion function and MGCEP-CycleGAN conversion function. After the 10-dimensional frequency F0 is converted into a 1-dimensional frequency F0, we send it to the WORLD/PML encoder with the 24-dimensional MGCEP to combine and encode the emotional wave file. Additionally, a linear function is used in this study's volume (intensity) to establish a conversion transformation function. Since the procedure is fairly straightforward, a detailed description will not be provided.

3.2.1. First Step: Speech Rate Conversion Model

In the first step, that is, the conversion of speech rate, this study developed a general emotional speech rate conversion. It can be applied to the conversion structure of a non-parallel corpus. The fact that this method can convert data without the use of an additional automatic speech recognition module is by far its biggest benefit. Furthermore, because the various shifts in the Initial, middle, and final positions of the speech rate will create the rhythm of the emotion in the emotional sentences of the story, a factor that dynamically changes the speech rate with time is added. To complete the construction of this general model, it is combined with the Initial–middle–final method.

In the model training phase, each sentence of the parallel corpus, such as neutral and happy voice files, is first processed by voice signal preprocessing. Then, using PML's (pulse model in log domain vocoder) phase distortion deviation, each sentence's voiced and unvoiced segments are determined. They are referred to as V/UV as a whole. Next, the sound frame alignment of parallel corpus V/UV is carried out through Dynamic Time Warping. The durations are then examined in light of the V/Initial, UV's middle, and final positions within the sentence. Finally, a speech rate conversion model is created using the mean mapping ratio of V/UV in each position interval, as shown in **Figure 3** below.

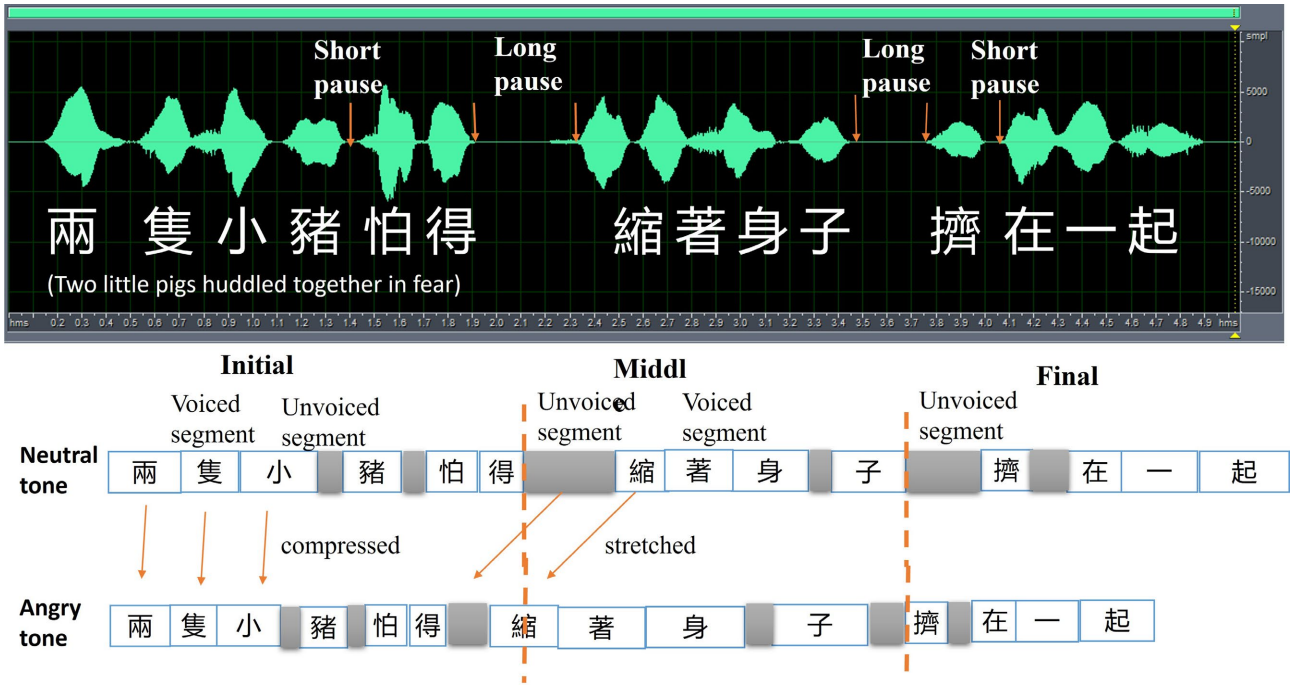

Figure 3. Initial-middle-final speech rate conversion method.

As part of the conversion process, the conversion model is used to calculate the voiced segment and unvoiced segment's shrinking/compressed, stretching, or kept parts. The speech frame sequence can then be increased and decreased using cubic spline interpolation to present the speech rate conversion. In the next stage, MGCEP and F0 conversion can be performed for the speech frame sequence after speech rate conversion.

3.3.2. Second Step: MGCEP and F0 Conversion Model

In this study, CycleGAN consists of two generators ($G_{x\to y}$ and $G_{y\to x}$) and two discriminators (D_y and D_x) as illustrated in **Figure 4** below. Taking X for neutral speech and Y for happy speech as an example, the target is to throw neutral and happy speech into training, so that these four neural networks ($G_{x\to y}$, $G_{y\to x}$, D_y, D_x) constituted CycleGAN, its loss function is minimized, and the overall objective formula is:

$$\begin{aligned} L_{full} &= L_{adv}\left(G_{x\to y}, D_y\right) + L_{adv}\left(G_{y\to x}, D_x\right) \\ &\quad + \lambda_{cyc} L_{cyc}\left(G_{x\to y}, G_{y\to x}\right) + \lambda_{id} L_{id}\left(G_{x\to y}, G_{y\to x}\right) \end{aligned} \tag{1}$$

