S & M 4387

Design of a Mandarin Spoken Dialogue System Using Tacotron2-based Speech Synthesis with Dialogist-aware System-speaking-style Switching

Ing-Jr Ding,[1]* Po-Jung Chen,[2] Xin-Bau Li,[1] and Yih-Her Yan[2]

[1]Department of Electronic Engineering, National United University,
No. 2, Lienda, Miaoli 360302, Taiwan
[2]Department of Electrical Engineering, National Formosa University,
No. 64, Wunhua Rd., Huwei Township, Yunlin County 632, Taiwan

(Received September 24, 2025; accepted February 20, 2026)

Keywords: spoken dialogue system, Tacotron2 speech synthesis, model fine tuning, synthetic speech evaluation, YOLO dialogist identification

As the global aging trend intensifies, the demand for long-term care systems will continue to rise, necessitating solutions to the problems of a shortage of manpower and excessive burdens on traditional human care. Among all care systems using AI techniques, the chatting system that can create a tight interaction between the aged and the system has inevitably become a necessary AI tool. However, for the aged, including those in Taiwan society, text-typing-based AI chatting systems with the interaction model of text-in–text-out are highly complicated and difficult to use. To tackle this issue, we will develop a Mandarin spoken dialogue system where chatting interactions will be in a simple and straight speech-to-speech mode. In addition, to provide emotion-connected voice interactions with psychological comfort and social companionship, the designed dialogue system will specifically contain the functionality of dialogist-aware system-speaking-style switching; in accordance with the system dialogist identity, the responding synthetic speech of the system will be in the style of a target speaker that is matched to the dialogist. The developed Mandarin spoken dialogue system in this study typically includes three computing modules, automatic speech recognition (ASR), semantics understanding of a large language model (LLM), and text-to-speech (TTS) speech synthesis. For the first two modules, the open source Google ASR and Google Gemma LLM are effectively employed and suitably integrated into the dialogue system. For TTS, to additionally perform system-speaking-style switching, the well-known Tacotron2 speech synthesis approach is adopted in this work. The Tacotron2 approach presented by Google is famous for its effectiveness in the deep learning of the speech database available. In this study, an initial Tacotron2 TTS model is first established using the Mandarin speech database 'Biaobei,' following which, a model fine-tuning procedure that uses small amounts of speech data from the specific target speaker to adjust the initial model parameters is designed. Aimed at the dialogist recognition of the dialogue system, You Only Look Once (YOLO)-based face detection is performed to classify the dialogist identity.

*Corresponding author: e-mail: eugen.ding@gmail.com

https://doi.org/10.18494/SAM5946

ISSN 0914-4935 © MYU K.K.
https://myukk.org/

With the recognized dialogist, the fine-tuned adaptation Tacotron2 model matched to this dialogist will then be used to perform speech synthesis. To evaluate the naturalness of the synthetic speech, various signal analysis evaluation metrics, including Mel-cepstral distortion (MCD), linear prediction code distortion (LPCD), and peak signal-to-noise ratio (PSNR), are also carried out in this work to investigate the effectiveness and compare the accuracy by the human-decision mean opinion score (MOS) approach.

1. Introduction

Intelligent accompany agents with the function of interactive chatting are an emerging technology in the current society of a large aging population. For the aged group, communicating with the system using the typical text-to-text mode will not be suitable because of the complexity and difficulty of utilizing computing devices. Chatting AI agents for care purposes are usually designed for more natural spoken interactions. Through speech-to-speech interactions with system inputs of natural speech from the dialogist and system outputs of synthetic speech responses created by the system, the spoken dialogue system can provide psychological comfort and social companionship to perform care of the aged. Early developed intelligent speech assistant systems, such as Apple Siri and Amazon Alexa, are mainly used for voice-command services and are based on limited-domain natural language processing and regular dialogue management.(1) However, such assistant systems have only limited interactive flexibility and have difficulty in understanding complex sentences. With the current rapid development of large language model (LLM) deep learning technology, dialogue systems will be able to understand more complex sentence structures, infer contextual relationships, and generate natural and semantically coherent responses, which will significantly enhance the practical help of spoken dialogue in elderly care.(2–4)

Currently, common LLMs include the Microsoft GPT series (such as GPT-3 and GPT-4) launched by OpenAI,(5) the LLaMA series published by Meta,(6) and the Gemini model of Google DeepMind.(7) Generally, in order to meet the needs of specific tasks, the fine-tuning technology of LLM has also been developed rapidly. The Gemma LLM, which belongs to the lightweight type compared with Gemini, is employed in developments of the spoken dialogue system in this work since such LLM may be suitable for smaller tasks such as chatting agent applications in this study. Furthermore, an additional consideration in integrating Gemma LLM into the spoken dialogue system is that improvements of the interpretability and performance of the model by the well-known low-rank adaptation (LoRA) fine-tuning approach have been successfully achieved in related applications.(8)

In this work, a Mandarin spoken dialogue system is developed for care of the elderly in Taiwan society. As mentioned, the open source Gemma LLM is adopted for semantic understanding, and for speech recognition for translating the speech of the dialogist into text, another open source Google automatic speech recognition (ASR) is utilized. For speech synthesis to generate semantic responses to the dialogist, the constructed spoken dialogue system in this work employs the Tacotron2-based approach. Although the early developed parametric synthesis approach (based on the characteristics of acoustic models, such as the statistical parametric synthesis) is a simple way without the need for exhausting model training,[9] such a method will inevitably encounter a serious problem of poor naturalness of the synthesized speech. With the rise of deep neural networks, end-to-end architectures have gradually become the mainstream technique in the development of speech synthesis owing to significant improvements of the naturalness and fluency of the synthesized speech. The core concept of such end-to-end speech synthesis is to convert text directly into acoustic features through the sequence-to-sequence model. The Tacotron2-based approach incorporated in the developed dialogue system is also categorized into such a type of end-to-end speech synthesis.[10,11] Following the fundamental end-to-end structure of Tacotron2, some Tacotron2-variant speech synthesis approaches are presented, such as the Transformer text-to-speech (Transformer–TTS) method in Ref. 12 that uses a self-attention mechanism to replace the traditional long short-term memoery (LSTM) architecture in Tacotron2 to improve the efficiency and stability of modeling long sequences and the FastSpeech-series method in Refs. 13 and 14 that further enhances Transformer–TTS by promoting typical autoregressive to non-autoregressive spectrogram generation using a duration predictor and position alignments for achieving fast speech synthesis. Although those Tacotron2-variant speech synthesis approaches are helpful in improving the data quality or inferring the speed of synthesis speech, difficulty will be encountered when further system development is required to adapt the TTS model to match speaking styles of the specific target speaker or to transfer emotion to the synthetic speech to form emotional synthesized speech.[15–17] To provide an emotion-connected voice interaction with psychological comfort and social companionship, speech synthesis with dialogist-aware system-speaking-style switching will be developed on the basis of the fundamental structure of Tacotron2 and properly incorporated into the Mandarin spoken dialogue system in this work. A You Only Look Once (YOLO)-based face detection scheme for classifying dialogist identity is constructed,[18] and then, in accordance with the recognized dialogist, the fine-tuned Tacotron2 model that has the speaking style of the target speaker and is matched to this dialogist will then be used to perform speech synthesis. Various signal analysis metrics will also be investigated to evaluate the naturalness of the synthetic speech. The fundamentals of Tacotron2 speech synthesis and related system development issues of the presented Mandarin spoken dialogue system with a target-speaker-adaptive Tacotron2 TTS for the identified system dialogist will be described in detail in the following sections.

