TaLNet: Voice Reconstruction from Tongue and Lip Articulation with Transfer Learning from Text-to-Speech Synthesis

Jing-Xuan Zhang[1,2], **Korin Richmond**[2], **Zhen-Hua Ling**[1], **Li-Rong Dai**[1]
[1]National Engineering Laboratory for Speech and Language Information Processing,
University of Science and Technology of China, P. R. China,
[2]The Center for Speech Technology Research, University of Edinburgh, UK
nosisi@mail.ustc.edu.cn, korin@cstr.ed.ac.uk, {zhling, lrdai}@ustc.edu.cn

Abstract

This paper presents TaLNet, a model for voice reconstruction with ultrasound tongue and optical lip videos as inputs. TaLNet is based on an encoder-decoder architecture. Separate encoders are dedicated to processing the tongue and lip data streams respectively. The decoder predicts acoustic features conditioned on encoder outputs and speaker codes. To mitigate for having only relatively small amounts of dual articulatory-acoustic data available for training, and since our task here shares with text-to-speech (TTS) the common goal of speech generation, we propose a novel transfer learning strategy to exploit the much larger amounts of acoustic-only data available to train TTS models. For this, a Tacotron 2 TTS model is first trained, and then the parameters of its decoder are transferred to the TaLNet decoder. We have evaluated our approach on an unconstrained multi-speaker voice recovery task. Our results show the effectiveness of both the proposed model and the transfer learning strategy. Speech reconstructed using our proposed method significantly outperformed all baselines (DNN, BLSTM and without transfer learning) in terms of both naturalness and intelligibility. When using an ASR model decoding the recovery speech, the WER of our proposed method shows a relative reduction of over 30% compared to baselines.

Introduction

Human speech production involves over 100 muscles (Anumanchipalli, Chartier, and Chang 2019). The tongue and lips play a crucial role to control the shape of vocal tract, so modifying its resonant properties in order to convey the different phonemes uttered. It is therefore natural to ask whether, and to what extent, speech can be reconstructed by computer model from observed tongue and lip articulations alone. In this paper, we study the issue of continuous and open-vocabulary speech generation using both ultrasound tongue and optical lip video. This task falls under the umbrella of articulatory-to-acoustic conversion, and may also be referred to as a silent speech interface (SSI) (Denby et al. 2010). Not only does work in this area improve our understanding of speech production mechanisms, but it also promises wide-ranging practical applications, such as restoring speech communication for patients who have undergone laryngectomy surgery, or as an aid for assisting communication in either high-noise or silent environments.

A reasonable amount of prior work has looked at voice reconstruction over the years, from either tongue or lip data, or both data streams combined. One early study adopted contour points extracted from ultrasound tongue imaging as inputs to a multi-layer perceptron (MLP) for acoustic feature prediction (Denby and Stone 2004). Eigentongues have also been proposed for feature extraction from tongue images (Hueber et al. 2007a). Naturally, deep neural networks (DNNs) have been used for tongue-to-speech conversion and F0 estimation (Csapó et al. 2017; Tóth et al. 2018; Grósz et al. 2018; Kimura, Kono, and Rekimoto 2019). Other work has focused on speech recovery from lip videos, for example: convolutional neural networks (CNNs) (Ephrat and Peleg 2017; Akbari et al. 2018; Kumar et al. 2019) and a dual CNN model (Ephrat, Halperin, and Peleg 2017) have been proposed for speech reconstruction. Prajwal et al. proposed a model for accurately learning a lip-to-speech mapping for an individual speaker using about 20 hours of that speaker's video (Prajwal et al. 2020). Some studies, meanwhile, have leveraged both tongue and lip data together. For example, (Hueber et al. 2011; Hueber and Bailly 2016) built and compared Gaussian mixture models (GMMs) and hidden Markov models (HMMs) for the articulatory-to-acoustic conversion task.

Previous studies like these have achieved impressive results, but have also had several limitations. First, the articulatory-to-acoustic conversion models have been trained only with data specifically designed and recorded with specialist articulography equipment for the task. This has strongly limited the amount of data available for training, certainly in comparison with standard audio-only TTS corpora which are much easier to collect. To address this, we argue that since both TTS and articulatory-to-acoustic conversion have in common the same objective of generating natural speech, we should seek to transfer learning from a TTS task to our model for articulatory-to-acoustic conversion. Second, the majority of work has focused exclusively either on tongue or lip inputs, rather than utilizing both features in tandem. Third, because articulatory-to-acoustic conversion remains challenging, many studies have focused only on narrow and constrained-vocabulary tasks, or indeed speaker-dependent models trained with data from a single speaker. Multi-speaker performance has not been well studied, and nor has performance on tasks with wide-ranging vocabularly.

To overcome these limitations, we propose *TaLNet* – a model for voice recovery from tongue and lip articulations which exploits transfer learning from TTS. TaLNet is based on an encoder-decoder architecture. Tongue and lip videos are processed by dedicated encoders based on three-dimensional (3D) CNNs. The hidden outputs from the tongue and lip encoders are then fused together along with a speaker code and fed into a decoder for predicting acoustic features. For transfer learning, a multi-speaker Tacotron 2 model is first trained on a large TTS corpus, and its decoder is then transferred to that of TaLNet. For fast, high-quality speech waveform recovery, we adopt a parallel-WaveGAN neural vocoder (Yamamoto, Song, and Kim 2020).

We have evaluated the TaLNet model on a multi-speaker, large-vocabulary speech task. The results of our experiments show the proposed method performs significantly better than all baselines in terms of both naturalness and intelligibility. We furthermore present the results of several ablation studies conducted to examine the characteristics and form of the TaLNet model. Finally, we also demostrate speech recovery from silent utterances.

