End-to-End Mandarin Speech Reconstruction Based on Ultrasound Tongue Images Using Deep Learning

Fengji Li, *Graduate Student Member, IEEE*, Fei Shen, *Graduate Student Member, IEEE*, Ding Ma, *Graduate Student Member, IEEE*, Jie Zhou, Shaochuan Zhang, Li Wang, *Member, IEEE*, Fan Fan, Tao Liu, Xiaohong Chen, Tomoki Toda, *Senior Member, IEEE*, and Haijun Niu, *Member, IEEE*

***Abstract*—The loss of speech function following a laryngectomy usually leads to severe physiological and psychological distress for laryngectomees. In clinical practice, most laryngectomees retain intact upper tract articulatory organs, emphasizing the significance of speech rehabilitation that utilizes articulatory motion information to effectively restore speech. This study proposed a deep learning-based end-to-end method for speech reconstruction using ultrasound tongue images. Initially, ultrasound tongue images and speech data were collected simultaneously with a designed Mandarin corpus. Subsequently, a speech reconstruction model was built based on adversarial neural networks. The model includes a pretrained feature extractor to process ultrasound images, an upsampling block to generate speech, and discriminators to ensure the similarity and fidelity of the reconstructed speech. Finally, both objective and subjective evaluations were conducted for the reconstructed speech. The reconstructed speech demonstrated high intelligibility in both Mandarin phonemes and tones. The character error rate of phonemes in automatic speech recognition was 0.2605, and tone error rate obtained from dictation tests was 0.1784, respectively. Objective results showed high similarity between the reconstructed and ground truth speech. Subjective perception results also indicated an acceptable level of naturalness. The proposed method demonstrates its capability to reconstruct tonal Mandarin speech from ultrasound tongue images. However, future research should concentrate on specific conditions of laryngectomees, aiming to enhance and optimize model performance. This will be achieved by enlarging training datasets, investigating the impact of ultrasound tongue imaging parameters, and further refining this method.**

***Index Terms*—Ultrasound tongue image, speech reconstruction, end-to-end, generative adversarial networks (GANs), Mandarin speech.**

I. Introduction

THE vocal cords are the most crucial organ in human speaking, serving as the source of speech production. However, most patients with laryngeal cancer have to undergo total laryngectomy, permanently losing their vocal cords and the ability to produce speech [1], resulting in severe physiological and psychological distress [2], [3]. In order to restore the ability to produce speech for such individuals with voice disabilities, many researchers have devoted significant efforts.

Laryngectomy primarily results in the loss of the voice source, prompting researchers to initially focus on methods to compensate for its loss and restore voice. Currently, three common methods of voice rehabilitation for laryngectomees are esophageal speech, tracheoesophageal speech, and electrolarynx (EL) speech [4]. In esophageal speech, phonation relies on an airflow from the stomach, often resulting in discontinuous voice due to the inadequate airflow. Tracheoesophageal speech addresses this issue by creating a fistula between the trachea and esophagus via tracheoesophageal puncture (TEP) and inserting a voice prosthesis. However, this method requires a two-step surgery and maintenance can be challenging. EL speech involves using an external vibrating device to produce speech. However, the speech generated by EL is often robotic, monotonic, and

Received 18 May 2024; revised 8 October 2024; accepted 17 December 2024. Date of publication 20 December 2024; date of current version 30 December 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 32071315 and Grant 12372299 and in part by Japan Science and Technology Agency (JST) Core Research for Evolutionary Science and Technology (CREST) Japan under Grant JPMJCR19A3. *(Corresponding author: Haijun Niu.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Beihang University Ethics Committee under Approval No. BM20230267.

Fengji Li, Fei Shen, Jie Zhou, Shaochuan Zhang, Li Wang, Fan Fan, Tao Liu, and Haijun Niu are with the School of Biological Science and Medical Engineering, Beihang University, Beijing 100191, China (e-mail: lifengji@buaa.edu.cn; wings@buaa.edu.cn; zhoujie18@buaa.edu.cn; zhangshaochuan@buaa.edu.cn; wangli4488@buaa.edu.cn; fanfan@buaa.edu.cn; tao.liu@buaa.edu.cn; hjniu@buaa.edu.cn).

Ding Ma is with the Graduate School of Informatics, Nagoya University, Nagoya 464-0823, Japan (e-mail: ding.ma@g.sp.m.is.nagoya-u.ac.jp).

Xiaohong Chen is with the Department of Otolaryngology, Head and Neck Surgery, Beijing Tongren Hospital, Capital Medical University, Beijing 100730, China (e-mail: trchxh@163.com).

Tomoki Toda is with the Information Technology Center, Nagoya University, Nagoya 464-0823, Japan (e-mail: tomoki@icts.nagoya-u.ac.jp).

Digital Object Identifier 10.1109/TNSRE.2024.3520498

lacks pitch modulation, which is particularly detrimental for tonal languages, leading to significantly poorer speech quality compared to natural speech [5].

In clinical practice, for most laryngectomees, despite the removal of the vocal cords, the articulatory organs responsible for phonation in the upper vocal tract remain intact. Exploring the possibility of utilizing neurophysiological signals and articulatory motion information to reconstruct high-quality speech has been concerned by scholars [6]. Some researchers have collected neurophysiological signals to conduct study related to speech, such as synthesizing speech from electromyography (EMG) [7] or electroencephalography (EEG) [8]. The research by Anumanchipalli et al. [9] on reconstructing high-quality speech using electrocorticography (ECoG) has gained considerable attention in recent years. They have explored the mechanism of neural activity associated with articulatory motion and synthesized speech based on it. Besides, some researchers have collected articulatory motion to reconstruct speech, with particular emphasis on capturing tongue motion. The tongue, as the most important articulatory organ, exhibits versatile and flexible movements that convey a wealth of articulatory information. The significant role of tongue in speech production has been extensively discussed by researchers such as Stone et al. [10], Hiiemae et al. [11], Badin et al. [12], and Chen et al. [13]. To date, researchers have explored various methods to record tongue motion: magnetic resonance imaging (MRI) [14] provides clear tongue image, but has long imaging time, making it difficult to capture rapid tongue movements in real-time; computed tomography (CT) [15] can record tongue movements clearly and rapidly, but it exposes the body to radiation, making it difficult to collect data for an extended period; electropalatography (EPG) [16] and electromagnetic articulography (EMA) [17] can accurately capture tongue movements, but sensors need to be placed inside the mouth during data collection, interfering with speech production movements.

Instead of the above-described approaches, ultrasound imaging [18], which boasts real-time, rapid, non-invasive and radiation-free attributes, is gradually becoming the preferred choice for many researchers due to its capability to accurately acquire tongue motion. Two decades ago, Denby and Stone were pioneers in utilizing a multilayer perceptron for mapping the features of ultrasound tongue images to vocal tract parameters [19]. Subsequently, researchers continuously explored to establish the relationship between dynamic ultrasound tongue images and speech by combining statistical methods or machine learning techniques. Hueber et al. proposed using a Hidden Markov Model (HMM) to establish the relationship between tongue image features and mel-frequency cepstral coefficients (MFCC) [20]. Csapó et al. used a ResNet model to map ultrasound images to mel-generalized cepstrum-based line spectral pair (MGC-LSP) [21]. Vocoders were further employed to synthesize speech after obtaining converted speech features. However, the main limitation of these works is solely using vocal tract parameters without exploring the fundamental frequency (F0) in tongue motion, resulting in a lack of naturalness in the reconstructed speech. To address this issue, Grósz et al. proposed a DNN model to predict F0 from ultrasound tongue images and utilized both the converted vocal tract parameters and F0 for speech reconstruction. Results of this study indicated that the predicted F0 contributes to enhancing the naturalness of the reconstructed speech [22].