2. Tacotron2 Speech Synthesis

As mentioned, the Tacotron2 framework is used for speech synthesis in the constructed Mandarin spoken dialogue system to generate voice responses to the utterances of the dialogist. Figure 1 depicts the overall structure of the Tacotron2 speech synthesis scheme, including mainly four computation parts: an encoder, a location-sensitive attention component, a decoder, and a vocoder integrated at the final output.[10,11] Tacotron2 is essentially a type of spectrogram deep learning model with the word token analysis of the corresponding text. In the training phase of Tacotron2 (i.e., model establishment), a database containing two separate sets, a set of a series of natural speech utterances and a set of a series of related text contents, will be necessary for model deep learning. Each of the text contents behaves as continuous word tokens and

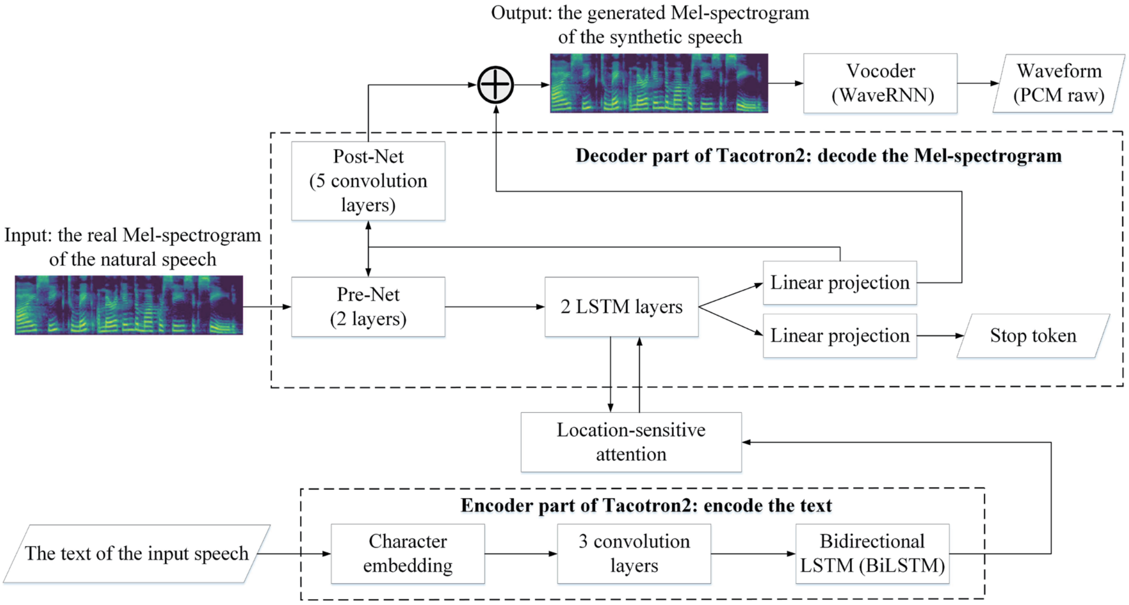

Output: the generated Mel-spectrogram of the synthetic speech
Vocoder (WaveRNN)
Waveform (PCM raw)
Post-Net (5 convolution layers)
Decoder part of Tacotron2: decode the Mel-spectrogram
Input: the real Mel-spectrogram of the natural speech
Pre-Net (2 layers)
2 LSTM layers
Linear projection
Linear projection
Stop token
Location-sensitive attention
Encoder part of Tacotron2: encode the text
The text of the input speech
Character embedding
3 convolution layers
Bidirectional LSTM (BiLSTM)

Fig. 1. (Color online) Framework of the Tacotron2 model for performing natural text-to-speech (the natural speech uttered in Mandarin in this work) translation.

corresponds to a certain speech utterance. When using Tacotron2 in the inference phase, only the text content desired to predict its spectrogram form is required.

As shown in Fig. 1, the encoder part of Tacotron2 mainly encodes the input text content and contains three operation modules: character embedding, three-layered convolution, and the bidirectional LSTM (BiLSTM) procedure. The main mission of the Tacotron2 decoder is to synthesize the spectrogram that corresponds to the input text content. In the training phase of Tacotron2 model establishments, both the input text content and its corresponding spectrogram are necessary; in the inference phase for online speech synthesis, only the input text content (obtained after speech recognition in this work) is required for decoding (inferring) its representative spectrogram. Because no input spectrogram is available in online inference [see Fig. (1)], zero vector is required as the initial input to guide the Pre-Net model of the decoder to start decoding. After feature extraction by the Pre-Net model, the extracted vector is concatenated with the context vector calculated by the location-sensitive attention mechanism. The concatenated vector is then passed to the LSTM to obtain the spectrogram hidden vector of the current frame. Such an LSTM-derived vector is sent to the linear projection layer to project the spectrum information of the current frame. In addition, the vector is simultaneously also sent to the Post-Net model with five convolution layers to compute the residual data, which will undergo an addition operation with the projected spectrum information of the current frame to further enhance the details of the current spectrum data. The current linearly projected spectrogram data will also be sent back to the Pre-Net model as input to derive the spectrogram information of the next frame. The spectrogram derivation will be repeated in a frame-by-frame manner until the appearance of the final "Stop token." This stop token indicates that the spectrogram decoding of the text content has been completed, that is, the final end of the text represented by a sequence of word tokens has been found. Note that the location-sensitive attention mechanism adopted by Tacotron2 can mitigate the undesired disadvantages of the traditional attention mechanism,(10,11) particularly the frequent appearances of errors such as repeated reading and word skipping owing to the lack of historical information memorized. To solve this problem, the location-sensitive attention mechanism additionally introduces the attention distribution of the previous time step (timestamp), which is shown as