Related Work

Other Articulography Features

In addition to the ultrasound tongue imaging and lip video used in this work, speech articulator movements can also be recorded by a range of other techniques, including: electromagnetic articulography (EMA) (Schönle et al. 1987); surface electromyography (sEMG) (Jorgensen and Dusan 2010); X-ray microbeam cinematography (Kiritani 1986); and magnetic resonance imaging (MRI) (Baer et al. 1987). Among these, EMA is perhaps the one that has been most frequently used in previous modelling work. (Kello and Plaut 2004; Ling et al. 2009; Toda, Black, and Tokuda 2008; Aryal and Gutierrez-Osuna 2016; Liu, Ling, and Dai 2016, 2018; Taguchi and Kaburagi 2018). EMA captures the location and orientation of sensor coils attached at selected fixed points on the articulators (typically tongue, lips, jaw, velum). Compared to EMA, tongue ultrasound and lip video have much higher dimensionality and greater noise to deal with. However, the advantage of tongue ultrasound and lip video is that they are non-invasive and do not require expensive bulky equipment, and so are far cheaper and more convenient to acquire.

Text Recognition from Tongue and Lips

Rather than reconstructing an audible speech signal, an alternative approach to SSIs that has attracted researchers' interest is recognition of text from tongue and lip data. Ultrasound tongue imaging (Xu et al. 2017; Ribeiro et al. 2019), lip video (Wand, Koutník, and Schmidhuber 2016; Assael et al. 2017; Afouras, Chung, and Zisserman 2018), or both of them together (Hueber et al. 2007b; Liu et al. 2016; Tatulli and Hueber 2017; Ji et al. 2018), have previously been adopted as model inputs. In particular, the conversion of lip video to text is often referred to *lipreading*. Compared to predicting text, speech reconstruction has two main attractive properties for SSI use. First, in addition to linguistic content, some prosodic characteristics of the speech can be preserved, which conveys supplementary information such as emotion and emphasis. Second, by regressing very short units of speech frame by frame, it can reconstruct words that are not present in the training set.

Text-to-speech Synthesis

Recently, sequence-to-sequence (seq2seq) models have been adopted for TTS, which predict acoustic features in an auto-regressive manner, for example Deep Voice 3 (Ping et al. 2018), Tacotron (Wang et al. 2017; Shen et al. 2018) or Tranformer-TTS (Li et al. 2019). High-fidelity neural vocoders, such as WaveNet (Den Oord et al. 2016), WaveRNN (Kalchbrenner et al. 2018) or parallel-WaveGAN (PWG) (Yamamoto, Song, and Kim 2020) are typically used for waveform reconstruction. Thanks to techniques such as these, synthesized speech is now reported to achieve performance that is very close to human quality (Li et al. 2019). Compared to TTS, our task in this paper shares the same goal of generating realistic speech signals, but differs in terms of input data streams (i.e. text for TTS versus articulation for articulatory-to-acoustic conversion). Our work here is inspired by the recent success of text-to-speech synthesis, and in particular because of their similarity. In our method, a seq2seq model and a high-fidelity neural vocoder are applied to the articulatory-to-acoustic conversion task. A strategy of transfer learning from a TTS task is also employed in order to increase performance.

Multi-speaker Tacotron

In this section, we briefly describe the Tacotron 2 model for multi-speaker text-to-speech synthesis. Tacotron 2 consists of a text encoder E and an acoustic decoder D_a. Let the text input sequence be $\mathbf{X} = [x_1, x_2, \ldots, x_N]$ and acoustic features be $\mathbf{A} = [\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_T]$, where N and T are lengths of text and acoustic sequence respectively. The text encoder first transforms the text inputs into linguistic representations $\mathbf{L} = [\mathbf{l}_1, \mathbf{l}_2, \ldots, \mathbf{l}_N]$ as

$$\mathbf{L} = E(\mathbf{X}). \tag{1}$$

The acoustic decoder accepts inputs of the linguistic representations $\mathbf{L}$ and speaker code $\mathbf{s}$ in the form of an x-vector (Snyder et al. 2018) in our experiments. The t-th acoustic frame is predicted by the decoder auto-regressively as

$$\mathbf{a}_t = D_a(\mathbf{L}, \mathbf{s}, \mathbf{a}_{<t}). \tag{2}$$

The auto-regressive inputs are first processed by a PreNet then sent to the decoder long short-term memory (LSTM). The decoder model is equipped with an attention block, which enables it to concentrate on a local context $\mathbf{c}_t$ of the linguistic sequence each step. This is achieved by a soft selection over the whole linguistic sequence as