The overall goals include 1). Adversarial loss: including forward $L_{adv}(G_{x\to y}, D_y)$ and reverse $L_{adv}(G_{y\to x}, D_x)$. Adversarial loss mainly measures the voice quality, that is, whether the converted voice resembles an emotional voice. 2). Cycle-consistency loss: $L_{cyc}(G_{x\to y}, G_{y\to x})$ is mainly to maintain the consistency of the contextual information of the input and output speech but also to prevent the generated speech being emotional. However, it is very different from the original sentence. 3). Identity-mapping loss: $L_{id}(G_{x\to y}, G_{y\to x})$ is mainly to maintain the linguistic-information consistency of input and output speech. The λ_{cyc} and λ_{id} represent trade-

off weight parameters. The adversarial loss function mainly includes two parts. It includes the adversarial function $L_{adv}(G_{x\to y}, D_y)$ formed by the generator $G_{x\to y}$ and the discriminator D_y of the neutral turning happy. It also includes the adversarial function $L_{adv}(G_{y\to x}, D_x)$ formed by the generator $G_{y\to x}$ and the discriminator D_x of happy to neutral. The neutral-to-happy adversarial function $L_{adv}(G_{x\to y}, D_y)$ is calculated as follows:

$$\begin{aligned} L_{adv}\left(G_{x\to y}, D_y\right) = {} & E_{y\sim PData}\left(y\right)\left[\log D_y\left(y\right)\right] \\ & + E_{x\sim PData}\left(x\right)\left[\log\left(1-D_y\left(G_{x\to y}\left(x\right)\right)\right)\right] \end{aligned} \tag{2}$$

$\mathrm{E}_{x\sim PData}(x)[log(1\text{-}\mathrm{D}_y(G_{x\to y}(x)))]$ denotes the happy voice $\hat{Y}$ generated by the neutral voice(x) through generator $G_{x\to y}$. In the discriminator D_y, the Y score is closer to 0, the better, indicating that the discriminator D_y can distinguish between real happy voice and generated happy voice in the data. $\mathrm{E}_{y\sim PData}(y)[logD_y(y)]$ means that the real happy voice data is thrown into the discriminator D_y, and the closer the discriminator D_y identifies the score to 1, the better. For the generator $G_{x\to y}$, the goal is to hope that the score of the synthesized happy voice $\mathrm{D}_y(G_{x\to y}(x))$ is as close to 1 as possible. Conversely, for the adversarial function of happy to neutral $L_{adv}(G_{y\to x}, D_x)$ is the same logic.

In cycle-consistency loss, to ensure that the neutral voice and the converted happy voice maintain consistent contextual information (same sentence), the neutral voice x is generated through the generator $G_{x\to y}$ to generate the happy voice $\hat{Y}$ and then restore the neutral speech x^1 through the generator $G_{y\to x}$. Then, we calculate the gap between the neutral voice x and the restored voice x^1. The better, the smaller the distance. Contrarily, the loss of happiness to neutral follows the same logic as the cycle-consistency. The formula is as follows:

$$\begin{aligned} L_{cyc}\left(G_{x\to y}, G_{y\to x}\right) = {} & E_{x\sim PData}\left(x\right)\left[\left\|G_{y\to x}\left(G_{x\to y}\left(x\right)\right)-x\right\|1\right] \\ & + E_{y\sim PData}\left(y\right)\left[\left\|G_{x\to y}\left(G_{y\to x}\left(y\right)\right)-y\right\|1\right] \end{aligned} \tag{3}$$

The goal of peer-to-peer mapping is that when the input is real data (emotional voice) to the generator, the output is still real data (emotional voice). It means that the voice information from the original voice won't be ignored, distorted, or retained by the generator. It is to ensure that the neutral voice of "Come on" will not produce the emotional voice of "How are you" (somewhat similar to supervised learning. the input of the discriminator must be a pair, including the output of the generator and the input linguistic-information. The formula is as follows:

$$\begin{aligned} L_{id}\left(G_{x\to y}, G_{y\to x}\right) := {} & E_{y\sim PData}\left(y\right)\left[\left\|G_{x\to y}\left(y\right)-y\right\|1\right] \\ & + E_{x\sim PData}\left(x\right)\left[\left\|G_{y\to x}\left(x\right)-x\right\|1\right] \end{aligned} \tag{4}$$

This study focuses on the network structure design of two generators ($G_{x\to y}$ and $G_{y\to x}$) and two discriminators (D_y and D_x). The generator uses a one-dimensional convolutional neural network (1D CNN). It includes down-sampling, residual learning, and up-sampling layers, while the discriminator uses a 2D convolutional

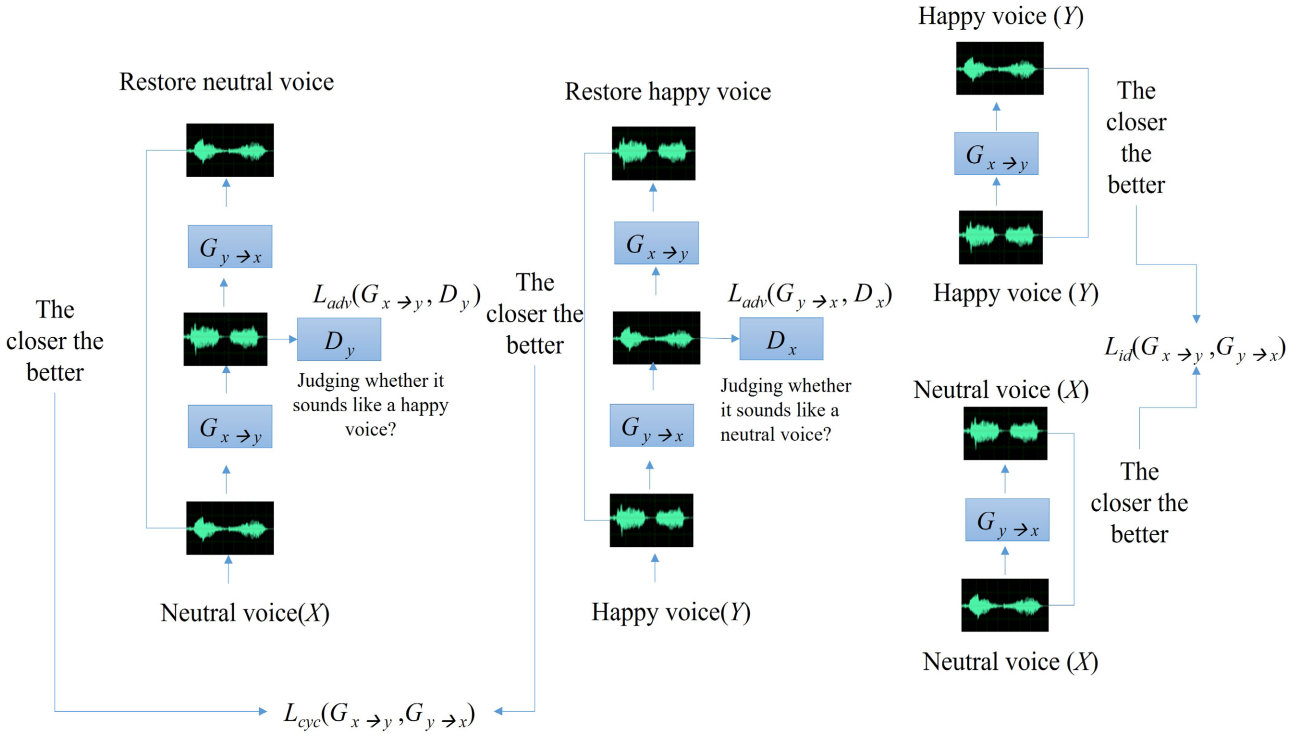

Figure 4. CycleGAN structure of emotional speech conversion.