$$f_t = ConV1D(\alpha_{t-1}), \tag{1}$$

where α_{t-1} denotes the attention distribution in the previous step, $ConV1D$ is a series of convolution filters with one dimension, and the final f_t derived from Eq. (1) represents convolutional location characteristics of the attention distribution in timestamp t. Equation (2) defines the attention score at each timestamp, $e_{t,i}$, which is also known as the alignment score:

$$e_{t,i} = V^T \tanh\left(W_s s_{t-1} + Vh_i + Uf_{t,i} + b\right), \tag{2}$$

where s_{t-1} denotes the hidden state of the decoder at timestamp $t-1$; h_i represents the i-th feature vector in the output sequence of the decoder; $f_{t,i}$ is the position feature obtained by the one-dimensional convolution of α_{t-1}; W_s, V, and U are the linear transformation matrices of s_{t-1}, h_i, and $f_{t,i}$, respectively; b indicates the bias value; and V^T denotes the attention score vector. On the basis of such a Tacotron2 deep learning framework, a Mandarin speech synthesis scheme with the ability to fine-tune the specific speaking style of the target speaker is constructed and properly incorporated into the Mandarin spoken dialogue system in this work. This will be described in detail in the following section.

3. Mandarin Spoken Dialogue System with Tacotron2 TTS of Tunable Target Speakers for Different System Dialogists

In this section, we will explain the designed Mandarin spoken dialogue system using Tacotron2-based speech synthesis with dialogist-aware system-speaking-style switching. Below we will introduce the Mandarin spoken dialogue system framework, the fine-tuning of the Tacotron2 TTS model for the speech synthesis of the target speaker, YOLO-based face recognition for dialogist awareness, and finally, the strategy adopted for naturalness evaluations of the synthesized speech in this work.

3.1 Designed framework of the Mandarin spoken dialogue system

As depicted in Fig. 2, the Mandarin spoken dialogue system designed in this work contains mainly three computation modules, ASR, LLM, and TTS. For the ASR module, the well-known cloud speech-to-text application program interface (API) released by Google is properly adopted and integrated in this study. The popular Google ASR is one of the web services of Google AI

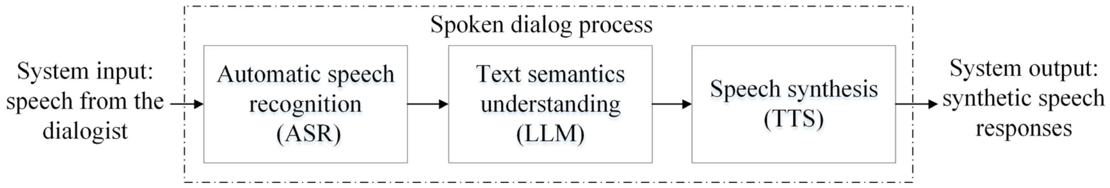

Spoken dialog process
System input: speech from the dialogist
Automatic speech recognition (ASR)
Text semantics understanding (LLM)
Speech synthesis (TTS)
System output: synthetic speech responses

Fig. 2. Classical spoken dialogue system with speech-to-speech interaction (generally only one speaking style in the synthetic speech output).

(i.e., a connection between the constructed dialogue system and the web of Google cloud over the IP networks required) and provides pretrained ASR models and related libraries for the spoken dialogue system (being one client of the connection-type web-client) to complete system integrations. As shown in Fig. 2, the system input of speech is first sent to ASR for translating the uttered speech from the dialogist to the correct text content, following which, the translated text content is then sent to the LLM module again to gain further text semantic understanding. Note that, besides using the open source of Google ASR, the LLM integrated in this dialogue system is also an open generation AI model, the Google Gemma LLM. Gemma was adopted for the reasoning of the speech-translated text content in this work because such a type of LLM is lightweight and capable of fine-tuning, which will therefore be appropriate in this work for rapidly inferring trustable semantics of the dialogist-spoken input.

For the TTS module interpolated at the end of the spoken dialogue system, as mentioned above, the Tacotron2 speech synthesis deep learning framework is used in this work (see Sect. 2). Note also that for the Tacotron2 system synthesis of Mandarin speech, the pre-processing of the Mandarin text will be required before the character embedding operation of the Tacotron2 encoder. The pre-processing of input text contents includes the phrase (or word) segmentation of the input text and “Hanyu Pinyin” of each separated word. Table 1 shows the text pre-processing performed on an input Mandarin word sequence. After text pre-processing, the character embedding of the Tacotron2 encoder can be performed to establish the embedded vector for each character. As can be seen in Table 1, the embedding vector of the Mandarin character ‘寶’ is denoted as “bao2”. Note that the number in the embedded vector represents the tone (there is a total of four basic tones, 1, 2, 3, and 4, in Mandarin). Mandarin characters with four different tones will be viewed as four separate character embedding vectors in speech synthesis. For all these constructed character embedding vectors, a symbol index table (see Table 2) to assign an index number to each different vector will be established so that the embedding vector of each Mandarin character can then be easily found during the text encoding procedure. Note that in Table 2, the symbol ‘_’ denotes a ‘space’, which is used for the purpose of ‘padding’ to align each input word sequence (i.e., padding token); the symbol ‘~’ is the “Stop token” that represents the end of the input text, notifying the encoder to stop text encoding in the training phase or reminding the decoder to stop spectrogram decoding in the inference phase, as previously described.

The Tacotron2-based TTS model for the Mandarin speech synthesis herein is first trained to be an initial TTS model using the database ‘Biaobei’, which contains 12 h of recordings of a total of 10000 Mandarin utterances. The initial Biaobei-derived TTS model can then be further fine-

Table 1
Example of pre-processing operations of the input text content by Mandarin Tacotron2 TTS.