$$\mathbf{c}_t = \sum_{n=1}^{N} \alpha_{n,t} * \mathbf{l}_n, \tag{3}$$

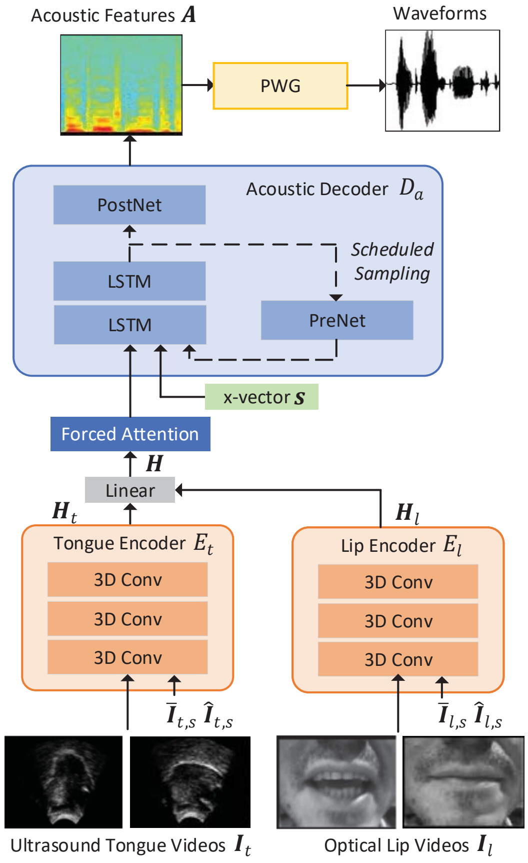

Acoustic Features $\boldsymbol{A}$
Waveforms
PWG
Acoustic Decoder D_a
PostNet
Scheduled Sampling
LSTM
LSTM
PreNet
x-vector $\boldsymbol{s}$
Forced Attention
$\boldsymbol{H}$
Linear
$\boldsymbol{H}_t$
$\boldsymbol{H}_l$
Tongue Encoder E_t
3D Conv
3D Conv
3D Conv
Lip Encoder E_l
3D Conv
3D Conv
3D Conv
$\bar{\boldsymbol{I}}_{t,s}$ $\hat{\boldsymbol{I}}_{t,s}$
$\bar{\boldsymbol{I}}_{l,s}$ $\hat{\boldsymbol{I}}_{l,s}$
Ultrasound Tongue Videos $\boldsymbol{I}_t$
Optical Lip Videos $\boldsymbol{I}_l$

Figure 1: An overview of the TaLNet model structure.

$$\alpha_{n,t} = SoftAttention(\mathbf{q}_t, \mathbf{l}_n), \tag{4}$$

where $\mathbf{q}_t$ is the query vector from the decoder LSTM state. In order to further boost the accuracy of acoustic feature prediction, a CNN-based PostNet is employed following the initial decoder outputs. We suggest the reader refers to the Tacotron 2 paper (Shen et al. 2018) for the full details.

Proposed Method

TaLNet has an encoder-decoder architecture, which includes a tongue encoder E_t for processing ultrasound images of the tongue, a lip encoder E_l for processing the images of the lips, and an acoustic decoder D_a for recovering acoustic features, as shown in Figure 1. We next describe each component in detail.

Encoder

Let $\mathbf{I}_t = [\mathbf{I}_{t,1}, \mathbf{I}_{t,2}, \ldots, \mathbf{I}_{t,T}]$ be ultrasound tongue frames and $\mathbf{I}_l = [\mathbf{I}_{l,1}, \mathbf{I}_{l,2}, \ldots, \mathbf{I}_{l,T}]$ be lip video images. For both tongue and lip images, pixel-wise mean and standard deviation are computed for each speaker. To obtain the inputs of the encoder, they are repeated and then appended as extra channels to the sequence of tongue and lip. The inputs are processed by the tongue encoder E_t and lip encoder E_l as

$$\mathbf{H}_t = E_t(\mathbf{I}_t, \bar{\mathbf{I}}_{t,s}, \hat{\mathbf{I}}_{t,s}), \tag{5}$$

$$\mathbf{H}_l = E_l(\mathbf{I}_l, \bar{\mathbf{I}}_{l,s}, \hat{\mathbf{I}}_{l,s}), \tag{6}$$

where $\mathbf{H}_t$ and $\mathbf{H}_l$ denote representations from tongue and lip respectively. $\bar{\mathbf{I}}_{t,s}$, $\hat{\mathbf{I}}_{t,s}$,$\bar{\mathbf{I}}_{l,s}$ and $\hat{\mathbf{I}}_{l,s}$ represent the mean tongue image, tongue standard deviation image, and lip mean and standard deviation images for speaker s respectively.

Ultrasound tongue and lip videos are challenging to deal with because of significant inter-speaker variation, arising from factors related to physiology, age or the recording process itself, including: tissue fat, shape of the oral cavity and face, facial hair, or placement of the ultrasound probe and camera. Therefore, by simply using the average and standard deviation images for the tongue and lips, we provide some speaker-specific information to better model inter-speaker variation. The tongue and lip encoders both adopt the same structure, which is based on stacks of 3D CNNs. The 3D CNN has been shown to be effective in multiple tasks involving spatio-temporal video data. The input to our network is a sequence of articulatory images with dimensions $T \times H \times W$, where T is the number of input frames, and H and W correspond to the spatial dimensions of the tongue and lip images.

Within the 3D CNN processing, the spatial dimensions H and W are reduced while the time dimension T is preserved. Maxpooling and dropout are also used to prevent overfitting. The final CNN layer outputs are flattened and passed through a fully connected layer to produce a single visual vector for each frame. Lastly, the vectors from tongue and lips are fused together to yield the final representations as

$$\mathbf{H} = \mathbf{W}\mathbf{H}_t + \mathbf{U}\mathbf{H}_l + \mathbf{b}, \tag{7}$$

where $\mathbf{W}$, $\mathbf{U}$ and $\mathbf{b}$ are model parameters.

Decoder

The decoder D_a is conditioned on the visual representations $\mathbf{H}$ and the speaker x-vector $\mathbf{s}$ to predict the acoustic features as

$$\mathbf{A} = D_a(\mathbf{H}, \mathbf{s}). \tag{8}$$

The acoustic decoder is transferred from a Tacotron 2 decoder pre-trained on a TTS task. The same structure is used, except we adopt a forced attention mechanism rather than the standard soft attention. Specifically, we force the decoder to attend to the corresponding visual vector along the time dimension by using an impulsive attention coefficient. Specifically, Equation 4 is replaced with

$$\alpha_{n,m} = \delta(m - n), \tag{9}$$

where $\delta(0) = 1$, otherwise $\delta(i) = 0$ when $i \neq 0$. Preliminary experiments showed it is necessary to use the forced attention. Otherwise, it proves difficult to ensure the the attention alignment converges after transferring from the TTS task. In contrast to the teacher forcing training typically used in Tacotron, we adopt a scheduled sampling strategy. During training, the model-predicted acoustic feature is selected with probability p as the auto-regressive input frame,