Towards the improvement of the generated speech quality, researchers have investigated the methods for estimating mel-spectrogram, which contains sufficient vocal tract and F0 information, from ultrasound tongue images. For instance, Kimura et al. utilized convolutional neural networks (CNN) [23], while Tóth et al. utilized Spatial Transformer Networks [24], to establish the relationship between ultrasound tongue images and mel-spectrogram. One typical work proposed by Csapó et al. employed CNNs to convert ultrasound tongue images into mel-spectrograms and fed them into the WaveGlow vocoder for speech synthesis [25]. The results suggested that this approach obviates the need for separately reconstructing vocal tract parameters and F0, yielding reconstructed speech with naturalness. However, there still exists obvious disparity between reconstructed speech and real speech, particularly in terms of F0. For tonal languages characterized by intricate F0 variations, such as Mandarin Chinese, the quality and intelligibility of it are notably sensitive to F0 nuances [26]. Therefore, exploring how to reconstruct tonal Mandarin speech directly form ultrasound tongue images emerges as a highly worthy and thought-provoking research question.

It is obvious that previous works have avoided reconstructing audio waveforms, and have focused on generating intermediate representations which are used for reconstructing speech. To the best of our knowledge, there has been no research work on directly reconstructing Mandarin speech from tongue ultrasound images. End-to-End speech synthesis is an emerging technique in the field of speech synthesis that directly maps the input of the model to speech waveforms without intermediate steps [27]. This technology has been widely applied in the fields such as text-to-speech conversion, human-computer interface, and real-time translation in recent years. Unlike traditional methods based on phoneme concatenation or parameter-based synthesis, this approach simplifies the synthesis process and reduces the accumulation of errors [28], [29]. Numerous studies have demonstrated that employing this technique significantly enhances the quality and naturalness of synthesized speech, enabling better preservation of speech characteristics such as pronunciation style and prosodic rhythm [30].

Considering the aforementioned challenges, we propose an end-to-end method for reconstructing Mandarin speech based on ultrasound tongue images. This approach eliminates the need for converting speech features, thereby preserving the information directly from ultrasound tongue image. We employ image autoencoder and neural vocoder techniques and construct the model based on Generative Adversarial Networks (GANs). The model integrates pre-trained feature extractors and feature upsampling modules, and includes discriminators to ensure the similarity and fidelity of the reconstructed speech. The speech reconstruction model in this work builds upon our previous research [31]. However,

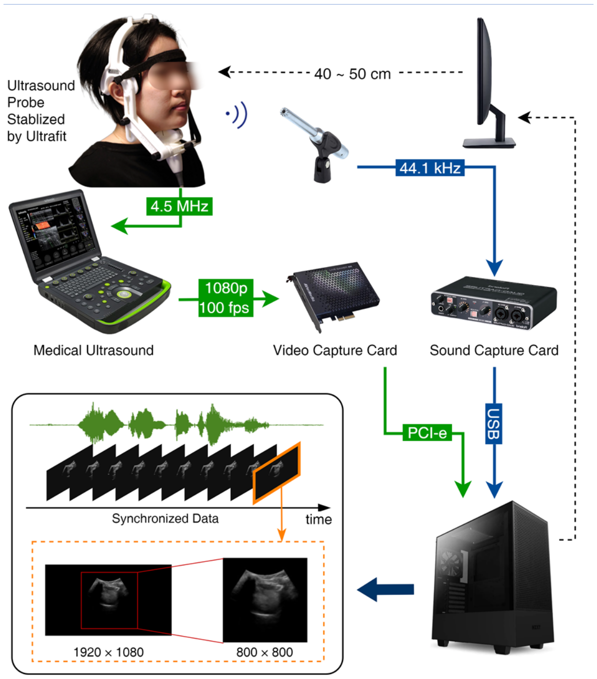

Fig. 1. Experimental system for audio and ultrasound tongue image acquisition and image preprocessing. The green line represents the acquisition process of ultrasound tongue images, while the blue line represents the audio acquisition process. Image preprocessing is depicted within the black frame.

the previous work did not use an end-to-end framework. In this paper, we reframe and deeply investigate the method for directly reconstructing Mandarin speech from ultrasound tongue images. To validate our proposed method, we collected synchronized data of ultrasound tongue images and speech, enabling the direct conversion of these images into fluent Mandarin speech. We then evaluated and analyzed the effectiveness of the reconstructed speech.

II. Materials and Methods

A. Experimental System

As shown in Fig. 1, the experimental system contains ultrasound tongue image and audio waveform acquisition. Ultrasound tongue images were obtained using Clover Medical B-mode ultrasound system with C5-1 curved transducer (Wisonic, China) and the GC573 video capture card (AverMedia, China). To stabilize the ultrasound transducer under the chin of the speaker to avoid deviations in measurement data, an Ultrafit headset (Articulate Instrument, UK) was used. Speech waveforms were captured using an ECM8000 microphone (Behringer, Germany) and a Quad-capture sound capture card (Roland, Japan). The synchronization of ultrasound video and audio signals was achieved through time-triggered coordination.

B. Data Acquisition

We built a corpus consisting of 1240 short sentences based on common Chinese daily expressions [32], with an average sentence length of 6 characters and a standard deviation of 3 characters. The entire corpus comprises 6858 Chinese characters. The data acquisition process involved a participation of a healthy native Mandarin speaker with no speech or hearing impairments. This participant signed written informed consent and the study was approved by Beihang University Ethics Committee (BM20230267). The entire experiment occurred in a soundproof recording studio. Throughout the process, the transducer was under the chin in a fixed position. The operating frequency of ultrasound transducer was 4.5 MHz. The ultrasound video stream was captured with 1080p (1920 × 1080 pixels) at a frame rate of 100 fps (frames per second), and the audio sampling frequency was set at 44.1 kHz. 1240 sentences were randomly presented on the screen, and speaker read the content with the display prompts.

C. Data Preparation

The collected data underwent preprocessing, which involved the following steps. After the separation of synchronized ultrasound video and audio, frames of the video were extracted to sequential images. Instead of using the full size of the image, we cropped the effective region with the central 800 × 800 pixels, excluding the surrounding black blank as depicted in Fig. 1. This process resulted in a dataset consisting of individual speech audio files matched with corresponding ultrasound tongue images. The dataset comprised a total of 1240 audio files and 422089 ultrasound tongue images.

D. Proposed Speech Reconstruction Architecture

The model for reconstructing speech based on ultrasound tongue images is illustrated in Fig. 2. Our model drew inspiration from advanced GAN-based vocoders, such as HiFi-GAN [33], MelGAN [34], and LSGAN [35], to construct a model for directly generating speech from features extracted from tongue motion ultrasound images. The ultrasound tongue images input into the model undergo feature extractor, encoding block, and upsampling modules to achieve end-to-end speech reconstruction. Specifically, the feature extractor is pretrained in a separately designed image autoencoder and then fine-tuned within the training process. Two independent discriminators are employed during the training of the generator to ensure the quality of the reconstructed speech.

1) Pretrained Feature Extraction: We designed the image autoencoder as shown in Fig. 2, comprising a feature extractor and a decoder. The autoencoder was used to extract efficient latent space representations of input in an unsupervised manner, and trained to restore the input images at the output layer. This forced the feature extractor to create compact representations. The feature extractor consists of four 2D convolution layers, each followed by rectified linear unit (ReLU) activation and 2D pooling. Meanwhile, the decoder consists of six 2D transposed convolution layers. Mean squared error (MSE) is used as a loss function. After training the autoencoder, the feature extractor acquired prior knowledge of extracting image features. By integrating the feature extractor into speech reconstruction model for fine-tuning, the prior knowledge from feature extractor was transferred into the speech reconstruction model training, resulting in enhanced performance.