<table>
<tr><td>Input word sequence (Mandarin)</td><td>寶馬配掛跛騾鞍，貂蟬怨枕董翁榻。</td></tr>
<tr><td>Word segmentations</td><td>寶馬 | 配掛 | 跛騾鞍 |，| 貂蟬 | 怨枕 | 董翁榻 |。|</td></tr>
<tr><td>Hanyu Pinyin</td><td>bao2 ma3 |pei4 gua4| bo3 luo2 an1|, |diao1 chan2| yuan4 zhen3 |dong3 weng1 ta4|。|</td></tr>
</table>

Table 2
Symbol index table established for making the dictionary of all character embedding vectors.

<table>
<tr><td>Hanyu Pinyin</td><td>bao2 ma3 |pei4 gua4| bo3 luo2 an1|, |diao1 chan2| yuan4 zhen3 |dong3 weng1 ta4|。|</td></tr>
<tr><td>Symbol index</td><td>‘_’:0,‘~’:1,“bao2”:2,“ma3”:3,“pei4”:4,“gua4”:5,“bo3”:6,“luo3”:7, “an1”:8, “diao1”:9, “chan2”:10, “yuan4”:11, “zhen3”:12, “dong3”:13, “weng1”:14, “ta4”:15, ‘，’:16, ‘。’:17</td></tr>
<tr><td>Index sequence</td><td>[2,3,4,5,6,7,8,16,9,10,11,12,13,14,15,17,0,0,...,1]</td></tr>
</table>

tuned as speaker-dependent TTS models. Each speaker-dependent TTS model corresponds to a specific target speaker with a distinctive speaking style, as will be explained in the following section.

3.2 Fast establishment of various target speakers with different speaking styles by tuning the initial Tacotron2 TTS model

The task of fine-tuning the initial Tacotron2 TTS model into a specific TTS model for the *i*-th speaker (i.e., Speaker-*i*) with an identical speaking style includes three main procedures: (1) the construction of the initial TTS model by Tacotron2, (2) the collection and pre-processing of small numbers of Mandarin utterances from a specific speaker, Speaker-*i* (i.e., datasets of Speaker-*i* for fine-tuning the initial TTS speech synthesis model), and (3) the establishment of the adaptive TTS model using speech datasets of a new target speaker (i.e., Speaker-*i*). Figure 3 shows the operation schematic of the Tacotron2 TTS model fine-tuning. As mentioned, in the first procedure, the large Biaobei database containing 10000 Mandarin utterances is employed for training the initial TTS model. Note that in this procedure, both text encoding (encoder) and spectrogram decoding (decoder) are performed. For the second procedure, various speakers were requested to be target speakers in this study, and only small numbers of speech data were recorded by each target speaker, 250 utterances from each target speaker. As shown in Fig. 3, three different adaptation datasets from three separate target speakers, Speaker-A, Speaker-B, and Speaker-C, will be used to tune the original Biaobei-derived TTS model to have identical model parameters to obtain three corresponding adaptive TTS models. Note also that for pre-processing the Mandarin text contents of the newly collected speech utterances of the target speaker, a similar task as the Biaobei database process mentioned in the previous section will also be carried out herein. As for model fine-tuning in the final procedure, different from the training of the initial TTS model, only model parameters of the decoder module are adjusted, and the entire encoder module is kept invariant (i.e., all the model parameters of the encoder are frozen) for the purpose of fast model adaptation. Freezing the encoder and tuning only the decoder parameters are essentially a type of transfer learning technique where the text encoder of target-speaker speech synthesis is trained on the relatively large database of Biaobei to provide the completeness of overall semantics.

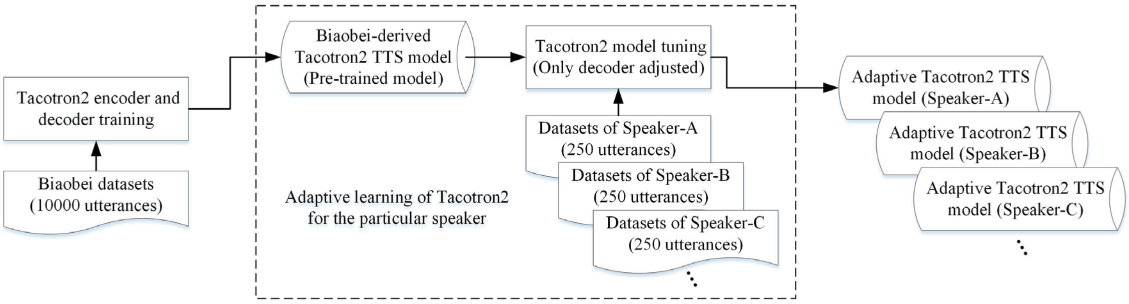

Tacotron2 encoder and decoder training
Biaobei datasets (10000 utterances)
Biaobei-derived Tacotron2 TTS model (Pre-trained model)
Tacotron2 model tuning (Only decoder adjusted)
Datasets of Speaker-A (250 utterances)
Datasets of Speaker-B (250 utterances)
Datasets of Speaker-C (250 utterances)
Adaptive learning of Tacotron2 for the particular speaker
Adaptive Tacotron2 TTS model (Speaker-A)
Adaptive Tacotron2 TTS model (Speaker-B)
Adaptive Tacotron2 TTS model (Speaker-C)

Fig. 3. Tacotron2 TTS system's fast establishment procedure of the target speaker with a particular speaking style using an adaptive learning scheme to tune the Biaobei-derived Tacotron2 TTS model using the collected speech utterances from the particular speaker.

3.3 Dialogist-aware spoken dialogue with speech synthesis of multiple target speakers by YOLO identity recognition

A series of adaptive Tactron2 TTS models constructed by model fine-tuning, as described above, can then be used in an online speech-to-speech dialogue system with the ability of dialogist awareness. As shown in Fig. 4, the developed spoken dialogue system in this work will perform the face recognition of the system dialogist before spoken interactions and choose one adaptive Tactron2 model that is completely matched to this dialogist to synthesize speech outputs in accordance with the recognized dialogist identity. Note that an identity mapping task between the current system dialogist and the corresponding target speaker with the particular speaking style will be properly set up in advance (for example, the relationships of "grandfather and grandson" and "mother and daughter"). The YOLO-based deep learning approach (mainly YOLOv4) will be employed in this study for the face recognition of system dialogists.(18) Users of our spoken dialogue system will be requested to scan RGB face images for the model training and inferrence by the YOLO face recognition deep learning scheme. The YOLOv4 structure used in this work for the face recognition of dialogists comprises mainly three computation components, 'Backbone', 'Neck', and 'Head'. The fine integration of these components will enable the high performance of feature extraction, fusion, and stacking of the extracted local and global features of the captured facial images of the system dialogist. Note that the bounded box for object detection (i.e., region of interest, ROI) in this work is the detected region of the dialogist's face in a captured speaking interaction image.