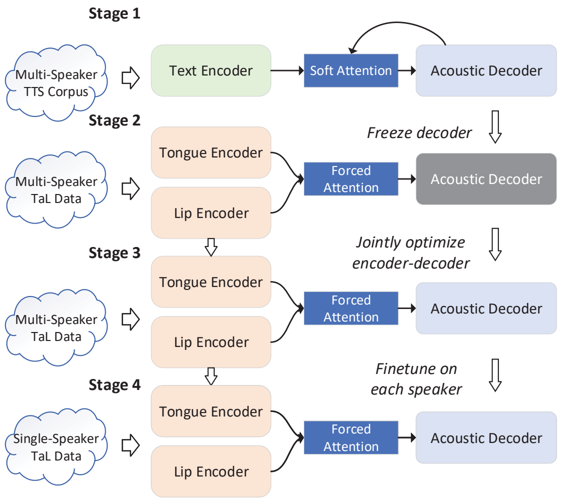


Figure 2: Training strategy of TaLNet. "TaL Data" represents the tongue and lip dataset used in the work here.

rather than always taking the natural one at each decoder step. Compared to always feeding back natural frames as in teacher forcing (i.e., p=0 in this case), scheduled sampling alleviates mismatch at training and testing times. The experiments later in the paper demonstrate the advantage of this.

Training Strategy

There are four stages involved in TaLNet training, as illustrated in Figure 2 and summarised as follows:

Stage 1 A multi-speaker Tacotron model, with text encoder and acoustic decoder, is trained on a multi-speaker TTS corpus. Mean absolute error (MAE) between predicted and target acoustic features is the optimization criterion.

Stage 2 The text encoder of the Tacotron model is discarded, while its decoder is transferred to be used as the TaLNet decoder. The soft attention module is substituted with forced attention. The decoder parameters are frozen and only the tongue and lip encoder parameters are updated at this stage. The encoders are therefore trained to fit the Tacotron decoder to minimize the MAE loss of acoustic feature prediction.

Stage 3 The encoders and the decoder of TaLNet are jointly optimized to fit each other with further training. Scheduled sampling is introduced with probability p linearly increased from 0 to 1.

Stage 4 TaLNet is fine-tuned on each speaker with a small learning rate. The model will thus be dedicated to one speaker to give further improvement in performance, as our results in the experiment section will show.

Experiments

Experiment Conditions

We used a multi-speaker dataset[1] containing ultrasound tongue imaging, optical lip video and audio for each utterance. The tongue was captured using a medical ultrasound transducer placed beneath the chin, and the lip videos were recorded using a video camera in front the mouth of the speakers. Further details about our dataset are described in our Technical Appendix. We used read-speech data from 75 speakers, which was split into a training set with a total of 10648 utterances, a validation set with a total of 750 utterances, and a test set with a total of 1800 utterances. The linguistic content of utterances was unique for each speaker in the training and validation sets, while it was the same across speakers in the test set (i.e. each speaker read the same set of test sentences). The content of the three sets was mutually exclusive with each other. The ultrasound tongue images, with a size of 64 × 842 for each frame, were recorded at 81.5 fps. The lip videos, with a shape of 240 × 320 for each frame, were recorded at 60 fps. Each ultrasound frame was resized to 64 × 128 pixels. Each video frame was randomly flipped horizontally, resized to 72 × 136 and then cropped to 64 × 128 pixels. The lip videos were resampled to 81.5 fps using *ffmpeg* in order to match the frame rate of the ultrasound stream. For acoustic features, 80-dimensional Mel-spectrograms were extracted with the Mel filter banks spanning from 80 Hz to 7600 Hz. They were then scaled by logarithm compression. For TTS pretraining, 460 hours data from 1150 speakers of the LibriTTS corpus[2] was used. Phoneme sequences were extracted using a grapheme-to-phoneme model as the inputs to the Tacotron 2 model. For speaker representation, x-vectors were extracted using the Kaldi toolkit[3].

Hyper-parameters in our experiments were determined according to how the model performed on our validation set. The Adam optimizer was used and the learning rate was determined by $lr = d_{model}^{-0.5} * min(step^{-0.5}, step * warmup^{-1.5})$, where d_{model} were set as 512, $step$ represents the number of training step counted from the second training stage. $warmup$ was set as 30k steps, which was also the beginning of the third stage. The scheduling sampling probability p increased linearly from 0 at 30k steps and reached 1 after 100k training steps. The model was kept training on the multi-speaker dataset until around 130k steps and then fine-tuned on each speaker. Details of the TaLNet encoders are presented in our Technical Appendix. The decoder structure follows that of Tacotron 2 except we have used a smaller LSTM layer with 512 units rather than 1024 in the original paper.

Comparison with Baselines

Three baseline methods were developed to compare with our proposed method, as follows:[4]

DNN Following recent DNN-based work on ultrasound-to-speech conversion (Csapó et al. 2017), 41-dimensional mel-cepstral coefficients (MCCs) and fundamental frequency (F0) were extracted by STRAIGHT (Kawahara,

[1] https://ultrasuite.github.io/data/tal_corpus/

[2] http://www.openslr.org/60/

[3] https://kaldi-asr.org/

[4] Samples from our experiments are available at https://jxzhanggg.github.io/TaLNet_demos/

Methods	MCD (dB)	STOI	CER (%)	WER (%)
DNN	3.81±0.31	0.52±0.05	52.7±12.2	75.7±14.2
BLSTM	3.31±0.15	0.59±0.04	35.9±9.1	53.6±11.8
TaLNet*	3.43±0.16	0.66±0.03	37.3±7.8	55.9±9.9
TaLNet	**3.22**±0.16	**0.69**±0.03	**23.3**±7.9	**36.5**±10.8

Table 1: Objective evaluation results comparing baselines and the proposed method for all speakers combined. Best results for each metric are highlighted in bold font. ± indicates the standard deviation of the metrics across speakers.