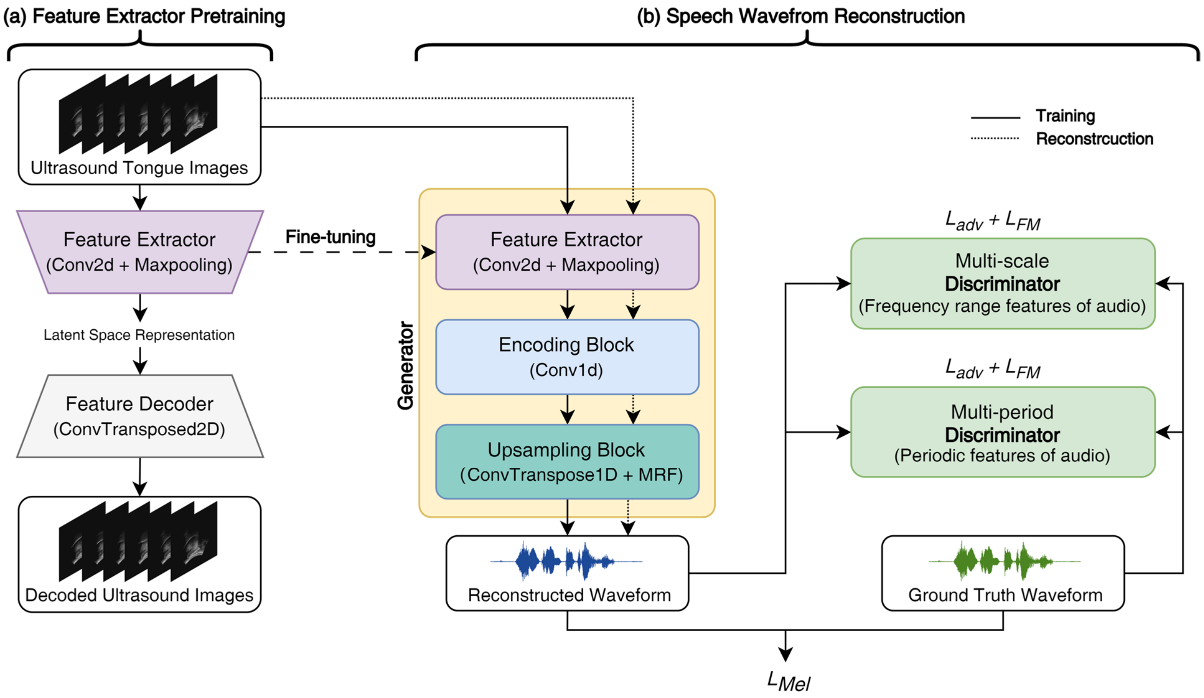

Fig. 2. The architecture of proposed method. (a) Feature extractor pretraining. (b) Speech waveform reconstruction.

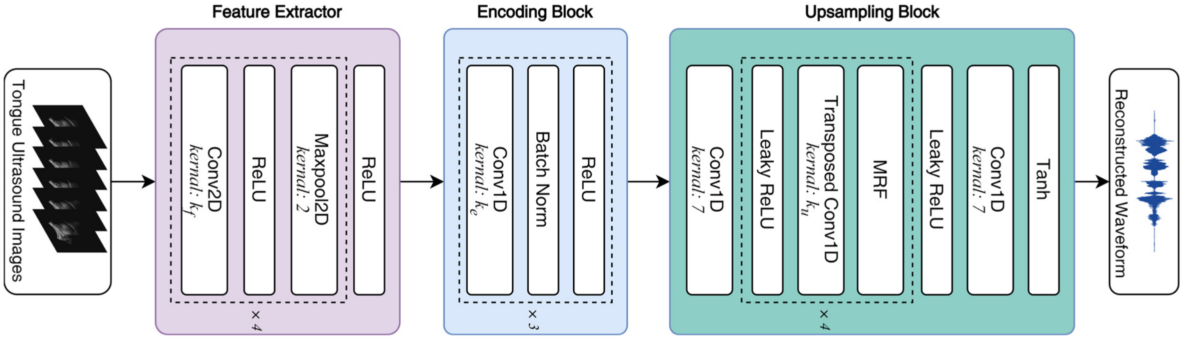

Fig. 3. The detailed architecture of generator consists of feature extractor, Encoding block, and Upsampling block.

2) Waveform Generator: Given that we aim to directly reconstruct speech from ultrasound tongue images, our generator accomplishes two sequential tasks: 1) decoding the temporal sequence of image features, and 2) upsampling the features into audio waveform. The specific structure of the generator is described in details in Fig. 3. First, the images are decoded into the latent space representations by feature extractor. Next, the representations undergo dimension adjustment before being fed into the encoding module, which consist of three 1D convolution layers, each followed with a batch normalization layer and ReLU activation. After this, the upsampling block performs upsampling until the length of the output sequence matches the temporal resolution of raw waveform. The upsampling block is primarily achieved by four layers of 1D transposed convolution layers, each followed by a multi-receptive field fusion (MRF) module [32]. The MRF utilizes a residual structure alternating between different convolutional kernels and dilation rates to enhance the upsampling performance. Some parameters in the generator are adjustable: kernel sizes *kf* of the feature extractor, kernel sizes *ke* of the encoding block, kernel sizes *ku* of the 1D transposed convolution in upsampling block.

3) Waveform Discriminators: We use two discriminators: 1) The multi-scale discriminator (MSD) [34] consisting of three scale discriminators, designed to operate on original, ×2 average-pooled, and ×4 average-pooled audio. Each of the sub-discriminators in MSD is a stack of 1D convolutional layers with leaky ReLU activation, focusing on the features across different frequency ranges in audios; 2) The multi-period discriminator (MPD) [33] consists of five period discriminators. Each sub-discriminator is a stack of 2D convolutional layers with leaky ReLU activation, processing reshaped and padded audio with period (2, 3, 5, 7, 11), focusing on periodicity features in audios.

4) Training Loss: The loss function comprises three components: GAN Loss, feature matching loss, and mel-spectrogram loss. The discriminator is trained to classify ground truth samples to 1, and the samples output from generator to 0.