3.4 Evaluation metrics of the naturalness of speech data synthesized by Tacotron2

The naturalness of speech synthesized by our Tacotron2-based approach is also investigated and compared with that of real (i.e., ground truth data) speech. Four speech naturalness evaluation metrics are employed: mean opinion score (MOS), Mel-cepstral distortion (MCD), linear prediction code distortion (LPCD), and peak signal-to-noise ratio (PSNR). Note that MOS is a measure of the average decision opinions of all collected multimedia data qualities (mainly the synthesis speech in this work) of test persons. The evaluated result of MOS on synthetic

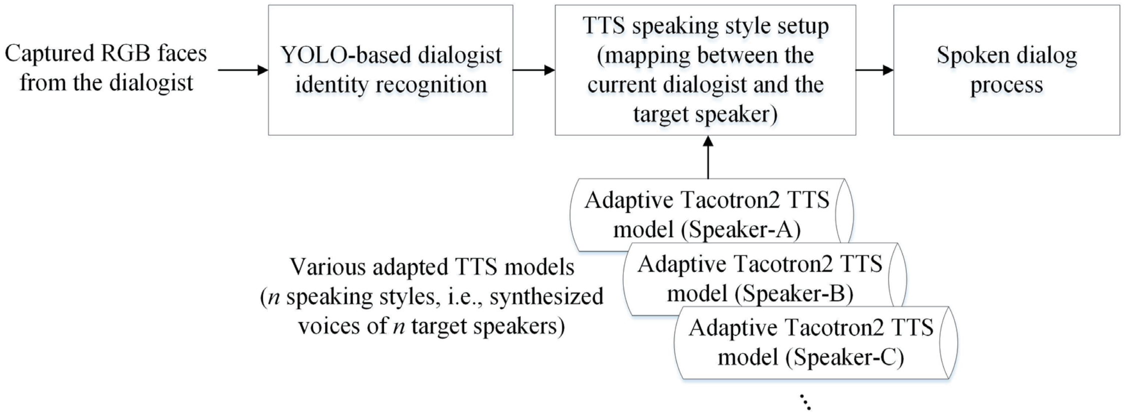

Captured RGB faces from the dialogist
YOLO-based dialogist identity recognition
TTS speaking style setup (mapping between the current dialogist and the target speaker)
Spoken dialog process
Adaptive Tacotron2 TTS model (Speaker-A)
Adaptive Tacotron2 TTS model (Speaker-B)
Adaptive Tacotron2 TTS model (Speaker-C)
Various adapted TTS models (*n* speaking styles, i.e., synthesized voices of *n* target speakers)

Fig. 4. Flowchart of the presented dialogist-identity-aware spoken dialogue system with switching of TTS speech synthesis of system speaking style (speech-to-speech interactions with synthesized voice of the particular speaker).

speech will be highly trustworthy. MCD, LPCD, and PSNR are determined by signal process computations and then used to compare the distortion (or difference) comparisons between real and synthetic speech to determine which evaluated result is closest to that of MOS (MCD in speech synthesis experiments in this work). Dynamic time warping (DTW), which is a type of dynamic programming, is used in MCD and LPCD metrics to compare distortion between the synthetic and real speech features.

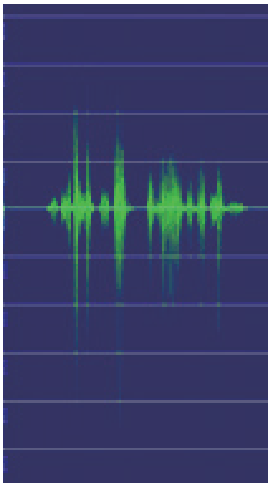
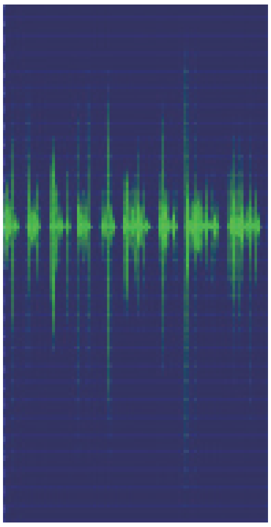
4. Experiments and Results

The system design and experimental construction of the developed Mandarin speech-to-speech dialogue system using Tacotron2-based speech synthesis with dialogist-aware speaking-style switching are carried out in a laboratory office environment. Training, inferring, and target speaker fine-tuning of the Tacotron2 speech synthesis model and YOLO-based face recognition deep learning for system-dialogist awareness are performed with a computer device comprising the hardware components of Intel i7 CPU, 128 GB RAM, and a GeForce RTX 4090 graphics card (containing mainly Nvidia GPU). The operating system on the computer is Ubuntu Linux (version 20.04). Both the speech data and the facial information of the dialogist are acquired using the properly deployed Kinect sensor device (version 1).(19) The Kinect sensor device developed by Microsoft is well known for its advantages of the simultaneous acquisition of video and voice data.

Table 3 shows an example of the speech-to-speech dialogue of our Mandarin spoken dialogue system developed in this work. As mentioned previously, the open source Goggle ASR and Gemma LLM were properly incorporated into this system to perform the speech recognition of input speech text and text generation after understanding of the semantics of the ASR-translated text. The synthetic speech RAW in Table 3 was generated by the well-trained Tacotron2 speech synthesis model.

In the Tarcotron2 speech synthesis experiments, the speech database BZNSYP (generally known as the Biaobei dataset, as mentioned previously) containing 10000 Mandarin utterances

Table 3
(Color online) Example of spoken dialogue. System inputs and outputs using the developed speech-to-speech Mandarin interaction dialogue (Google Gemma LLM integrated for understanding the semantics of the ASR-translated text) are listed.