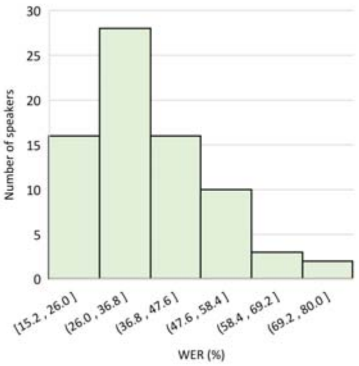
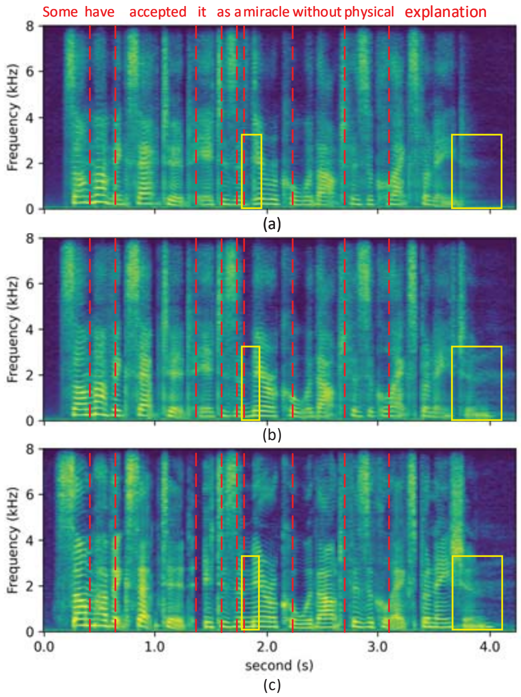


Figure 3: Histogram of per-speaker word error rate (WER) for TaLNet.



(a)

(b)

(c)

Figure 4: Spectrogram visualization of: (a) voice reconstructed by TaLNet without transfer learning (TaLNet*) (b) voice reconstructed by TaLNet, (c) natural speech. The color boxes indicate two pronunciation errors that occurred in (a), which correspond to phones /m/ and /ən/ respectively.

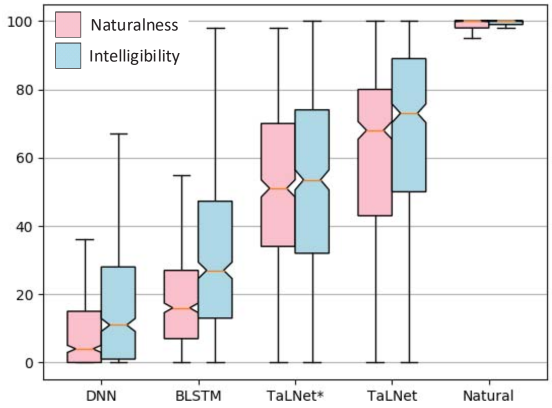


Figure 5: Naturalness and intelligibility ratings of the proposed method and three baselines obtained in a MUSHRA listening test. Natural denotes the natural target speech.

Masuda-Katsuse, and Cheveigné 1999) as acoustic features. Eigentongue and eigenlip transformations were used to extract 1000-dimensional feature vectors, and 5 consecutive frames were concatenated as inputs. The tongue and lips were processed by separate 3×1024 layer DNNs. Their outputs were concatenated and sent to another 3×1024 layer DNN to predict the acoustic features. Batch normalization and dropout were used before and after ReLU activation in each DNN layer.

BLSTM This model was made similar to the previously reported LipNet (Assael et al. 2017) and Lipper (Kumar et al. 2019). MCCs and F0 were extracted by STRAIGHT (Kawahara, Masuda-Katsuse, and Cheveigné 1999). The tongue and lip encoders were based on the same 3D CNN structures as in TaLNet. The visual representations from tongue and lips were then concatenated to be consumed by a two-layer bi-directional LSTM-RNN. Each LSTM layer had 256 units for each direction.

TaLNet* This model was the same as the proposed method except that the transfer learning strategy was not used. The model was randomly initialized and pretrained on the multi-speaker tongue and lip dataset, then finetuned on each speaker.

For objective evaluation, we report mel-cepstral distortion (MCD) and short-time objective intelligibility (STOI) as metrics. In order to further evaluate intelligibility, we have decoded the synthesized speech using an open-source auto-

# of spk.	Methods	MCD (dB)	STOI	CER (%)	WER (%)
75	TaLNet*	3.22	0.71	20.8	33.7
	TaLNet	**3.03**	**0.74**	**9.7**	**17.5**
25	TaLNet*	3.36	0.68	30.8	48.9
	TaLNet	**3.11**	**0.72**	**13.4**	**24.2**
9	TaLNet*	3.49	0.66	35.1	55.1
	TaLNet	**3.18**	**0.71**	**15.4**	**27.9**
3	TaLNet*	3.52	0.65	36.7	57.1
	TaLNet	**3.25**	**0.70**	**23.9**	**42.6**
1	TaLNet*	3.64	0.63	39.4	62.6
	TaLNet	**3.29**	**0.69**	**25.4**	**41.1**

Table 2: Objective evaluation results when varying the number of training speakers. All results calculated using the test set of speaker *70ms*, with the best highlighted in bold font.

matic speech recognition (ASR) model[5] based on ESPnet (Watanabe et al. 2018) and report character error rate (CER) and WER. For reference, the natural test speech measured 1.55% (CER) and 3.55% (WER).