neural network. **Figure 5** depicts the design in detail. The generator's one-dimensional convolutional neural network is designed using a gated-1D CNN, and the convolutional layer design is based on work by [15]. Its structure is a deep neural network containing many different modules. In the input layer and input layer design (as shown in the gray block), the upper h, w, and c represent height, width, and channel. In the convolution layer, the variables k, c, and s stand for the size of the filter or kernel used in the convolution operation, the number of channels, and the stride length, respectively. The generator uses one-dimensional convolution so that both parameters will be $1*n$, $n \in N$. For example, $h = 1$, $w =$ T, $c = 24$ on the input mean that there will be a total of T frames of 24-dimensional Mel cepstral coefficients. So, it can be regarded as a feature with a height of 1, a width of T, and 24 channels. In the second convolutional layer, $k = 1 \times 15$ means that the filter refers to 15 sound frames at a time, and $c = 256$ means that there are 256 filters in total. So, the number of channels of the resulting feature will be 256, and $s = 1 \times 2$ represents that the filter convolution operation spans two frames simultaneously.

In terms of number setting, the weight λ_{cyc} is 10, L_{id} sets the weight λ_{id} to 5 in the first 10,000 training sessions, and 0 after that. Adam optimizer is used, and the batch size is set to 1. The generator's learning rate is set to 0.0002, the discriminator is set to 0.0001, the momentum term β is set to 0.5, the first 200,000 iterations are kept the same, and the number of iterations decreases linearly after each 200,000. This study does not require time alignment and makes use of a non-parallel corpus.

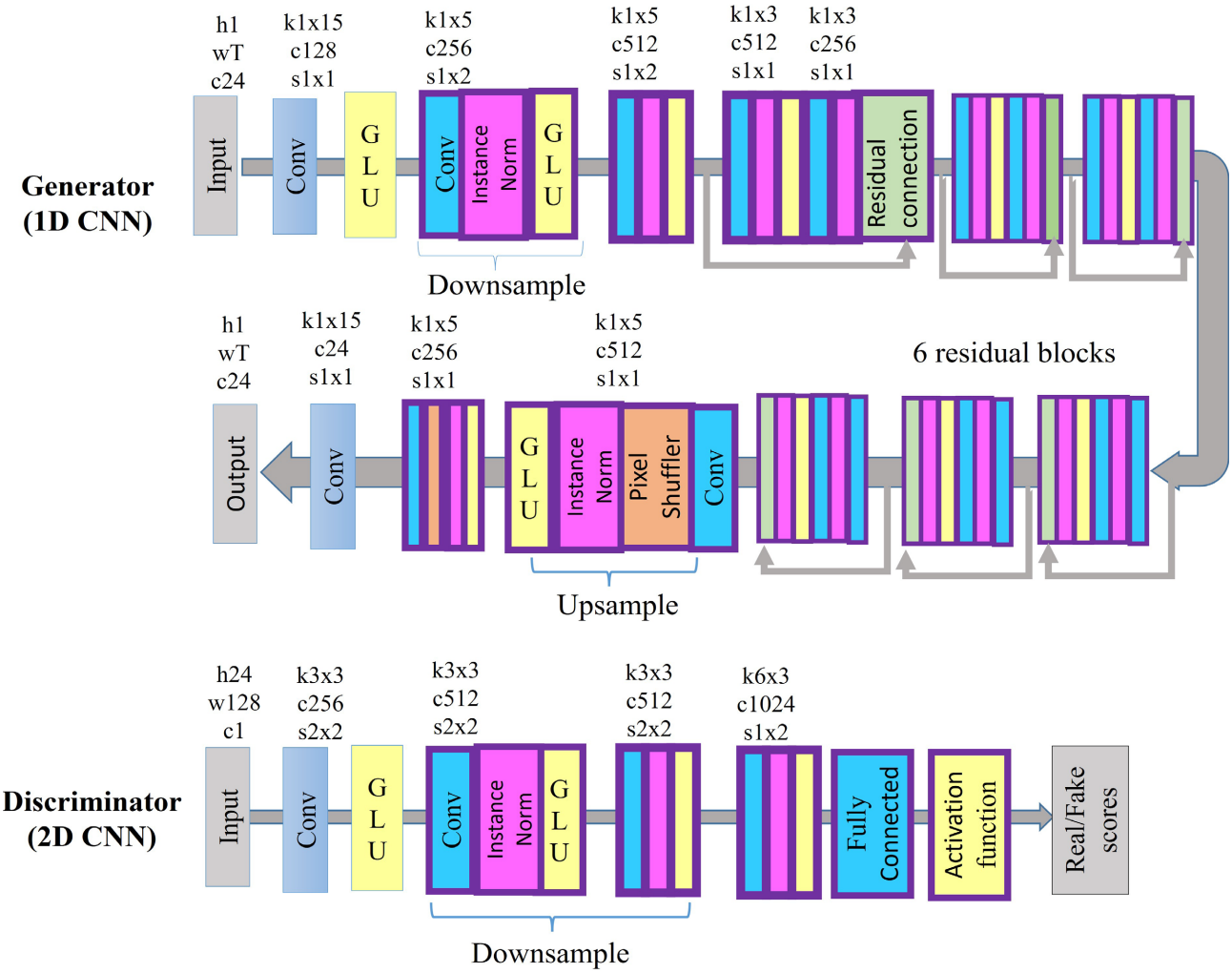

Figure 5. CNN design architecture for generator and discriminator.