System user (dialogist in speaking)		Spoken dialogue system (generates responses to utterances of the dialogist)	
Natural speech RAW (input)	Input speech text (by ASR)	Output speech text (from Gemma LLM)	Synthetic speech RAW (output)
	In Mandarin: 我覺得好孤單，都沒有任何人來看我。 (I feel so lonely, and nobody comes to see me.)	In Mandarin: 我理解您的感覺，在這個世界中，每個人都有自己的孤單，但您可以知道，您並不在單獨。很多人都經歷過類似的感覺。您可以嘗試與朋友、家人或其他信任的人分享您的感受，或者在網上上線與其他的人交流。 (I understand how you feel. In this world, everyone experiences loneliness at some point. Please know that you are not alone. Many people have faced similar feelings. You can try talking to family members or trusted individuals about your emotions, or connect with others online to share your feelings and experiences.)	

recorded by a woman speaking in standard Chinese is employed for training the initial Tacotron2 TTS model. Figure 5 depicts the convergence conditions in a series of iterative training. As depicted in Fig. 5, the total loss, which represents the differences between all synthetic and real spectrograms, can attain a satisfactory convergence value approaching 0 when the iterative training is sufficient (over 100000 steps). In the TTS inference phase for evaluating the performances of the constructed Tacotron2 models, text contents of the first 20 utterances in the BZNSYP database are used for text-to-speech translation. A total of 20 synthetic utterances, each of which corresponds to one of the text contents, are generated for test evaluation. As mentioned in Sect. 3.4, the four evaluation metrics of MOS, MCD, LPCD, and PSNR are utilized to evaluate the quality of these 20 synthetic utterances. Table 4 shows the averaged evaluation result of the 20 utterances for each metric. As observed from Table 4, the MOS score achieves an ideal value of 4.32 ± 0.085. Note that in the MOS evaluation task, 30 participating subjects aged 20 to 25 years were recruited and requested to rate each of the 20 synthetic utterances. In the MOS evaluation task, participants rated the naturalness and clarity of the synthetic speech (listen to the wav-file composed of PCM raw data) on a scale of 1 to 5, in which 5 indicates that the speech was closest to the human voice and 1 that the speech was completely unnatural or difficult to understand. In the two evaluations where distortion was compared, MCD shows an average distortion value of only 5.55, which indicates that the degree of similarity between the

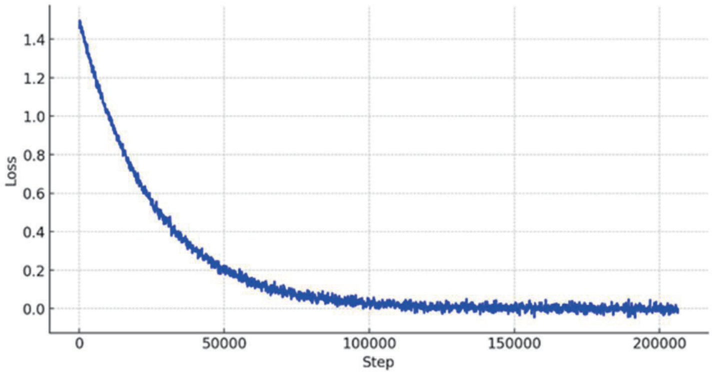


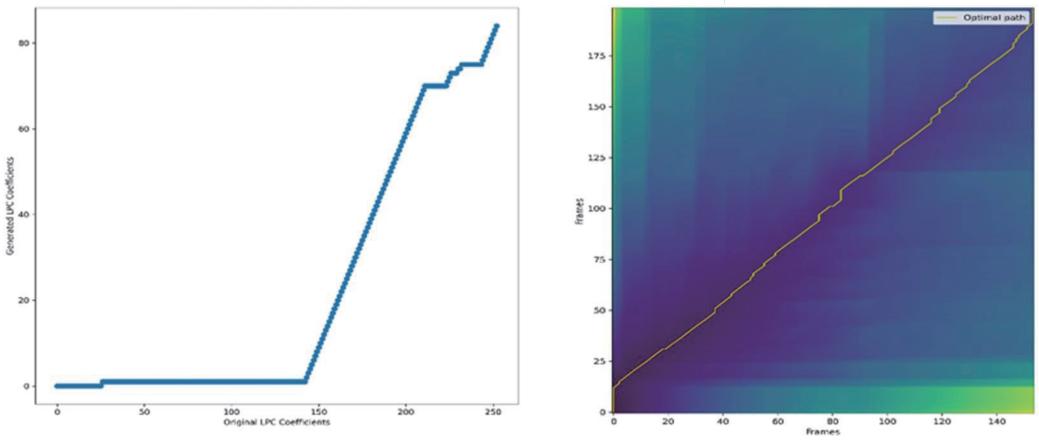
Fig. 5. (Color online) Loss curve of the initial Tacotron2 model in the training phase using the open Mandarin dataset ‘BZNSYP’ (10000 utterances) developed by the Data-Baker company.

Table 4
Averaged evaluation results of 20 Tacotron2 synthesized speech utterances for various evaluation metrics including human-decision MOS and three signal-analysis-based approaches.

Evaluation metrics	Average of 20 different utterances
MOS (human decision)	4.32 ± 0.085
MCD	5.55
LPCD	193.03
PSNR (dB)	16.97

synthetic speech and the real speech is very high. However, the evaluation result of LPCD approaches an obviously large value of 193.03, indicating a dissatisfactory speech outcome. Distortion curves between natural and synthetic speech data (the 20th utterance) derived by MCD and LPCD evaluations are illustrated in Fig. 6. The PSNR of the synthetic speech is also substandard, only 16.97 dB, which is far below the ideal 30 dB of the real-speech-like data. Therefore, experimental results demonstrate that the MCD comparison can reveal a more accurate evaluation result and express viewpoints similar to the human-decision.