As presented in Table 1, the DNN baseline obtained lower performance compared to the sequential models. The bidirectional long short-term memory (BLSTM) baseline achieved slightly better results than TaLNet* except for the STOI metric. The proposed method (TaLNet) achieved the best performance. In particular, there was a significant improvement of CER and WER over the baselines, which reflects higher intelligibility of reconstructed speech. WER distribution among speakers is further presented in Figure 3. These results show large variance across speakers, among which the least intelligible achieved a WER of 80% and the most intelligible a WER of 15.2%. Closer examination revealed that the least intelligible speakers often suffered from bad imaging quality, such as an unclear tongue edge or tilting of lip videos. Focusing on bad speaker data and improving performance there will be important future work. Figure 4 shows spectrograms of one utterance in the test set. We observe that the spectrogram of speech synthesized by the proposed method is close to the natural one. TaLNet performed better than the equivalent model without transfer learning in reconstruction accuracy. The colored boxes indicate two examples of pronunciation errors for TaLNet*, which were not present when transfer learning was used.

For evaluating the method subjectively, we also conducted a MUSHRA (MUlti-Stimulus test with Hidden Reference and Anchor) listening test to measure both naturalness and intelligibility. 100 test utterances from 25 speakers were randomly selected. Twenty native British English listeners were recruited on the Prolific[6] crow-sourcing platform. The stimuli were presented in random order and listeners were asked to give a score ranging from 0 to 100 (from least to most natural/intelligible). The results are shown in Figure 5. We can clearly see that the proposed TaLNet method outperformed the baselines significantly in terms of both naturalness and intelligibility. There is still, however, a gap between the recovered and natural utterances.

Varying the Number of Speakers

In this section, we investigate the relationship between the amount of training data and the resulting performance of the proposed method. To this end, we gradually removed speakers from the original training set, resulting in a series of training subsets, which contained 75, 25, 9, 3 and 1 speaker respectively. For equal comparison across the different configurations, performance was evaluated on the test set of the final remaining speaker (speaker ID *70ms*). In Table 2, we see that performance degraded as the number of speakers used for training decreased. The results also demonstrate the effectiveness of using the transfer learning strategy, which consistently improved model performance by a large margin when varying the amount of training data.

Ablation Studies

To evaluate the relative contributions of tongue ultrasound and lip video, ablation studies were conducted without using tongue data ("*w/o tongue*") or lip images ("*w/o lip*") respectively. We also tried to remove the average and standard deviation tongue and lip images ("*w/o stat.*"). To gauge the effectiveness of the scheduled sampling strategy, the model was trained with only teacher forcing ("*w/o ss*"). Results before fine-tuning (i.e., the last training stage) are also reported ("*w/o finetuning*").

We see in Table 3 that "*w/o lip*" greatly outperformed "*w/o tongue*". Best performance, though, was achieved when combining both features (TaLNet), indicating they complement each other well. "*w/o stat.*" achieved very close performance to the proposed method. This may be because the speaker-specific inputs were less useful after the model was fine-tuned on each speaker. The performance of "*w/o ss*" declined significantly compared to the proposed method, underlining its importance. Comparing "*w/o finetune*" with the proposed method demonstrates how fine-tuning can further improve the quality of reconstructed speech.

Voice Reconstruction from Silent Utterances

In this section, we explore voice reconstruction from silent utterances (i.e. speakers were asked to articulate with-

[5] https://drive.google.com/file/d/1BtQvAnsFvVi-dp_qsaFP7n4A_5cwnlR6/view?usp=drive_open

[6] https://www.prolific.co/

Methods	MCD (dB)	STOI	CER (%)	WER (%)
TaLNet	**3.22**±0.16	**0.69**±0.03	**23.3**±7.9	**36.5**±10.8
w/o tongue	4.20 ±0.21	0.52±0.04	68.3±3.3	90.9±3.7
w/o lip	3.38±0.16	0.66±0.03	35.3±8.0	53.0±10.2
w/o stat	3.22±0.15	0.69±0.03	23.5±7.0	36.9±9.9
w/o ss	3.55±0.19	0.63±0.04	36.8±8.5	54.3±10.6
w/o finetune	3.34±0.16	0.68±0.03	26.4±8.6	40.7±11.3

Table 3: Objective evaluation results of the proposed method in ablation studies. Best results for each metric are highlighted in bold font. ± indicates the standard deviation of the metrics across speakers.

Type of utt.	CER (%)	WER (%)
audible	22.4±8.4	34.4±10.6
silent	60.3±8.6	77.3±9.3

Table 4: Objective evaluation results of audible and silent utterances. ± indicates the standard deviation of the metrics across speakers.

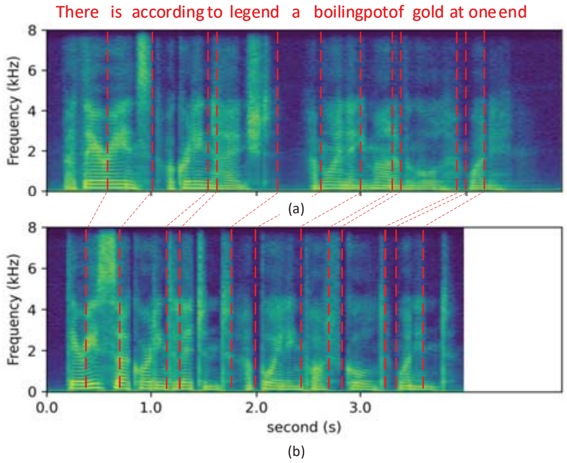


Figure 6: Spectrogram of (a) speech reconstructed from a silently articulated utterance, and (b) speech reconstructed from an equivalent audibly spoken utterance.

out emitting sound). Figure 6 compares spectrograms of utterances recovered from silently-articulated and normal speech. We observe that speech generated from silent speech has similar spectral patterns to audible speech, despite the model not having been trained on silent sentences. We also observe an interesting phenomenon, however, whereby silent utterances were 20% longer than the corresponding audible utterances on average. This indicates speakers spent more effort to speak silently without auditory feedback.