4. Experimental Design and Results

4.1. Corpus Construction

In this research, Chinese voice models of three characters were produced: a cute little girl, a clever little boy, and a neutral narrator. Since this study is aimed at children's storytelling, the timbres that are similar to the voices of cartoon characters will be chosen in the selection of timbres. Table 1 below displays each character's pitch and speech rate. We selected professional voice actors with professional recording studios for emotional voice recording in the recording corpus because this study proposes a technical demonstration that can be imported into the story robot. It is hoped that the recording quality can be optimized to avoid issues such as inconsistent pauses in words caused by the unstable situation of the recorder, the speed of speech fluctuating quickly and slowly, incorrect pronunciation, excessive breath sounds, and the generation of legato or ambient sound interference. These issues may make it impossible to distinguish between different algorithm optimizations in the follow-up, which would prevent the student re-

cording method from being used.

The recording cost of a professional voice actor can reach up to 4 Taiwan region dollars per word. As the training corpus increases with the recording of more voices, the quality of the voice should also improve. However, an excessive number of identical corpus recordings will raise the price of recording and manual transcription while gradually lowering the marginal benefit of voice quality improvement. The balanced corpus can cover about 1340 Chinese pronunciation units on average by recording fewer sentences. The creation of a balanced corpus can cut down on the number of sentences that must be recorded while maintaining the same voice quality. It can also lower the cost of creating future multi-role voices.

This study uses a set of balanced corpus integrals to determine the optimal number of sentences to be recorded. The obtained 200 story text files include Hans Christian Andersen, Grimm's fairy tales, Aesop's fables, Hsin Yi's self-made stories, etc., totaling about 3000 sentences. The average 500 sentences covering 1340 pronunciation units are screened out by calculating the balance corpus integral. The voice actor for a character will record 500 sentences in each of the following tones: neutral for neutral sentences, happy for happy sentences, sad for sad sentences, and angry for angry sentences. The number of sentences in our database is comparable to the size of emotional databases used in past studies. These include the German emotional database EmoDB, the Indian emotional database IITKGP in Telugu, and the English emotional database SAVEE.

Table 1. Acoustic characteristics and corpus recordings of the three characters.

Character name	Imitated cartoon character	Pitch range	Speech rate (seconds/word)
Cute little girl	Friendly: Sounds like Magical DoReMi	150~550 Hz	Fast 0.25
Clever boy	Smart/talented: Similar to Kenichi Ono's voice in Chibi Maruko Chan	140~450Hz	Slow 0.35
Neutral narration	Open, similar to the voice of Jin in Robocar Poli	200~400 Hz	Moderate 0.3

4.2. Experimental Designs

This study performs speech synthesis on three multi-character and emotionally rich children's stories, including Going to the Beach Together, A Little Match Girl, and Watermelon Girl Chooses the Groom. The average story length is about 2~3 minutes; each story is about 20~30 sentences, as shown in Table 2 below. This study chose Hsin Yi Family Square in Taipei, Taiwan region because parents frequently bring their young children there to play and read.

In this research, 25 pairs of children aged 4 to 8 and their parents were recruited for field studies. Each pair, consisting of a parent and their child, served as a unit. The experiment was conducted with one such unit at a time and lasted for approximately 15 minutes. The child's gender, age, and whether or not their home has a story robot were all filled out at the beginning after we explained the test procedure, obtained their consent, and signed the consent form for the video and

personal information. We prepared a 10-inch flat-panel story robot on site. The tablet story robot includes 9 story voice files and recorded 3 stories with real-person dubbing, google TTS speech synthesis, B company TTS speech synthesis, and speech synthesizer by this study. The tablet interface has 12 selection buttons for recording files, with next, previous, pause, and volume control buttons. This study measured immersion for the young children of those parents after designing a MOS evaluation for those parents.

In this study, four emotions are categorized for each sentence in the story. The emotions include happy, sad, angry, and neutral. The emotional voices that the sentence should be changed into are determined by averaging the scores, as seen in Table 2. After that, we use the character voice model and emotional voice conversion module to convert each character's emotional sentences, after which we combine all of the sentences into a story. As seen from Table 2, Story 1, Going to the Beach Together, has the happiest sentences in the story and is a happy story. Story 2, A Little Match Girl, has the saddest sentences in the story, which is a sad story. Story 3, Watermelon Girl Chooses the Groom, has the angriest sentences in the story and is an angry story. This study investigates the effects of emotional speech synthesis in this system using various emotion-oriented stories.

Table 2. Test stories.

Story name	Character	Number of sentences	Story length (minutes)	Happy/Angry/Sad tone (sentences)
Going to the Beach Together	Narrator, rabbit, goat	16	2	9/2/1
A Little Match Girl	Narrator, little girl, boss	40	4	3/5/10
Watermelon Girl Chooses the Groom	Narrator, Watermelon Girl, Mr. Papaya	30	3	2/8/3

The 5-point Mean Opinion Score (MOS) is to evaluate the difference in the scores of parents on naturalness, word error rate, and preference for the same story, using google TTS speech synthesis, B company TTS speech synthesis, and SSPGS synthesis. It is a 5-point scale, the questionnaire design is depicted in Annex, and the story files are given randomly. The way to assess story immersion is to ask children to comment on the differences in engagement, liking, and empathizing between real-person dubbing and voice synthesizer in this study for the same story.

In terms of evaluation procedures, this study refer to Costa' questionnaire items [36] and evaluation procedures. In addition to using questionnaires, the evaluation also includes observation (video recording). Only one story was read to each child (two audio files). During the storytelling process, the observers made observations and videos at a distance of about 2 to 3 meters. The observers would approach for an interview right away after hearing a story file and request to fill out a questionnaire. Following the conclusion of the experiment for the day, the observers would document the children's body language and facial expressions as they watched the story in the video to further support the measurement of story engagement, liking, and empathy.

The questionnaire consists of six items, measuring engagement (Question 1 & Question 4), Liking (Question 3 & Question 6), and empathizing (Question 2 & Question 5). Each question is on a seven-point Likert scale, with 1 denoting strong disagreement and 7 denoting strong agreement.