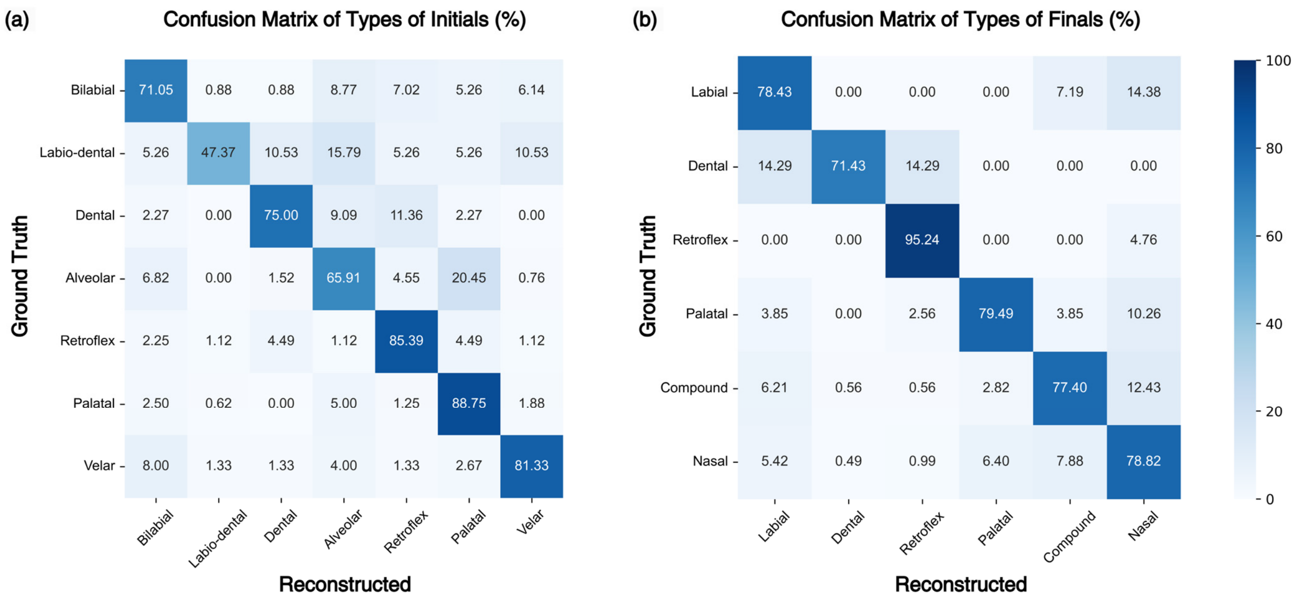

Fig. 4. Confusion matrix of types of mandarin pinyin results (ASR). (a) Confusion matrix of types of initials. (b) Confusion matrix of types of finals.

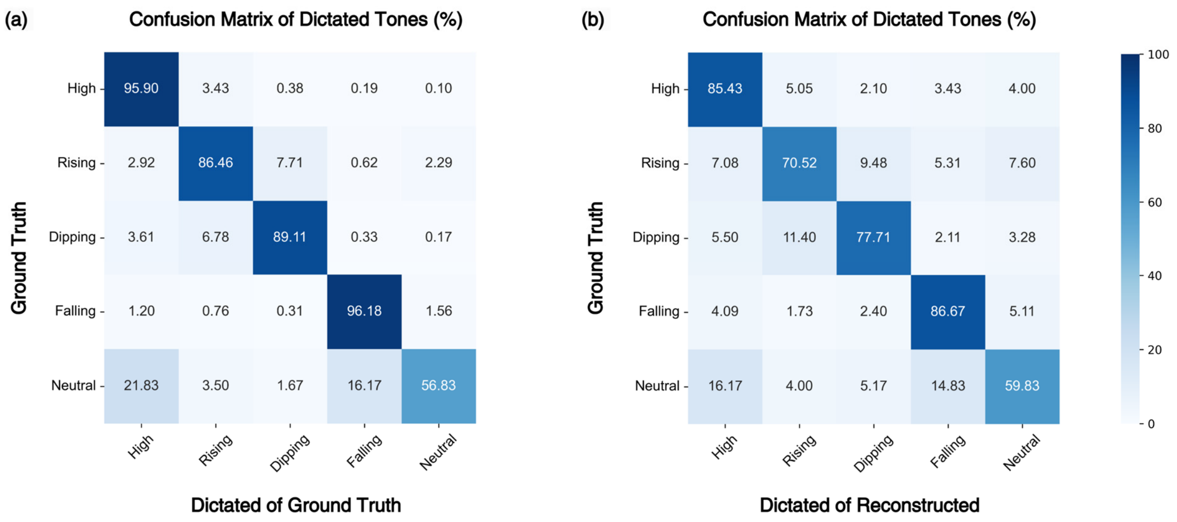

Fig. 5. Confusion matrix of mandarin tone results (Dictated). (a) Confusion matrix of ground truth. (b) Confusion matrix of reconstructed.

The generator is trained to fake the discriminator by updating the output quality to be classified to a value equal to 1. The GAN loss utilizes the least squares loss function, for generator and discriminator are defined as

$$\mathcal{L}_{Adv}(G; D) = \mathbb{E}_s\left[(D(G(u)) - 1)^2\right] \quad (1)$$

$$\mathcal{L}_{Adv}(D; G) = \mathbb{E}_{(x,s)}\left[(D(x) - 1)^2 + (D(G(u)))^2\right] \quad (2)$$

where Gdonates the generator, D donates the discriminator, x donates ground truth audio, and u donates the input of the ultrasound image. The feature matching loss quantifies the L1 distance between the output feature of each layer in the discriminator for ground truth audio and generated audio, assessing the similarity between them from a feature perspective. It is defined as

$$\mathcal{L}_{FM}(G; D) = \mathbb{E}_{(x,s)}\left[\sum_{i=1}^{T}\frac{1}{N_i}\left\| D^i(x) - D^i(G(u))\right\|_1\right] \quad (3)$$

where T donates the number of layers in the discriminator, D^i and N^i donate the features and the number of features in the i-th layers of the discriminators, respectively. The mel-spectrogram loss calculates the L1 distance between the mel-spectrograms of generated and original waveform, positively influencing the efficiency of the generator and the fidelity of the generated waveform.

$$\mathcal{L}_{Mel}(G) = \mathbb{E}_{(x,s)}\left[\| \phi(x) - \phi(G(u))\|_1\right] \quad (4)$$

where ϕ donates the function that extract mel-spectrogram from waveform. Finally, the loss of generator and discrimina-

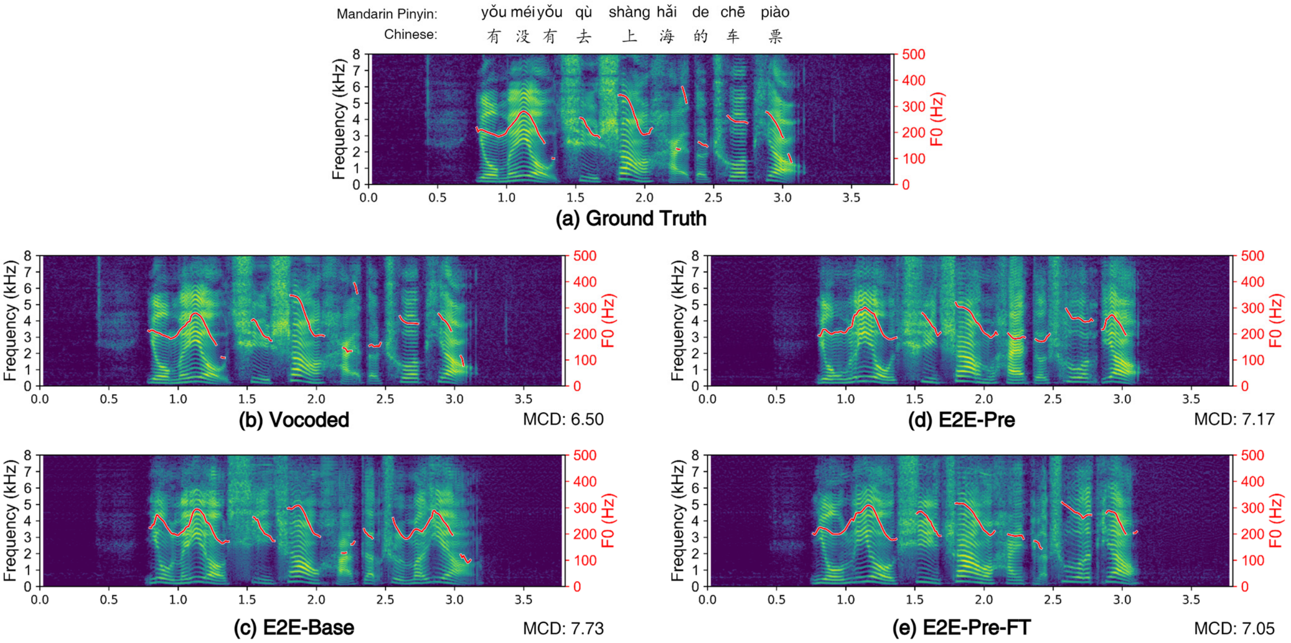

Fig. 6. Spectrograms and F0 contours for an example “有没有去上海的车票” under different evaluation conditions: (a) Ground truth, (b) vocoded, (c) E2E-Base, (d) E2E-Pre, and (e) E2E-Pre-FT.