The silent and corresponding normal test utterances were evaluated, and CER and WER are reported in Table 4. MCD and STOI were omitted because the target speech signal was of course not available. We observe that the intelligibility of recovered speech was significantly degraded for silently spoken utterances. A similar effect was also observed in a previous study (Kimura, Kono, and Rekimoto 2019). Compared to speaking normally, speakers were observed to utter slightly differently when doing so silently. This may be caused by the lack of auditory feedback. However, Kimura et al. reported seeing speakers attempt to change their mouth movement over time to obtain better results, and generated speech improved after several trials (Kimura, Kono, and Rekimoto 2019). It is possible to implement our method online to introduce auditory feedback, which will be tackled in future work.

Conclusions

In this paper, we have proposed TaLNet, an articulatory-to-acoustic model with both tongue ultrasound and lip video as inputs. The model is based on the encoder-decoder structure. Transfer learning from text-to-speech models for the decoder has also been presented. Our method achieved the best performance in terms of naturalness and intelligibility compared to all baselines. We have used ablation studies to demonstrate the effectiveness of our proposed method. We have also presented promising results on silent utterance recovery. To further improve the intelligibility of silent utterances by introducing auditory feedback of speakers' speech will be our future work.

Acknowledgements

This work was partially funded by the National Key R&D Program of China (Grant No. 2019YFF0303001 and 2017YFB1002202), the National Nature Science Foundation of China (Grant No. 61871358), and the UK EPSRC Healthcare Partnerships Programme grant EP/P02338X/1 (Ultrax2020: http://www.ultrax-speech.org). The first author gratefully acknowledges financial support from the China Scholarship Council. Many thanks to the Ultrax2020 project team for technical help and feedback, and in particular to Manuel Sam Ribeiro for giving early access to the TaL corpus and for his many useful suggestions.

References

Afouras, T.; Chung, J. S.; and Zisserman, A. 2018. Deep Lip Reading: a comparison of models and an online application. In *INTERSPEECH*.

Akbari, H.; Arora, H.; Cao, L.; and Mesgarani, N. 2018. Lip2audspec: Speech reconstruction from silent lip movements video. In *2018 IEEE International Conference on*

Acoustics, Speech and Signal Processing (ICASSP), 2516–2520. IEEE.

Anumanchipalli, G. K.; Chartier, J.; and Chang, E. F. 2019. Speech synthesis from neural decoding of spoken sentences. *Nature* 568(7753): 493–498.

Aryal, S.; and Gutierrez-Osuna, R. 2016. Data driven articulatory synthesis with deep neural networks. *Computer Speech & Language* 36: 260–273.

Assael, Y. M.; Shillingford, B.; Whiteson, S.; and de Freitas, N. 2017. LipNet: Sentence-level lipreading. In *ICLR*, 1–13.

Baer, T.; Gore, J.; Boyce, S.; and Nye, P. 1987. Application of MRI to the analysis of speech production. *Magnetic resonance imaging* 5(1): 1–7.

Csapó, T. G.; Grósz, T.; Gosztolya, G.; Tóth, L.; and Markó, A. 2017. DNN-based ultrasound-to-speech conversion for a silent speech interface. In *INTERSPEECH*, 3672–3676. International Speech Communication Association (ISCA).

Den Oord, A. V.; Dieleman, S.; Zen, H.; Simonyan, K.; Vinyals, O.; Graves, A.; Kalchbrenner, N.; Senior, A. W.; and Kavukcuoglu, K. 2016. WaveNet: A Generative Model for Raw Audio. In *9th ISCA Speech Synthesis Workshop (SSW9)*, 125–125.

Denby, B.; Schultz, T.; Honda, K.; Hueber, T.; Gilbert, J. M.; and Brumberg, J. S. 2010. Silent speech interfaces. *Speech Communication* 52(4): 270–287.

Denby, B.; and Stone, M. 2004. Speech synthesis from real time ultrasound images of the tongue. In *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, I–685–I–688. IEEE.

Ephrat, A.; Halperin, T.; and Peleg, S. 2017. Improved speech reconstruction from silent video. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 455–462.

Ephrat, A.; and Peleg, S. 2017. Vid2speech: speech reconstruction from silent video. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5095–5099. IEEE.

Grósz, T.; Gosztolya, G.; Tóth, L.; Csapó, T. G.; and Markó, A. 2018. F0 estimation for DNN-based ultrasound silent speech interfaces. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 291–295. IEEE.

Hueber, T.; Aversano, G.; Cholle, G.; Denby, B.; Dreyfus, G.; Oussar, Y.; Roussel, P.; and Stone, M. 2007a. Eigentongue feature extraction for an ultrasound-based silent speech interface. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, volume 1, I–1245. IEEE.

Hueber, T.; and Bailly, G. 2016. Statistical conversion of silent articulation into audible speech using full-covariance HMM. *Computer Speech & Language* 36: 274–293.

Hueber, T.; Benaroya, E.-L.; Denby, B.; and Chollet, G. 2011. Statistical mapping between articulatory and acoustic data for an ultrasound-based silent speech interface. In *Twelfth Annual Conference of the International Speech Communication Association*.

Hueber, T.; Chollet, G.; Denby, B.; Dreyfus, G.; and Stone, M. 2007b. Continuous-speech phone recognition from ultrasound and optical images of the tongue and lips. In *Eighth Annual Conference of the International Speech Communication Association*.

Ji, Y.; Liu, L.; Wang, H.; Liu, Z.; Niu, Z.; and Denby, B. 2018. Updating the silent speech challenge benchmark with deep learning. *Speech Communication* 98: 42–50.

Jorgensen, C.; and Dusan, S. 2010. Speech interfaces based upon surface electromyography. *Speech Communication* 52(4): 354–366.

Kalchbrenner, N.; Elsen, E.; Simonyan, K.; Noury, S.; Casagrande, N.; Lockhart, E.; Stimberg, F.; van den Oord, A.; Dieleman, S.; and Kavukcuoglu, K. 2018. Efficient Neural Audio Synthesis. In Dy, J.; and Krause, A., eds., *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, 2410–2419. Stockholmsmässan, Stockholm Sweden: PMLR.