4.3. Experimental Results

4.3.1. 5-Point Mean Opinion Score (MOS)

The MOS of this study are depicted in Table 3 below. SSPGS has good performance in speech naturalness, word error rate, and preference. First of all, we can see that the three speech synthesis systems all achieved scores above 4 for storytelling voice comprehension. Miss Google outperformed the others, scoring 4.2 points for storytelling voice intelligibility. The fact that everyone is accustomed to Miss Google's voice is one of the main causes. Although this study is based on the voice of A, B, and C systems, almost all respondents can recognize Miss Google, and some expressed some intimacy. Second, in the speech naturalness score, SSPGS has the highest score, reaching 4.2 points, while Miss Google has the lowest score, only 3.58 points. We can see that even though Google has the highest word error rate score in the application of storytelling speech synthesis, storytelling is not appropriate because Google uses an anchor style. Finally, in the preference score, it is evident that SSPGS conducted the best, reaching 4.45 points, higher than Google's 3.4 points and B company's 3.9 points. It demonstrates that using the voices of cartoon characters to create a voice model can produce effective results.

Table 3. MOS evaluation results.

Voice system	Company B			SSPGS			Google		
Indicator	Word error rate	Naturalness	Preference	Word error rate	Naturalness	Preference	Word error rate	Naturalness	Preference
Average score	4.03	4.08	3.9	4.08	4.21	4.45	4.2	3.58	3.4
Standard deviation	0.718	0.918	0.852	0.768	0.688	0.759	0.649	0.513	0.503

4.3.2. Story Immersion Measurement

The immersion measurement result is illustrated in Table 4. We can see that SSPGS performs better than real-person dubbing in terms of engagement, liking, and empathizing for stories with more upbeat sentences (Going to the Beach Together). However, in the comparison of stories with more sad sentences (A Little Match Girl), the scores were lower than those of real-person dubbing in terms of engagement, liking, and empathizing. There was no difference in the engagement, Liking, and empathizing scores in the story (Watermelon Girl Chooses the Groom) contained more angry sentences, according to an analysis of variance.

This study employed paired *t*-tests to examine whether there were significant differences in story immersion among children when listening to stories narrated by real-person dubbing and SSPGS across three different emotional story types: happy, sad, and angry.

For the happy story type (*Going to the Beach Together*), the paired *t*-test revealed a significant difference in immersion levels between real-person dubbing and SSPGS ($p < 0.001$). A possible reason why SSPGS performed better is that the three voice models used in this study were trained on speech data recorded by professional voice actors imitating popular cartoon characters such as DoReMi from Magical DoReMi, Kenichi Ono from Chibi Maruko Chan, and Jin from Robocar Poli. These characters are well-known and favored by children aged 4 – 8, and their voices are often associated with happy emotions in cartoons. In contrast, the real-person dubbing was performed by general voice actors, whose voices were less familiar to children. This difference in familiarity may have contributed to the significant variation in immersion levels.

For the sad story type (*A Little Match Girl*), the paired *t*-test also revealed a significant difference in immersion levels between real-person dubbing and SSPGS ($p < 0.05$). A possible explanation for SSPGS's lower performance is that sad stories often contain crying sounds, such as sobbing or choking voices, which SSPGS struggled to synthesize naturally. In some cases, these artificially generated sobbing effects sounded unnatural or even distorted, making them distracting or unpleasant for children, thereby reducing immersion. In contrast, real-person dubbing naturally conveyed emotional variations in crying voices, allowing children to genuinely feel the sadness in the story.

For the angry story type (*Watermelon Girl Chooses a Groom*), the paired *t*-test found no significant difference in immersion levels between real-person dubbing

Table 4. Immersion measurement results.

	Real-person dubbing			SSPGS		
Question item	Story1: Going to the Beach Together	Story2: A Little Match Girl	Story3: Watermelon Girl Chooses the Groom	Story1: Going to the Beach Together	Story2: A Little Match Girl	Story3: Watermelon Girl Chooses the Groom
Q1: The story is interesting.	4.6 (1.67)	4.6 (1.67)	4.6 (1.59)	5.0 (2.0)	3.8 (1.79)	4.4 (1.89)
Q2: At the beginning of the story, I feel sorry for the little match girl.	4.2 (1.95)	4.4 (1.95)	4.3 (1.72)	4.0 (2.0)	2.2 (1.64)	3.1 (1.82)
Q3: I enjoy listening to the stories.	4.4 (1.30)	3.8 (1.30)	4.1 (1.49)	4.8 (2.17)	3.2 (1.30)	4.0 (1.74)
Q4: I really want to know how this story ends.	5.2 (2.35)	5.0 (2.35)	5.1 (1.59)	5.6 (1.14)	4.0 (1.87)	4.8 (1.51)
Q5: At the end of the story, I'm happy for the little match girl.	5.4 (1.30)	4.8 (1.30)	5.1 (1.22)	6.4 (1.34)	4.4 (1.95)	5.4 (1.65)
Q6: I like this story.	4.8 (1.67)	4.6 (1.67)	4.7 (1.98)	5.6 (1.52)	4.2 (2.17)	4.9 (1.84)
Liking(Q3 + Q6)	4.6 (1.67)	4.2 (1.48)	4.4 (1.67)	5.2 (1.81)	3.7 (1.77)	4.5 (1.90)
Engagement(Q1 + Q4)	4.9 (1.20)	4.8 (1.93)	4.9 (1.57)	5.3 (1.57)	3.9 (1.73)	4.6 (1.76)
Empathizing (Q3 + Q6)	4.8 (1.40)	4.6 (1.58)	4.7 (1.45)	5.2 (2.04)	3.3 (2.06)	4.3 (2.22)

and SSPGS ($p > 0.05$). During the experiment, we observed polarized reactions among children toward angry speech. Through interviews with their parents, we found that some children were uncomfortable with angry tones, especially if they had negative associations with hearing their parents or teachers express anger. This factor may have interfered with the effect of voice type on immersion levels, diminishing the potential differences between SSPGS and real-person dubbing. Based on these findings, future research may consider using a surprised tone instead of an angry tone to minimize external biases and better assess the impact of voice synthesis on immersion.

5. Conclusions

This study proposed a Chinese storytelling Style speech generation system composed of a text-to-speech system and emotional voice conversion module and developed three role-neutral speech models. The voice model includes a female voice (as a narrator), a cute girl voice (female protagonist), and a low-pitched boy voice (male protagonist). As a result, using a CycleGAN, we were able to create an emotional voice conversion module without the need for a parallel corpus. The models include neutral→happy, neutral→angry, and neutral→sad models, a total of 9 transformation models to achieve the happy, angry, and sad voices most often needed in stories.