During the fast establishment of TTS models adapted to the target speaker with a specific speaking style by fine-tuning the initially trained Tacotron2 TTS model, as described in Sect. 3.2, the system developer and non-system-developer test users are requested to record speech data. Each system test dialogist is requested to speak 250 different utterances to construct the speech database for TTS model adaptation. The text contents of these 250 recorded utterances completely correspond to those of speech data of the 750th to 999th utterances in the BZNSYP database. Figure 7 shows the curve of loss values in fine-tuning the decoder parameters of the initial Tacotron2 TTS model using 250 utterances of the target speaker (the system developer). An ideal convergence of loss values can be observed in Fig. 7, indicating fast model adaptability using only small amounts of adaptive speech data. Figure 8 shows the curves of attention alignment between the encoder and the decoder when fine-tuning the initial Tacotron2 model. As shown in Fig. 8, the attention alignment curve is rough up to the 100th iteration of training. By increasing the number of fine-tuning model training iterations to 1000, the attention alignment curve becomes significantly improved. At 10000 iterations, a standard attention



(a) (b)

Fig. 6. (Color online) Distortion curves between real and synthetic speech data (20th utterance in the dataset ‘BZNSYP’) obtained for the two evaluation metrics (a) LPCD and (b) MCD.

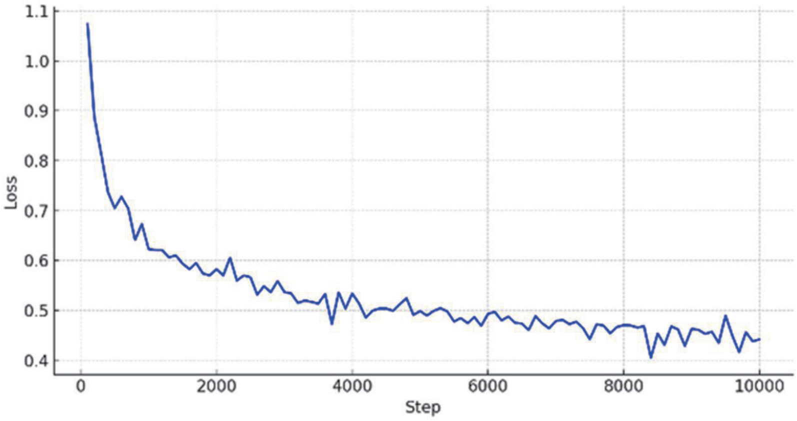


Fig. 7. (Color online) Loss curve of adaptive Tacotron2 model in the model fine-tuning phase using the database containing 250 utterances recorded by the target speaker (system developer).

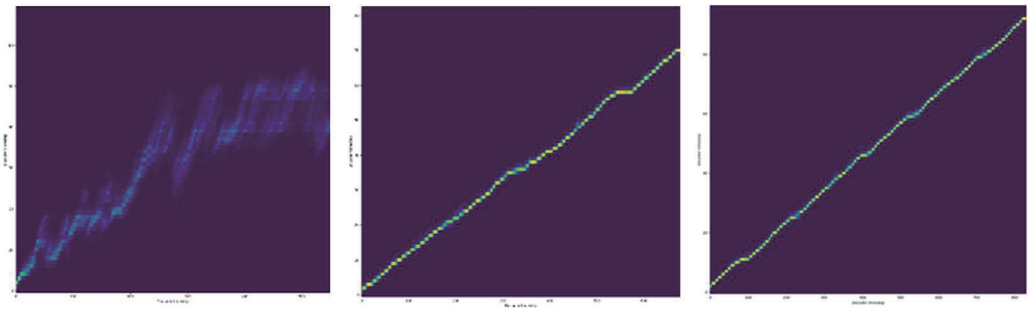


Fig. 8. (Color online) Attention alignment between encoding (*y*-axis) and decoding (*x*-axis) procedures of model fine-tuning of the initial Tacotron2 model (from left to right: 100, 1000, and 10000 training iterations).

alignment curve can be seen, indicating the attainment of excellent model fine-tuning. Figure 9 depicts the originally real (target) and generated spectrograms by the fine-tuned Tacotron2 model with 100, 1000, and 10000 iterations. To compare the naturalness of the initial and the adaptive Tacotron2 model synthesized speech, the MCDs of the different types of synthesized speech are evaluated, and the results are listed in Table 5. Fine-tuned Tacotron2 models show outstanding performances for both system-developer and non-system-developer target speakers. These close MCD distortion values in Table 5 indicate the effectiveness of the fast TTS model adaptation for the target speaker.

Finally, in the experiment using the YOLO-based deep learning for the face recognition of system users in order to generate dialogist-aware spoken dialogue, a typical model training procedure is performed to establish the YOLO-based face detection model. The dialogist identity recognition using the collected images of the faces of different users (the simulated system dialogists) achieved perfect accuracy. Figure 10(a) depicts the loss curve in the training phase of the face detection model (YOLOv4).[18] Figure 10(b) shows an example of dialogist identity

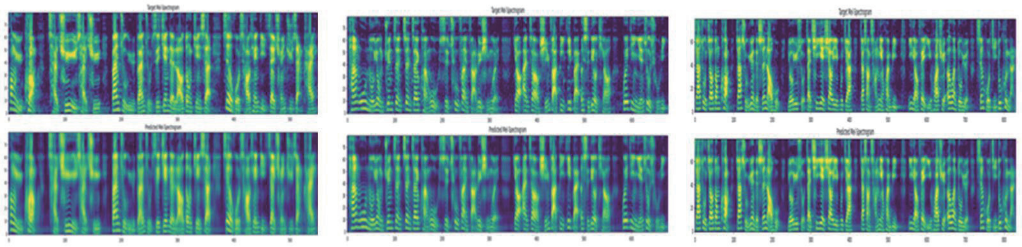
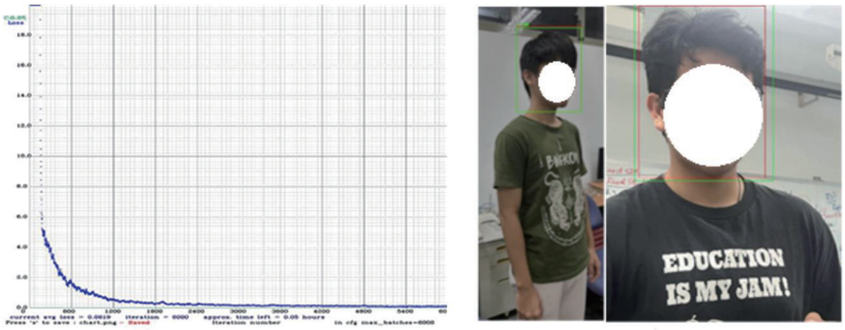


Fig. 9. (Color online) Speech spectrograms of real speech of the target speaker and synthesized speech generated by the fine-tuned Tacotron2 model (from left to right: 100, 1000, and 10000 training iterations; upper and lower: real and generated spectrograms).