tor are as

$$\mathcal{L}_G = \sum_{k=1}^{K} \left[\mathcal{L}_{Adv}(G;D) + \lambda_{FM}\mathcal{L}_{FM}(G;D)\right] + \lambda_{Mel}\mathcal{L}_{Mel}(G) \tag{5}$$

$$\mathcal{L}_D = \sum_{k=1}^{K} \mathcal{L}_{Adv}(D_m;G) \tag{6}$$

where K donates the number of discriminators, D_k donates the m-th sub-discriminator, and λ_{FM} and λ_{Mel} denote the weights assigned to feature matching loss and mel-spectrogram loss, set to 2 and 45, respectively.

5) Training Details: configured the model as follows: we set $k_f = [7, 5, 3, 3]$ in feature extractor, $k_e = [11, 9,7]$ in encoding block, and $k_u = [20, 12, 4, 4]$ in upsampling block. The model was trained using AdamW optimizer with optimizer parameters $\beta_1 = 0.8$, $\beta_2 = 0.99$, and weight decay of 0.01. The learning rate was starting from 0.0002 with decay by a factor of 0.999 in each epoch. The model training was conducted on a single GPU (NVIDIA RTX 3090) and CPU (AMD EPYC 7302) with a batch size of 16.

Additionally, to assess the significance and effectiveness of pretraining the feature extractor, we also trained models with feature extractors that were trained from scratch, as well as models where the feature extractor was pretrained but not fine-tuned. The structures and parameter settings were same with the speech reconstruction model containing a pretrained feature extractor.

E. Evaluation Metrics

Data were randomly divided into training set (1100), validation set (100), and test set (40). To evaluate the model's generalization performance, a modified three-fold cross-validation was employed. In each fold, the test set and validation set were distinct and non-overlapping, with each sample used once as part of the test set and once as part of the validation set. The training set remained fixed at 1100 samples across all folds. Objective and subjective evaluations were then performed, as described in the following sections.

1) Objective Measurements: We evaluated our speech reconstruction model using three kinds of objective metrics which capture different properties of the audio: 1) speech recognition accuracy, 2) spectrogram similarity parameter, and 3) F0 similarity parameter.

To objectively evaluate the intelligibility of the reconstructed speech, a pretrained automatic speech recognition (ASR) model was employed to transcribe both the reconstructed speech and the original speech samples [36]. Given that Mandarin syllables are primarily composed of initials (consonants) and finals (vowels), an in-depth analysis of the performance of the reconstructed speech was conducted by calculating the character error rate (CER) of mandarin syllables based on ASR results. Additionally, since this study focuses on speech reconstruction base on articulatory motions, different articulatory gestures and phonetic positions were analyzed, Mandarin syllables were categorized as shown in Table I. Mandarin initials were categorized into seven types: bilabial, labio-dental, dental, alveolar, retroflex, palatal, and velar. Mandarin finals were categorized into six types: labial, dental, retroflex, palatal, compound, and nasal. Analyzing such results can provide insights into the performance of reconstructed speech across various articulatory gestures and phonetic positions.

Speech quality was evaluated by spectrogram similarity parameter and F0 similarity parameters. Mel-cepstral distance

TABLE I
MANDARIN PINYIN INITIAL-FINAL MAP TABLE WITH INTERNATIONAL PHONETIC ALPHABET (IPA)

Types of Initials	Initials	Types of Finals	Finals
Bilabial	b ([p]), p ([p^h]), m ([m]), w ([w])	Labial	a ([a]), o ([o]), e ([ɤ]), u ([u])
Labio-dental	f ([f])	Dental	i ([ɹ])
Dental	z ([ts]), c([ts^h]), s([s])	Retroflex	er ([ɤ]), i ([ɻ̩])
Alveolar	d ([t]), t ([t^h]), n ([n]), l ([l])	Palatal	i ([i]), ü ([y])
Retroflex	zh ([tʂ]), ch ([ts^h]), sh ([ʂ]), r ([ʐ])	Compound	ai ([aɪ]), ao ([ɑʊ]), ei ([eɪ]), ou ([oʊ]), ia ([ja]), ie ([jɛ]), ua ([wa]), uo ([wo]), üe ([ɥɛ]), iao ([jɑʊ]), iu ([joʊ]), uai ([waɪ]), ui ([weɪ])
Palatal	j ([tɕ]), q ([tɕ^h]), x ([ɕ]), y ([j])	Nasal	an ([an]), en ([ən]), in ([in]), ün ([yn]), üan ([ɥæn]), un ([wən]), uan ([wan]), ian ([jɛn]), ang ([ɑŋ]), eng ([ɤŋ]), ing ([iŋ]), ong ([uŋ]), iang ([jɑŋ]), uang ([wɑŋ]), ueng ([wɤŋ]), iong ([iuŋ])
Velar	g ([k]), k ([k^h]), h ([h])		

(MCD) is designed to evaluate speech quality based on cepstrum distance on mel-scale [37]. F0 contours of the source speech signals were automatically extracted using a robust algorithm for pitch tracking [38]. The F0 logarithmic root mean square error (Log F0 RMSE), F0 correlation coefficient (F0 CORR), and F0 voiced/unvoiced accuracy (F0 V/U) are used to measure the consistency and accuracy of the F0. In practice, they work quite reliably in measuring quality in reconstructed speech compared to ground truth speech.

2) Subjective Listening Tests: We conducted listening tests to obtain evaluations of the speech, which comprised Mean Opinion Score (MOS) tests and Mandarin tone dictation tests.

In the MOS tests, listeners were instructed to rate the naturalness of both the ground truth and reconstructed utterances without prior knowledge of the sources. Ratings range from 1 to 5, following the criteria: 5 - Excellent, 4 - Good, 3 - Fair, 2 - Bad, 1 - Poor. Before the tests, listeners were provided with details written explanation of these ratings, covering aspects such as speech quality, clarity, the presence of mechanical noises or discontinuity, and whether the audio is perceived as speech.

The Mandarin tone dictation tests evaluate the performance of tone in the reconstructed speech. Mandarin tones can be categorized into five types: high (Tone 1), rising (Tone 2), dipping (Tone 3), falling (Tone 4), and neutral (Tone 0). During this test, participants listened to utterances and transcribed the tones they heard.

F. Evaluation Conditions

We prepared three training settings: 1) "E2E-Base" used a randomly initialized feature extractor in the speech reconstruction model, serving as baseline of our method. 2) "E2E-Pre" utilized a pretrained feature extractor in the model without fine-tuning, aimed at investigating the effect of fine-tuning. 3) "E2E-Pre-FT" represents the complete method proposed in this paper, where the feature extractor was pretrained and fine-tuned in the model. Additionally, we included the following methods for reference: "Ground Truth" represents real and natural speech, and "Vocoded" refers to the speech reconstructed using HiFi-GAN vocoder from mel-spectrogram of ground truth speech.