This study model is tested on the ground with children and parents. The system performed well in MOS's naturalness, word error rate, and preference. Additionally, there is no discernible difference between listening to the story and real-person dubbing in terms of engagement, liking, and empathy when the analysis of the difference in story immersion is done.

5.1. Practical Contribution

This study aims at the high cost, and time-consuming real-person dubbing pain points domestic story robot manufacturers face. Furthermore, it investigates the possibility of applying speech synthesis to story robots. As a result, this study created an SSPGS with a function for converting emotional speech. The study also pioneered the use of parent-child demonstrations with young children as the subject to demonstrate the distinction between the real-person voice dubbing currently used and the story voice synthesized by this system. Finally, it attempted to address the high cost and time-consuming problem of real-person dubbing. Additionally, it offered voice-over services to children's audiobook publishers so they could use speech synthesis to import stories.

To construct a voice model for a single character in this study, professional voice actors were required to record 300 sentences per emotion: neutral, happy, sad, and angry, totaling 1200 sentences. The final dataset consisted of approximately 4800 words. Given that the cost of professional voice recording is $0.15 per word, the total cost of recording amounts to $720. Additionally, the annotation of 4800 words requires a professional annotator working for seven days, leading to

an estimated cost of $1000. Furthermore, the training and fine-tuning of the voice model incur an additional $1000 to $1200.

As a result, the estimated total cost for developing a single character's voice model is around $3000. Considering that a single story typically consists of three characters, the cost to fully develop voice models for one story amounts to approximately $9000. However, once these character voice models are established, they can be utilized for an unlimited number of stories without additional recording costs.

In contrast, if human voice actors were to record each story manually, a budget of $9000 would allow for the recording of approximately 60000 words. Given that each story consists of around 600 words, this budget would only cover 100 stories. Thus, in terms of cost efficiency, the proposed model in this study demonstrates a significant advantage over traditional human voice recordings, making it a viable alternative for scaling up content production in storytelling applications.

5.2. Academic Contribution

This study is the first to develop a speech synthesis and conversion model for the Chinese language family in the context of storytelling and to investigate the feasibility of its speech quality. Based on the established technology of cycle generative adversarial networks, an emotional voice conversion model was developed. Taking advantage of CycleGAN's ability to model without recording parallel corpora, an emotional voice conversion framework is proposed. The framework is an emotional speech conversion model that can transform the timbre (spectrum), pitch (fundamental frequency F0), and speech rate required by emotional speech under the condition of no parallel corpus. Additionally, it makes it possible to change neutral speech into happy, angry, and sad emotions, which exacerbates the issue of a fixed speech rate/rhythm in CycleGAN-based emotional speech conversion regardless of emotion.

5.3. Future Work

This study successfully developed a Chinese storytelling-style speech generation system and evaluated its performance against Google TTS and a commercial system. While these comparisons provide valuable insights, incorporating StarGAN could further enhance the evaluation of our system's performance. StarGAN allows for multi-emotion conversion using a single model, while CycleGAN requires training separate models for each emotional transformation. However, StarGAN's shared feature learning approach can sometimes lead to blended or inconsistent emotions, whereas CycleGAN, by training dedicated models, ensures more stable and distinct emotional expression. Additionally, StarGAN typically requires a larger, well-annotated dataset to achieve optimal performance, whereas CycleGAN is more adaptable to smaller datasets, making it more practical for the limited availability of high-quality Chinese emotional speech corpora. Moreover, StarGAN's higher computational cost can make real-time speech synthesis more

challenging, while CycleGAN's emotion-specific models offer better processing efficiency, making it a more suitable choice for real-time storytelling applications.

While CycleGAN has demonstrated strong performance in this study, we recognize the potential advantages of StarGAN and plan to conduct a direct comparative study in the future. This evaluation will focus on speech naturalness, emotional expressiveness, computational efficiency, and the ability to transition smoothly between multiple emotions. In summary, CycleGAN was chosen for its stability, efficiency, and adaptability to limited datasets, making it the most suitable framework for this study. However, future research will investigate StarGAN's potential to further enhance emotional speech synthesis, ensuring the system meets the highest standards of expressiveness, efficiency, and real-world applicability.

Acknowledgements

The authors would like to thank the National Science Council of the Republic of China, Taiwan region for financially supporting this research under Contract No. 110-2410-H-035-011-.

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

References

[1] Leite, I., McCoy, M., Lohani, M., Ullman, D., Salomons, N., Stokes, C., *et al.* (2015) Emotional Storytelling in the Classroom: Individual versus Group Interaction between Children and Robots. *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, Portland, 2-5 March 2015, 75-82. https://doi.org/10.1145/2696454.2696481

[2] Sarkar, P. and Rao, K.S. (2015) Modeling Pauses for Synthesis of Storytelling Style Speech Using Unsupervised Word Features. *Procedia Computer Science*, **58**, 42-49. https://doi.org/10.1016/j.procs.2015.08.007

[3] Sisman, B., Yamagishi, J., King, S. and Li, H. (2021) An Overview of Voice Conversion and Its Challenges: From Statistical Modeling to Deep Learning. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **29**, 132-157. https://doi.org/10.1109/taslp.2020.3038524

[4] Luo, Z., Chen, J., Takiguchi, T. and Ariki, Y. (2017) Emotional Voice Conversion Using Neural Networks with Arbitrary Scales F0 Based on Wavelet Transform. *EURASIP Journal on Audio, Speech, and Music Processing*, **2017**, Article No. 18. https://doi.org/10.1186/s13636-017-0116-2

[5] Luo, Z., Chen, J., Takiguchi, T. and Ariki, Y. (2019) Neutral-to-Emotional Voice Conversion with Cross-Wavelet Transform F0 Using Generative Adversarial Networks. *APSIPA Transactions on Signal and Information Processing*, **8**, e10. https://doi.org/10.1017/atsip.2019.3

[6] Luo, Z., Chen, J., Takiguchi, T. and Ariki, Y. (2019) Emotional Voice Conversion Using Dual Supervised Adversarial Networks with Continuous Wavelet Transform F0 Features. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **27**, 1535-1548. https://doi.org/10.1109/taslp.2019.2923951

[7] Xue, Y., Hamada, Y. and Akagi, M. (2018) Voice Conversion for Emotional Speech: Rule-Based Synthesis with Degree of Emotion Controllable in Dimensional Space. *Speech Communication*, **102**, 54-67. https://doi.org/10.1016/j.specom.2018.06.006