Table 5
Average performances of three different types of synthesis speech (PCM raw) generated by the initial, system-developer-adaptive, and non-system-developer–adaptive Tacotron2 models.

Tacotron2 TTS model used for speech synthesis	MCD distortion
Initial TTS model using ‘BZNSYP’ database (WaveRNN vocoder)	5.55
Adaptive TTS model of system developer (Griffin–Lim vocoder)	10.02
Adaptive TTS model of non-system developer (Griffin–Lim vocoder)	10.68



(a) (b)

Fig. 10. (Color online) (a) Loss curve of YOLO-based face detection during training. (b) Dialogist awareness by identity recognition of system users (bounding box is perfectly positioned in the face region of the dialogist).

recognition using the trained YOLO model to detect and classify the identity of the dialogist, after which, the system target speaker can be matched with the specific adaptive Tacotron2 TTS model in advance before starting speech-to-speech dialogue, as mentioned in Sect. 3.3.

5. Conclusions

In this work, a Mandarin spoken dialogue system was developed in which a scheme of Tacotron2-based speech synthesis with dialogist-aware system-speaking-style switching is properly incorporated. By properly integrating open source ASR and LLM, the developed dialogue system can perform speech-to-speech interactions with dialogist speech inputs and system-synthesized speech outputs. To achieve awareness of the system dialogist and then generate responses of synthetic speech of the target speaker with a specific speaking style, the adaptation of the initial Tacotron2 model by fine-tuning and dialogist identity recognition by YOLO-based face recognition are also embedded in the system. Tacotron2 speech synthesis experiment results demonstrate that MCD can reveal accurate results and indicates judgements close to those of human-decision MOS in synthetic speech evaluations. The dialogue system designed in this study will greatly benefit the aged population that experiences difficulty in using text-typing methods to interact with AI chatting systems.

Acknowledgments

We acknowledge the support given by the National Science and Technology Council (NSTC), Taiwan, under Grant No. NSTC 114-2221-E-239-014.

References

1 M. B. Hoy: Med. Ref. Serv. Q. **37** (2018) 81. https://doi.org/10.1080/02763869.2018.1404391

2 H. Guo, S. Zhang, F. K. Soong, L. He, and L. Xie: Proc. 8th IEEE Spoken Language Technology Workshop (IEEE, 2021) 403–409. https://doi.org/10.1109/SLT48900.2021.9383460

3 X. Zhou, D. Chen, and Y. Chen: Proc. 6th Int. Conf. Nat. Lang. Speech Process. (ACL, 2023) 274–281. https://doi.org/10.48550/arXiv.2309.11000

4 Q. Fang, Y. Zhou, S. Guo, S. Zhang, and Y. Feng: Proc. 63rd Annual Meeting of the Association for Computational Linguistics (Association for Computational Linguistics, 2025) 18617–18629. https://doi.org/10.18653/v1/2025.acl-long.912

5 T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei: Proc. 34th Int. Conf. Neural Information Processing Systems (Curran Associates Inc., 2020) 1877–1901. https://doi.org/10.48550/arXiv.2005.14165

6 D. Huang, Z. Hu and Z. Wang: Proc. 2024 IEEE Int. Conf. Artificial Intelligence (IEEE, 2024) 1081–1085. https://doi.org/10.1109/CAI59869.2024.00108

7 Y. Kim and J. Heer: IEEE Trans. Vis. Comput. Graph. **27** (2021) 485. https://doi.org/10.1109/TVCG.2020.3030360

8 K. Mo, W. Liu, X. Xu, C. Yu, Y. Zou, and F. Xia: Proc. 2024 IEEE 4th Int. Conf. Electronic Technology, Communication and Information (2024) 130–135. https://doi.org/10.1109/ICETCI61221.2024.10594605

9 H. Zen, K. Tokuda, and A. W. Black: Speech Commun. **51** (2009) 1039. https://doi.org/10.1016/j.specom.2009.04.004

10 Y. Wang, R. J. Skerry-Ryan, D. Stanton, and Y. Wu: Proc. 18th Annual Conf. Interspeech 2017 (ISCA, 2017) 4006–4010. https://doi.org/10.21437/Interspeech.2017-1452

11 J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Chang, Y. Wang, R. Skerrv-Ryan, R. A. Saurous, Y. Agiomvrgiannakis, and Y. Wu: Proc. 2018 IEEE Int. Conf. Acoustics, Speech and Signal Processing (IEEE, 2018) 4779–4783. https://doi.org/10.1109/ICASSP.2018.8461368

12 N. Li, S. Liu, Y. Liu, S. Zhao, M. Liu and M. Zhou: Proc. 33th AAAI Conf. Artificial Intelligence (AAAI Press, 2019) 6706–6713. https://doi.org/10.1609/aaai.v33i01.33016706

13 Y. Ren, Y. Ruan, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T. Y. Liu: Proc. 33rd Int. Conf. Neural Information Processing Systems (Curran Associates Inc., 2019). https://doi.org/10.48550/arXiv.1905.09263

14 Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T. Y. Liu: Proc. 9th Int. Conf. Learning Representations (ICLR, 2021). https://doi.org/10.48550/arXiv.2006.04558

15 T. Li, S. Yang, L. Xue, and L. Xie: Proc. 12th IEEE Int. Symp. Chinese Spoken Language Processing (IEEE, 2021) 1–5. https://doi.org/10.1109/ISCSLP49672.2021.9362069

16 T. Li, X. Wang, Q. Xie, Z. Wang, and L. Xie: IEEE/ACM Trans. Audio Speech Lang. Process. **30** (2022) 1448–1460. https://doi.org/10.1109/TASLP.2022.3164181

17 O. Kwon, I. Jang, C. Ahn, and H. G. Kang: Proc. 34th IEEE Int. Tech. Conf. Circuits/Systems, Computers and Communications (IEEE, 2019) 1–4. https://doi.org/10.1109/ITC-CSCC.2019.8793393

18 M. L. Ali and Z. Zhang: Computers **13** (2024) 336. https://doi.org/10.3390/computers13120336

19 S. G. Heng, R. Samad, M. Mustafa, N. R. H. Abdullah, and D. Pebrianti: Proc. 2019 IEEE 9th Int. Conf. System Engineering and Technology (IEEE, 2019) 17–22. https://doi.org/10.1109/ICSEngT.2019.8906419