TABLE II
RESULTS OF CER BY ASR (MEAN ± STANDARD DEVIATION ACROSS CROSS-VALIDATION)

Method	Pretraining Feature Extractor	Fine-tuning Feature Extractor	CER of Mandarin Pinyin ↓
Ground Truth	-	-	0.0036
Vocoded	-	-	0.0036
E2E-Base	-	-	0.4563 ± 0.0197
E2E-Pre	✓	-	0.3012 ± 0.0153
E2E-Pre-FT	✓	✓	0.2605 ± 0.0165

TABLE III
RESULTS OF TER BY DICTATION TESTS

Method	TER of Mandarin Tones↓
Ground Truth	0.0959
E2E-Pre-FT	0.1784

III. RESULTS

A. Intelligibility Evaluation

The evaluation of Mandarin speech intelligibility contains both phoneme and tone recognition accuracy. Phoneme-level result was obtained by transcribed Mandarin pinyin through ASR. Tone result was obtained through subjective dictation test.

1) Results of Phoneme by ASR: Initially, the Baidu AI ASR system was employed to transcribe both the reconstructed and original speech into Mandarin Pinyin. Then we calculated the CER as shown in Table II. Among the results, E2E-Pre-FT method proposed in this paper achieved the best at 0.2544. The result of E2E-Pre method without fine-tuning closed to E2E-Pre-FT. Both of these significantly outperformed the E2E-Base method.

To assess the accuracy of reconstructed speech across various articulatory gestures and phonetic positions, results of E2E-Pre-FT method were analyzed. Based on the accuracy of Mandarin initials and finals and Table I, we designed confusion matrixes as shown in Fig. 4. Sub figures (a) and (b) illustrate the specific accuracy performance of initials and finals, respectively, where the color depth on the diagonal

TABLE IV
RESULTS OF MCD AND F0 PARAMETERS (MEAN ± STANDARD DEVIATION ACROSS CROSS-VALIDATION)

Method	Pretraining Feature Extractor	Fine-tuning Feature Extractor	MCD [dB] ↓	Log F0 RMSE ↓	F0 CORR ↑	F0 V/U ↑
Vocoded	-	-	6.42	0.28	0.87	0.83
E2E-Base	-	-	7.83 ± 0.02	0.36 ± 0.01	0.73 ± 0.02	0.79 ± 0.00
E2E-Pre	✓	-	7.45 ± 0.07	0.34 ± 0.00	0.75 ± 0.00	0.81 ± 0.01
E2E-Pre-FT	✓	✓	7.43 ± 0.10	0.35 ± 0.01	0.77 ± 0.00	0.82 ± 0.00

TABLE V
RESULTS OF MOS TEST WITH 95% CONFIDENCE INTERVAL

Method	MOS
Ground Truth	4.98 ± 0.01
E2E-Pre-FT	3.30 ± 0.13

indicates the recognition accuracy of each phoneme, and the remaining areas illustrate instances of phoneme confusion.

2) Results of Tones by Dictation: Thirty healthy Mandarin-speaking listeners participated in the dictation tests. Based on the dictation results, the tone error rate was calculated as shown in Table III. The proposed E2E-Pre-FT method achieved a tone error rate (TER) of 0.1784. The confusion matrix is designed and presented in Fig. 5.

B. Speech Quality Evaluation

1) Results of MCD and F0: After dynamic time warping (DTW), the similarity parameters of the spectrogram and fundamental frequency between the reconstructed and original speech were calculated separately, with the results presented in Table IV. The E2E-Pre-FT method achieved an MCD, Log F0 RMSE, F0 CORR, and F0 V/U of 7.56, 0.35, 0.77, and 0.82, respectively. These values are close to those obtained with E2E-Pre method and significantly better than those with the E2E-Base method.

Additionally, for visualizing performance evaluation of MCD and F0, spectrogram and F0 contour for speech were plotted. Examples of one speech sample under all conditions are as shown in Fig. 6. It can be observed that the speech reconstructed by our proposed methods closely resembles the original and vocoded speech in both spectral and F0. Particularly, clear harmonic structures are visible in the low-frequency region, and the positions of resonance peaks and segmentation of F0 are consistent. In contrast, the E2E-Base method exhibits the poorest performance in both spectral and F0 contours, showing more significant differences.

2) Results of MOS Scores: Thirty healthy Mandarin-speaking listeners participated in the subjective listening tests. They listened to totaling 80 of natural and reconstructed speech utterances and rated them accordingly. The MOS test results are presented in Table V. The average MOS result for the original speech was 4.98 ± 0.01, while that for the reconstructed speech was 3.30 ± 0.13.

IV. DISCUSSION

Since the last century, scholars have been exploring speech-related research based on ultrasound tongue images [39], [40], [41]. In recent years, with the advancement of new digital signal processing and learning techniques, more and more researchers have explored the feasibility and methods of reconstructing speech from ultrasound tongue images [42]. In these studies, speech reconstruction based on ultrasound tongue images involves two separate steps: 1) the conversion from image to speech features (such as MGC-LSP, Formant, MFCC, and Mel-spectrogram), and 2) the use of a vocoder to convert speech features to speech waveforms. These two independent steps lead to the loss of information from the image during the entire process of speech reconstruction, and the separate models cannot achieve unified optimization during training, thereby increasing the complexity of the system. To address this issue, this paper explores the feasibility of directly reconstructing speech from ultrasound tongue images. The end-to-end based model can directly convert images into speech, thereby enabling more comprehensive utilization of the information contained in ultrasound tongue images. Moreover, the end-to-end model can achieve more consistent and unified optimization of the entire reconstruction process during training to minimize adverse effects of error propagation on the final performance [43], and can also simplify the speech reconstruction process, reducing the complexity and computational overhead of the system [27].

The development of neural vocoders has facilitated the emergence of high-performance vocoders such as Mel-GAN, HiFi-GAN etc., which can deeply consider the complex structure of dynamic changes of speech during the training process, thereby improving the quality of reconstructed speech [44]. In this paper, we were inspired by advanced GAN vocoders, and established the speech reconstruction model based on GANs. In the generator, the convolutional layers in the encoding block can effectively capture the features of input, providing additional information for upsampling. Meanwhile, the MRF in the upsampling block contributed to observing patterns of various lengths in parallel. For discriminators, MSD and MPD focused on different frequency ranges and periodic features in the audio. All these aspects of model help in reconstructing Mandarin speech with diverse F0 variations.

We designed and trained an autoencoder for the pretrained feature extractor. Autoencoder method are commonly used for feature extraction in images [45], [46], with the aim of accurately restore the input images at the output layer to force the feature extractor to create latent features that present sufficient information in an unsupervised way. For our method, the feature extractor had already learned the prior knowledge of extracting ultrasound tongue image features during the pretraining. After integrating it into the generator for fine-tuning, the model improved its ability to understand and analyze ultrasound image data by leveraging the prior knowledge, thereby enhancing the overall performance of the model. This aspect also validated by the results: the model with pretrained feature extractor performed better in all results compared to using the model with unpretrained feature extractor. Additionally, the fine-tuning of the feature extractor slightly improved the results, although the enhancement was not significant.

The results from the cross-validation demonstrated the stability and reliability of the proposed model. The low standard deviations across key metrics indicated that the model consistently performs well across different data folds.

In terms of the intelligibility of the reconstructed speech, our method achieved a Pinyin CER of 0.2605 through ASR assessment, indicating that our approach can generate easily understandable speech. The results of phoneme revealed that the accuracy of phonemes was closely related to tongue motion and phonetic positions, such as “retroflex”, “palatal”, and “velar”, exceeded 80% for Mandarin Initials and Finals. However, the accuracy for phonemes related to the tongue tip, such as “labial”, “labio-dental” and “alveolar”, was less satisfactory, possibly due to potential information loss about the tongue tip in ultrasound tongue imaging [47]. This also indicates that the phonemes of the reconstructed speech are influenced by the limited view of ultrasound tongue image acquisition.