[8] Vekkot, S., Gupta, D., Zakariah, M. and Alotaibi, Y.A. (2020) Emotional Voice Conversion Using a Hybrid Framework with Speaker-Adaptive DNN and Particle-Swarm-Optimized Neural Network. *IEEE Access*, **8**, 74627-74647. https://doi.org/10.1109/access.2020.2988781

[9] Elgaar, M., Park, J. and Lee, S.W. (2020) Multi-Speaker and Multi-Domain Emotional Voice Conversion Using Factorized Hierarchical Variational Autoencoder. *ICASSP* 2020-2020 *IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), Barcelona, 4-8 May 2020, 7769-7773. https://doi.org/10.1109/icassp40776.2020.9054534

[10] Kim, T., Cho, S., Choi, S., Park, S. and Lee, S. (2020) Emotional Voice Conversion Using Multitask Learning with Text-to-Speech. *ICASSP* 2020-2020 *IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), Barcelona, 4-8 May 2020, 7774-7778. https://doi.org/10.1109/icassp40776.2020.9053255

[11] Jia, N., Zheng, C. and Sun, W. (2019) A Model of Emotional Speech Generation Based on Conditional Generative Adversarial Networks. 2019 11*th International Conference on Intelligent Human-Machine Systems and Cybernetics* (*IHMSC*), Hangzhou, 24-25 August 2019, 106-109. https://doi.org/10.1109/ihmsc.2019.00033

[12] Kameoka, H., Tanaka, K., Kwasny, D., Kaneko, T. and Hojo, N. (2020) ConvS2S-VC: Fully Convolutional Sequence-to-Sequence Voice Conversion. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **28**, 1849-1863. https://doi.org/10.1109/taslp.2020.3001456

[13] Singh, J.B. and Lehana, P. (2017) Straight-Based Emotion Conversion Using Quadratic Multivariate Polynomial. *Circuits, Systems, and Signal Processing*, **37**, 2179-2193. https://doi.org/10.1007/s00034-017-0660-0

[14] Robinson, C., Obin, N. and Roebel, A. (2019) Sequence-to-Sequence Modelling of F0 for Speech Emotion Conversion. *ICASSP* 2019-2019 *IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), Brighton, 12-17 May 2019, 6830-6834. https://doi.org/10.1109/icassp.2019.8683865

[15] Kaneko, T. and Kameoka, H. (2018) CycleGAN-VC: Non-Parallel Voice Conversion Using Cycle-Consistent Adversarial Networks. 2018 26*th European Signal Processing Conference* (*EUSIPCO*), Rome, 3-7 September 2018, 2100-2104. https://doi.org/10.23919/eusipco.2018.8553236

[16] Asakura, T., Akama, S., Shimokawara, E., Yamaguchi, T. and Yamamoto, S. (2019) Emotional Speech Generator by Using Generative Adversarial Networks. *Proceedings of the Tenth International Symposium on Information and Communication Technology*, Hanoi, 4-6 December 2019, 9-14. https://doi.org/10.1145/3368926.3369662

[17] Zhou, K., Sisman, B., Zhang, M. and Li, H. (2020) Converting Anyone's Emotion: Towards Speaker-Independent Emotional Voice Conversion. *Interspeech* 2020, Shanghai, 25-29 October 2020, 3416-3420. https://doi.org/10.21437/interspeech.2020-2014

[18] Rizos, G., Baird, A., Elliott, M. and Schuller, B. (2020) Stargan for Emotional Speech Conversion: Validated by Data Augmentation of End-to-End Emotion Recognition. *ICASSP* 2020-2020 *IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), Barcelona, 4-8 May 2020, 3502-3506. https://doi.org/10.1109/icassp40776.2020.9054579

[19] He, X., Chen, J., Rizos, G. and Schuller, B.W. (2021) An Improved Stargan for Emotional Voice Conversion: Enhancing Voice Quality and Data Augmentation. *Interspeech* 2021, Brno, 30 August-3 September 2021, 821-825. https://doi.org/10.21437/interspeech.2021-1253

[20] Moritani, A., Sakamoto, S., Ozaki, R., Kameoka, H. and Taniguchi, T. (2021) StarGAN-Based Emotional Voice Conversion for Japanese Phrases. *Proceedings of the* 2021 *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference* (*APSIPA ASC*), Tokyo, 14-17 December 2021, 836-840.

[21] Theune, M., Meijs, K., Heylen, D. and Ordelman, R. (2006) Generating Expressive Speech for Storytelling Applications. *IEEE Transactions on Audio*, *Speech and Language Processing*, **14**, 1137-1144. https://doi.org/10.1109/tasl.2006.876129

[22] Kaliyev, A., Rybin, S.V., Matveev, Y.N., Kaziyeva, N. and Burambayeva, N. (2018) Modeling Pause for the Synthesis of Kazakh Speech. *Proceedings of the Fourth International Conference on Engineering & MIS* 2018, Istanbul, 19-20 June 2018, 1-4. https://doi.org/10.1145/3234698.3234699

[23] Putra Negara, A.B., Magdalena, Y., Nyoto, R.D. and Sujaini, H. (2019) Chunking Phrase to Predict Pause Break in Pontianak Malay Language. *Lontar Komputer: Jurnal Ilmiah Teknologi Informasi*, **10**, 128-139. https://doi.org/10.24843/lkjiti.2019.v10.i03.p01

[24] Negara, A.B.P., Muhardi, H. and Muniyati, E.F. (2020) Prediksi Jeda dalam Ucapan Kalimat Bahasa Melayu Pontianak Menggunakan Hidden Markov Model Berbasis Part of Speech. *Jurnal Teknologi Informasi dan Ilmu Komputer*, **7**, 755-764. https://doi.org/10.25126/jtiik.2020742166

[25] Ramli, I., Jamil, N., Seman, N. and Ardi, N. (2016) An Improved Pitch Contour Formulation for Malay Language Storytelling Text-to-Speech (TTS). 2016 *IEEE Industrial Electronics and Applications Conference* (*IEACon*), Kota Kinabalu, 20-22 November 2016, 250-255. https://doi.org/10.1109/ieacon.2016.8067387

[26] Ramli, I., Seman, N., Ardi, N. and Jamil, N. (2016) Rule-Based Storytelling Text-to-Speech (TTS) Synthesis. *MATEC Web of Conferences*, **77**, Article 04003. https://doi.org/10.1051/matecconf/20167704003