Kawahara, H.; Masuda-Katsuse, I.; and Cheveigné, A. D. 1999. Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency based F0 extraction: Possible role of a repetitive structure in sounds. *Speech Communication* 27(3–4): 187–207.

Kello, C. T.; and Plaut, D. C. 2004. A neural network model of the articulatory-acoustic forward mapping trained on recordings of articulatory parameters. *The Journal of the Acoustical Society of America* 116(4): 2354–2364.

Kimura, N.; Kono, M.; and Rekimoto, J. 2019. SottoVoce: an ultrasound imaging-based silent speech interaction using deep neural networks. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–11.

Kiritani, S. 1986. X-ray microbeam method for measurement of articulatory dynamics-techniques and results. *Speech Communication* 5(2): 119–140.

Kumar, Y.; Jain, R.; Salik, K. M.; Shah, R. R.; Yin, Y.; and Zimmermann, R. 2019. Lipper: Synthesizing thy speech using multi-view lipreading. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 2588–2595.

Li, N.; Liu, S.; Liu, Y.; Zhao, S.; and Liu, M. 2019. Neural speech synthesis with transformer network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 6706–6713.

Ling, Z.-H.; Richmond, K.; Yamagishi, J.; and Wang, R.-H. 2009. Integrating articulatory features into HMM-based parametric speech synthesis. *IEEE Transactions on Audio, Speech, and Language Processing* 17(6): 1171–1185.

Liu, L.; Ji, Y.; Wang, H.; and Denby, B. 2016. Comparison of DCT and autoencoder-based features for DNN-HMM multimodal silent speech recognition. In *2016 10th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, 1–5. IEEE.

Liu, Z.-C.; Ling, Z.-H.; and Dai, L.-R. 2016. Articulatory-to-Acoustic Conversion with Cascaded Prediction of Spectral and Excitation Features Using Neural Networks. In *Interspeech*, 1502–1506.

Liu, Z.-C.; Ling, Z.-H.; and Dai, L.-R. 2018. Articulatory-to-acoustic conversion using blstm-rnns with augmented input representation. *Speech Communication* 99: 161–172.

Ping, W.; Peng, K.; Gibiansky, A.; Arik, S. O.; Kannan, A.; Narang, S.; Raiman, J.; and Miller, J. P. 2018. Deep Voice 3: 2000-Speaker Neural Text-to-Speech. *International Conference on Learning Representations* .

Prajwal, K.; Mukhopadhyay, R.; Namboodiri, V. P.; and Jawahar, C. 2020. Learning Individual Speaking Styles for Accurate Lip to Speech Synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13796–13805.

Ribeiro, M. S.; Eshky, A.; Richmond, K.; and Renals, S. 2019. Speaker-independent classification of phonetic segments from raw ultrasound in child speech. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1328–1332. IEEE.

Schönle, P. W.; Gräbe, K.; Wenig, P.; Höhne, J.; Schrader, J.; and Conrad, B. 1987. Electromagnetic articulography: Use of alternating magnetic fields for tracking movements of multiple points inside and outside the vocal tract. *Brain and Language* 31(1): 26–35.

Shen, J.; Pang, R.; Weiss, R. J.; Schuster, M.; Jaitly, N.; Yang, Z.; Chen, Z.; Zhang, Y.; Wang, Y.; Skerry-Ryan, R. J.; et al. 2018. Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4779–4783.

Snyder, D.; Garcia-Romero, D.; Sell, G.; Povey, D.; and Khudanpur, S. 2018. X-vectors: Robust dnn embeddings for speaker recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5329–5333. IEEE.

Taguchi, F.; and Kaburagi, T. 2018. Articulatory-to-speech Conversion Using Bi-directional Long Short-term Memory. In *INTERSPEECH*, 2499–2503.

Tatulli, E.; and Hueber, T. 2017. Feature extraction using multimodal convolutional neural networks for visual speech recognition. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2971–2975. IEEE.

Toda, T.; Black, A. W.; and Tokuda, K. 2008. Statistical mapping between articulatory movements and acoustic spectrum using a Gaussian mixture model. *Speech Communication* 50(3): 215–227.

Tóth, L.; Gosztolya, G.; Grósz, T.; Markó, A.; and Csapó, T. G. 2018. Multi-Task Learning of Speech Recognition and Speech Synthesis Parameters for Ultrasound-based Silent Speech Interfaces. In *INTERSPEECH*, 3172–3176.

Wand, M.; Koutník, J.; and Schmidhuber, J. 2016. Lipreading with long short-term memory. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6115–6119. IEEE.

Wang, Y.; Skerry-Ryan, R. J.; Stanton, D.; Wu, Y.; Weiss, R. J.; Jaitly, N.; Yang, Z.; Xiao, Y.; Chen, Z.; Bengio, S.; et al. 2017. Tacotron: Towards End-to-End Speech Synthesis. In *Annual Conference of the International Speech Communication Association, INTERSPEECH*, 4006–4010.

Watanabe, S.; Hori, T.; Karita, S.; Hayashi, T.; Nishitoba, J.; Unno, Y.; Enrique, N.; Soplin, Y.; Heymann, J.; Wiesner, M.; Chen, N.; Renduchintala, A.; and Ochiai, T. 2018. ESPnet: End-to-End Speech Processing Toolkit. In *INTERSPEECH*, 2207–2211.

Xu, K.; Roussel, P.; Csapó, T. G.; and Denby, B. 2017. Convolutional neural network-based automatic classification of midsagittal tongue gestural targets using B-mode ultrasound images. *The Journal of the Acoustical Society of America* 141(6): EL531–EL537.

Yamamoto, R.; Song, E.; and Kim, J.-M. 2020. Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6199–6203. IEEE.