Regarding the performance of Mandarin tone in the reconstructed speech, the TER obtained from dictation tests was 0.1784, indicating that our method can achieve a varied and accurate reconstruction of tones, even better than phonemes with CER of 0.2605. Indeed, many researchers believe that the F0 of speech is mainly related to vocal fold vibration, but studies have also shown a connection between articulatory motion and F0. Chen et al. found a positive correlation between tongue activity and F0 [13], and recent studies such as Zhao et al. [48] and Grósz et al. [22] have also indicated that F0 can be obtained from ultrasound tongue images. Therefore, it is reasonable to achieve tone reconstruction through our method. Additionally, in the results of tone dictation, the TER in the original speech was 0.0959, relatively higher than that of phoneme. We believe this is mainly because Mandarin tones exhibit tone sandhi, such as, phonological rule that changes the first of two “dipping” tones to a “rising” tone [49], there is also considerable controversy over the recognition of “neutral” tone [50]. In Fig. 5, it can also be observed that there is a significant confusion between the “rising” tone and “dipping” tone, as well as between the “neutral” tone and other tones in the confusion matrix.

In terms of the speech quality, the results of MCD and F0 indicate a high degree of similarity between the reconstructed speech and the original speech. The MCD value of the reconstructed speech is less than 8, indicating that the quality of the reconstructed speech is acceptable [51]. Concerning the F0 parameters, the difference between the reconstructed speech and the ground truth speech is small, which also provides another perspective to verify the high-quality tone reconstruction. Additionally, the observation of spectrogram and F0 contours further validates the quality of the reconstructed speech. In Fig. 6, it can be observed that the spectrogram of the reconstructed speech closely resembles that of the ground truth speech, with no significant differences in harmonic shapes, resonance peak positions, and voiced/unvoiced segmentation. Moreover, according to the results of the listening test, the MOS of the naturalness of the reconstructed speech is above 3, indicating that it is subjectively acceptable. Based on informal feedback from listeners, most of the speech found to be very close to normal speech in terms of tone, rhythm, and timbre.

This study primarily developed an end-to-end method for reconstructing Mandarin speech from ultrasound tongue images based on deep learning techniques. The preliminary results demonstrate the feasibility of the proposed method. These results indicate that the quality of the reconstructed speech is deemed acceptable. However, there remains a gap between the reconstructed speech and natural speech. Further research should address the specific needs of laryngectomee by expanding the dataset, tailor the model architecture, and optimizing model hyperparameters. Currently, only data from a single participant has been used, and the generalizability of the method has not been extensively explored. Hence, future work should also consider improvements for personalized modeling. To this end, we plan to include multiple participants in future studies, including both healthy individuals and laryngectomee, to more comprehensively evaluate the model’s performance across diverse speaker populations. Moreover, factors such as tongue motion patterns and parameters of ultrasound imaging (e.g., resolution, frame rate, and imaging focal position) may also influence the quality of the reconstructed speech, which warrants further exploration. Additionally, we have updated the discussion section to highlight the potential clinical applications of this method.

References

[1] C. G. Tang and C. F. Sinclair, “Voice restoration after total laryngectomy,” *Otolaryngologic Clinics North Amer.*, vol. 48, no. 4, pp. 687–702, Aug. 2015.

[2] E. Babin, D. Beynier, D. Le Gall, and M. Hitier, “Psychosocial quality of life in patients after total laryngectomy,” *Rev. Laryngol. Otol. Rhinol.*, vol. 130, no. 1, pp. 29–34, Feb. 2009.

[3] B. Polat, K. S. Orhan, M. C. Kesimli, Y. Gorgulu, M. Ulusan, and K. Deger, “The effects of indwelling voice prosthesis on the quality of life, depressive symptoms, and self-esteem in patients with total laryngectomy,” *Eur. Arch. Oto-Rhino-Laryngol.*, vol. 272, no. 11, pp. 3431–3437, Nov. 2015.

[4] S. Bien et al., “History of voice rehabilitation following laryngectomy,” *Laryngoscope*, vol. 118, no. 3, pp. 453–458, Mar. 2008.

[5] R. Kaye, C. G. Tang, and C. F. Sinclair, “The electrolarynx: Voice restoration after total laryngectomy,” *Med. Devices, Evidence Res.*, vol. 10, pp. 133–140, Jun. 2017.

[6] T. Schultz, M. Wand, T. Hueber, D. J. Krusienski, C. Herff, and J. S. Brumberg, “Biosignal-based spoken communication: A survey,” *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 25, no. 12, pp. 2257–2271, Dec. 2017.

[7] K. Scheck, D. Ivucic, Z. Ren, and T. Schultz, “Stream-ETS: Low-latency end-to-end speech synthesis from electromyography signals,” in *Proc. 15th ITG Conf. Speech Commun.*, Sep. 2023, pp. 200–204.

[8] J. S. Brumberg, K. M. Pitt, and J. D. Burnison, "A noninvasive brain-computer interface for real-time speech synthesis: The importance of multimodal feedback," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 26, no. 4, pp. 874–881, Apr. 2018.

[9] G. K. Anumanchipalli, J. Chartier, and E. F. Chang, "Speech synthesis from neural decoding of spoken sentences," *Nature*, vol. 568, no. 7753, pp. 493–498, Apr. 2019.

[10] M. Stone, M. A. Epstein, and K. Iskarous, "Functional segments in tongue movement," *Clin. Linguistics Phonetics*, vol. 18, nos. 6–8, pp. 507–521, Sep. 2004.

[11] K. M. Hiiemae and J. B. Palmer, "Tongue movements in feeding and speech," *Crit. Rev. Oral Biol. Med.*, vol. 14, no. 6, pp. 413–429, Nov. 2003.

[12] P. Badin, Y. Tarabalka, F. Elisei, and G. Bailly, "Can you 'read' tongue movements? Evaluation of the contribution of tongue display to speech understanding," *Speech Commun.*, vol. 52, no. 6, pp. 493–503, Jun. 2010.

[13] W.-R. Chen, D. H. Whalen, and M. K. Tiede, "A dual mechanism for intrinsic f0," *J. Phonetics*, vol. 87, Jul. 2021, Art. no. 101063.

[14] A. D. Scott, M. Wylezinska, M. J. Birch, and M. E. Miquel, "Speech MRI: Morphology and function," *Phys. Medica*, vol. 30, no. 6, pp. 604–618, Sep. 2014.

[15] V. V. Wear, J. W. Allred, D. Mi, and M. K. Strother, "Evaluating 'Eee' phonation in multidetector CT of the neck," *Amer. J. Neuroradiology*, vol. 30, no. 6, pp. 1102–1106, Jun. 2009.

[16] B. Jiang, J. Kim, and H. Park, "Palatal electrotactile display outperforms visual display in tongue motor learning," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 30, pp. 529–539, 2022.

[17] T. Toda, A. W. Black, and K. Tokuda, "Statistical mapping between articulatory movements and acoustic spectrum using a Gaussian mixture model," *Speech Commun.*, vol. 50, no. 3, pp. 215–227, Mar. 2008.

[18] B. Denby, T. Schultz, K. Honda, T. Hueber, J. M. Gilbert, and J. S. Brumberg, "Silent speech interfaces," *Speech Commun.*, vol. 52, no. 4, pp. 270–287, Apr. 2010.

[19] B. Denby and M. Stone, "Speech synthesis from real time ultrasound images of the tongue," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 1, May 2004, p. 4.