[27] Ramli, I., Jamil, N., Seman, N. and Ardi, N. (2018) The First Malay Language Storytelling Text-to-Speech (TTS) Corpus for Humanoid Robot Storytellers. *Journal of Fundamental and Applied Sciences*, **9**, 340-358. https://doi.org/10.4314/jfas.v9i4s.20

[28] Sarkar, P. and Rao, K.S. (2015) Data-Driven Pause Prediction for Synthesis of Storytelling Style Speech Based on Discourse Modes. 2015 *IEEE International Conference on Electronics*, *Computing and Communication Technologies* (*CONECCT*), Bangalore, 10-11 July 2015, 1-5. https://doi.org/10.1109/conecct.2015.7383906

[29] Sarkar, P. and Rao, K.S. (2015) Analysis and Modeling Pauses for Synthesis of Storytelling Speech Based on Discourse Modes. 2015 *Eighth International Conference on Contemporary Computing* (*IC3*), Noida, 20-22 August 2015, 225-230. https://doi.org/10.1109/ic3.2015.7346683

[30] Sarkar, P. and Sreenivasa Rao, K. (2015) Data-Driven Pause Prediction for Speech Synthesis in Storytelling Style Speech. 2015 *Twenty First National Conference on Communications* (*NCC*), Mumbai, 27 February-1 March 2015, 1-5. https://doi.org/10.1109/ncc.2015.7084924

[31] Chen, Y., Wu, C., Huang, Y., Lin, S. and Wang, J. (2016) Candidate Expansion and Prosody Adjustment for Natural Speech Synthesis Using a Small Corpus. *IEEE/ACM Transactions on Audio*, *Speech*, *and Language Processing*, **24**, 1052-1065. https://doi.org/10.1109/taslp.2016.2537982

[32] Huang, Y., Wu, C. and Weng, S. (2016) Improving Mandarin Prosody Generation Using Alternative Smoothing Techniques. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **24**, 1897-1907. https://doi.org/10.1109/taslp.2016.2588727

[33] Huang, Y., Wu, C., Chen, Y., Shie, M. and Wang, J. (2017) Personalized Spontaneous Speech Synthesis Using a Small-Sized Unsegmented Semispontaneous Speech. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **25**, 1048-1060. https://doi.org/10.1109/taslp.2017.2679603

[34] Takatsu, H., Fukuoka, I., Fujie, S., Iwata, K. and Kobayashi, T. (2019) Speech Synthesis for Conversational News Contents Delivery. *Transactions of the Japanese Society for Artificial Intelligence*, **34**, 1-15. https://doi.org/10.1527/tjsai.b-i65

[35] Kato, S., Yasuda, Y., Wang, X., Cooper, E., Takaki, S. and Yamagishi, J. (2020) Modeling of Rakugo Speech and Its Limitations: Toward Speech Synthesis That Entertains Audiences. *IEEE Access*, **8**, 138149-138161. https://doi.org/10.1109/access.2020.3011975

[36] Costa, S., Brunete, A., Bae, B. and Mavridis, N. (2018) Emotional Storytelling Using Virtual and Robotic Agents. *International Journal of Humanoid Robotics*, **15**, Article 1850006. https://doi.org/10.1142/s0219843618500068

[37] Vuppala, A.K. and Kadiri, S.R. (2014) Neutral to Anger Speech Conversion Using Non-Uniform Duration Modification. 2014 9*th International Conference on Industrial and Information Systems* (*ICIIS*), Gwalior, 15-17 December 2014, 1-4. https://doi.org/10.1109/iciinfs.2014.7036614

[38] Vydana, H.K., Kadiri, S.R. and Vuppala, A.K. (2015) Vowel-Based Non-Uniform Prosody Modification for Emotion Conversion. *Circuits, Systems, and Signal Processing*, **35**, 1643-1663. https://doi.org/10.1007/s00034-015-0134-1

[39] Chatziagapi, A., Paraskevopoulos, G., Sgouropoulos, D., Pantazopoulos, G., Nikandrou, M., Giannakopoulos, T., *et al.* (2019) Data Augmentation Using Gans for Speech Emotion Recognition. *Interspeech* 2019, Graz, 15-19 September 2019, 171-175. https://doi.org/10.21437/interspeech.2019-2561

Annex: Design of MOS Evaluation Questionnaire

Indicator	System sound A	System sound B	System sound C
1. Sound liking (Preference)	☐I really like this sound ☐I like this sound ☐I like the sound moderately ☐I don't like this sound very much ☐I don't like this sound	☐I really like this sound ☐I like this sound ☐I like the sound moderately ☐I don't like this sound very much ☐I don't like this sound	☐I really like this sound ☐I like this sound ☐I like the sound moderately ☐I don't like this sound very much ☐I don't like this sound
Order (represented by 1, 2 and 3)			
2. Sound clarity (Word error rate)	☐Excellent, can be completely relaxed, does not require concentration ☐Not bad, needs attention, does not require special concentration ☐Fair, moderate concentration ☐Not very good, need to concentrate ☐Poor, difficult to understand even if you try	☐Excellent, can be completely relaxed, does not require concentration ☐Not bad, needs attention, does not require special concentration ☐Fair, moderate concentration ☐Not very good, need to concentrate ☐Poor, difficult to understand even if you try	☐Excellent, can be completely relaxed, does not require concentration ☐Not bad, needs attention, does not require special concentration ☐Fair, moderate concentration ☐Not very good, need to concentrate ☐Poor, difficult to understand even if you try
Oder (represented by 1, 2 and 3)			
3. Degree of human-sounding (Naturalness)	☐Excellent, it's almost like listening to a real-person, sounds natural ☐Not bad, a bit like a real-person telling a story ☐It's okay, it's a little fake, it's moderately natural ☐Not very good, sounds a little weird ☐Poor, very unnatural, very robot-like	☐Excellent, it's almost like listening to a real-person, sounds natural ☐Not bad, a bit like a real-person telling a story ☐It's okay, it's a little fake, it's moderately natural ☐Not very good, sounds a little weird ☐Poor, very unnatural, very robot-like	☐Excellent, it's almost like listening to a real-person, sounds natural ☐Not bad, a bit like a real-person telling a story ☐It's okay, it's a little fake, it's moderately natural ☐Not very good, sounds a little weird ☐Poor, very unnatural, very robot-like
Oder (represented by 1, 2 and 3)			