[20] T. Hueber, E.-L. Benaroya, G. Chollet, B. Denby, G. Dreyfus, and M. Stone, "Development of a silent speech interface driven by ultrasound and optical images of the tongue and lips," *Speech Commun.*, vol. 52, no. 4, pp. 288–300, Apr. 2010.

[21] T. G. Csapó, G. Gosztolya, L. Tóth, A. H. Shandiz, and A. Markó, "Optimizing the ultrasound tongue image representation for residual network-based articulatory-to-acoustic mapping," *Sensors*, vol. 22, no. 22, p. 8601, Nov. 2022.

[22] T. Grósz, G. Gosztolya, L. Tóth, T. G. Csapó, and A. Markó, "F0 estimation for DNN-based ultrasound silent speech interfaces," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 291–295.

[23] N. Kimura, M. Kono, and J. Rekimoto, "SottoVoce: An ultrasound imaging-based silent speech interaction using deep neural networks," in *Proc. CHI Conf. Human Factors Comput. Syst.* Glasgow, U.K.: ACM, May 2019, pp. 1–11.

[24] L. Tóth, A. H. Shandiz, G. Gosztolya, and T. G. Csapó, "Adaptation of tongue ultrasound-based silent speech interfaces using spatial transformer networks," in *Proc. INTERSPEECH*, Aug. 2023, pp. 1169–1173.

[25] T. G. Csapó, C. Zainkó, L. Tóth, G. Gosztolya, and A. Markó, "Ultrasound-based articulatory-to-acoustic mapping with WaveGlow speech synthesis," in *Proc. Interspeech*, Oct. 2020, pp. 2727–2731.

[26] A. Jongman, Y. Wang, C. B. Moore, and J. A. Sereno, "Perception and production of Mandarin Chinese tones," in *The Handbook of East Asian Psycholinguistics*, P. Li, L. H. Tan, E. Bates, and O. J. L. Tzeng, Eds., Cambridge, U.K.: Cambridge Univ. Press, 2006, pp. 209–217.

[27] Y. Wang et al., "Tacotron: Towards end-to-end speech synthesis," in *Proc. Interspeech*, Stockholm, Sweden, Aug. 2017, pp. 4006–4010.

[28] Z. Mu, X. Yang, and Y. Dong, "Review of end-to-end speech synthesis technology based on deep learning," Apr. 2021, *arXiv:2104.09995*.

[29] O. Nazir and A. Malik, "Deep learning end to end speech synthesis: A review," in *Proc. 2nd Int. Conf. Secure Cyber Comput. Commun. (ICSCCC)*. Jalandhar, India: IEEE, May 2021, pp. 66–71.

[30] X. Tan et al., "NaturalSpeech: End-to-end text-to-speech synthesis with human-level quality," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 6, pp. 4234–4245, Jun. 2024.

[31] F. Li et al., "Mandarin speech reconstruction from tongue motion ultrasound images based on generative adversarial networks," in *Proc. 46th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*. Orlando, FL, USA: IEEE, Jul. 2024, pp. 1–4.

[32] P. Li, *Everyday Chinese: 900 Sentences Chinese*. Beijing, China: Foreign Language Teaching and Research Press, 2007.

[33] J. Kong, J. Kim, and J. Bae, "HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis," in *Proc. 34th Int. Conf. Neural Inf. Process. Syst. (NIPS)*. Red Hook, NY, USA: Curran Associates, Jan. 2020, pp. 17022–17033.

[34] K. Kumar et al., "MelGAN: Generative adversarial networks for conditional waveform synthesis," in *Proc. Adv. Neural Inf. Process. Syst.*, Jan. 2019, pp. 14910–14921.

[35] X. Mao, Q. Li, H. Xie, R. Y. K. Lau, Z. Wang, and S. P. Smolley, "Least squares generative adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2813–2821.

[36] C. Spille, S. D. Ewert, B. Kollmeier, and B. T. Meyer, "Predicting speech intelligibility with deep neural networks," *Comput. Speech Lang.*, vol. 48, pp. 51–66, Mar. 2018.

[37] R. Kubichek, "Mel-cepstral distance measure for objective speech quality assessment," in *Proc. IEEE Pacific Rim Conf. Commun. Comput. Signal Process.*, vol. 1. Victoria, BC, Canada: IEEE, May 1993, pp. 125–128.

[38] C. Gussenhoven, "The phonology of tone and intonation," in *Linguistics* (Research Surveys). Cambridge, U.K.: Cambridge Univ. Press, 2004.

[39] C. A. Kelsey, F. D. Minifie, and T. J. Hixon, "Applications of ultrasound in speech research," *J. Speech Hearing Res.*, vol. 12, no. 3, pp. 564–575, Sep. 1969.

[40] B. C. Sonies, T. H. Shawker, T. E. Hall, L. H. Gerber, and S. B. Leighton, "Ultrasonic visualization of tongue motion during speech," *J. Acoust. Soc. Amer.*, vol. 70, no. 3, pp. 683–686, Sep. 1981.

[41] M. Stone, "A guide to analysing tongue motion from ultrasound images," *Clin. Linguistics Phonetics*, vol. 19, nos. 6–7, pp. 455–501, Jan. 2005.

[42] J. Cleland, "Ultrasound tongue imaging," in *Manual of Clinical Phonetics*. Evanston, IL, USA: Routledge, 2021.

[43] Y. Ren et al., "FastSpeech 2: Fast and high-quality end-to-end text to speech," in *Proc. 9th Int. Conf. Learn. Represent. (ICLR)*, Europe, Austria, May 2020, pp. 1–15.

[44] X. Tan, T. Qin, F. Soong, and T.-Y. Liu, "A survey on neural speech synthesis," Jul. 2021, *arXiv:2106.15561*.

[45] G. Gosztolya, Á. Pintér, L. Tóth, T. Grósz, A. Markó, and T. G. Csapó, "Autoencoder-based articulatory-to-acoustic mapping for ultrasound silent speech interfaces," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*. Budapest, Hungary: IEEE, Jul. 2019, pp. 1–8.

[46] K. Xu, T. G. Csapo, D. Feng, and H. Mi, "Ultrasound tongue gestural sequence classification using convolutional auto-encoder and recurrent neural network," in *Proc. 12th Int. Seminar Speech Prod. (ISSP)*, 2020, pp. 1–2.

[47] K. Al-Hammuri, F. Gebali, I. T. Chelvan, and A. Kanan, "Tongue contour tracking and segmentation in lingual ultrasound for speech recognition: A review," *Diagnostics*, vol. 12, no. 11, p. 2811, Nov. 2022.

[48] C. Zhao, L. Wang, J. Dang, and R. Yu, "Prediction of F0 based on articulatory features using DNN," in *Studies on Speech Production* (Lecture Notes in Computer Science), Q. Fang, J. Dang, P. Perrier, J. Wei, L. Wang, and N. Yan, Eds. Cham, Switzerland: Springer, 2018, pp. 58–67.

[49] S. R. Speer, C.-L. Shih, and M. L. Slowiaczek, "Prosodic structure in language understanding: Evidence from tone sandhi in Mandarin," *Lang. Speech*, vol. 32, no. 4, pp. 337–354, Oct. 1989.

[50] L. Depuydt, "Neutral tone in Chinese: A comprehensive theory bridging East and West," *Open J. Modern Linguistics*, vol. 12, no. 6, pp. 768–789, 2022.

[51] CMU Speech Group. (2012). *Machine Learning in Speech Synthesis*. [Online]. Available: http://festvox.org/11752/slides/lecture11a.pdf