Speech Audio Synthesis from Tagged MRI and Non-negative Matrix Factorization via Plastic Transformer

Xiaofeng Liu[1](✉), Fangxu Xing[1], Maureen Stone[2], Jiachen Zhuo[2], Sidney Fels[3], Jerry L. Prince[4], Georges El Fakhri[1], and Jonghye Woo[1]

[1] Massachusetts General Hospital and Harvard Medical School, Boston, MA, USA
xliu61@mgh.harvard.edu
[2] University of Maryland, Baltimore, MD, USA
[3] University of British Columbia, Vancouver, BC, Canada
[4] Johns Hopkins University, Baltimore, MD, USA

Abstract. The tongue's intricate 3D structure, comprising localized functional units, plays a crucial role in the production of speech. When measured using tagged MRI, these functional units exhibit cohesive displacements and derived quantities that facilitate the complex process of speech production. Non-negative matrix factorization-based approaches have been shown to estimate the functional units through motion features, yielding a set of building blocks and a corresponding weighting map. Investigating the link between weighting maps and speech acoustics can offer significant insights into the intricate process of speech production. To this end, in this work, we utilize two-dimensional spectrograms as a proxy representation, and develop an end-to-end deep learning framework for translating weighting maps to their corresponding audio waveforms. Our proposed plastic light transformer (PLT) framework is based on directional product relative position bias and single-level spatial pyramid pooling, thus enabling flexible processing of weighting maps with variable size to fixed-size spectrograms, without input information loss or dimension expansion. Additionally, our PLT framework efficiently models the global correlation of wide matrix input. To improve the realism of our generated spectrograms with relatively limited training samples, we apply pair-wise utterance consistency with Maximum Mean Discrepancy constraint and adversarial training. Experimental results on a dataset of 29 subjects speaking two utterances demonstrated that our framework is able to synthesize speech audio waveforms from weighting maps, outperforming conventional convolution and transformer models.

1 Introduction

Intelligible speech is produced by the intricate three-dimensional structure of the tongue, composed of localized functional units [26]. These functional units, when measured using tagged magnetic resonance imaging (MRI), exhibit cohesive displacements and derived quantities that serve as intermediate structures linking

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14226, pp. 435–445, 2023.
https://doi.org/10.1007/978-3-031-43990-2_41

tongue muscle activity to tongue surface motion, which in turn facilitates the production of speech. A framework based on sparse non-negative matrix factorization (NMF) with manifold regularization can be used to estimate the functional units given input motion features, which yields a set of building blocks (or basis vectors) and a corresponding sparse weighting map (or encoding) [27]. The building blocks can form and dissolve with remarkable speed and agility, yielding highly coordinated patterns that vary depending on the specific speech task at hand. The corresponding weighting map can then be used to identify the cohesive regions and reveal the underlying functional units [25]. As such, by elucidating the relationship between the weighting map and intelligible speech, we can gain valuable insights for the development of speech motor control theories and the treatment of speech-related disorders.

Despite recent advances in cross-modal speech processing, translating between varied-size of wide 2D weighting maps and high-frequency 1D audio waveforms remains a challenge. The first obstacle is the inherent heterogeneity of their respective data representations, compounded by the tendency of losing pitch information in audio [1,6]. By contrast, transforming a 1D audio waveform into a 2D spectrogram provides a rich representation of the audio signal's energy distribution over the frequency domain, capturing both pitch and resonance information along the time axis [9,12]. Second, the input sizes of the weighting maps vary between $20 \times 5{,}745$ and $20 \times 11{,}938$, while the output spectrogram has a fixed size for each audio section. Notably, fully connected layers used in [1] require fixed size input, while the possible fully convolution neural networks (CNN) can have varied output sizes and unstable performance [23]. Third, modeling global correlations for the long column dimension of the weighting map and the lack of spatial local neighboring relationships in the row dimension presents further difficulties for conventional CNNs that rely on deep hierarchy structure for expanding the reception field [2,21]. Furthermore, the limited number of training pairs available hinders the large model learning process.

To address the aforementioned challenges, in this work, we propose an end-to-end translator that generates 2D spectrograms from 2D weighting maps via a heterogeneous plastic light transformer (PLT) encoder and a 2D CNN decoder. The lightweight backbone of PLT can efficiently capture the global dependencies with a wide matrix input in every layer [14]. Our PLT module is designed with directional product relative position bias and single-level spatial pyramid pooling to enable flexible global modeling of weighting maps with variable sizes, producing fixed-size spectrograms without information loss or dimension expansion due to cropping, padding, or interpolation for size normalization. To deal with a limited number of training samples, we explore pair-wise utterance consistency as prior knowledge with Maximum Mean Discrepancy (MMD) [8] in a disentangled latent space as an additional optimization objective. Additionally, a generative adversarial network (GAN) [10] can be incorporated to enhance the realism of the generated spectrograms.

The main contributions of this work are three-fold:

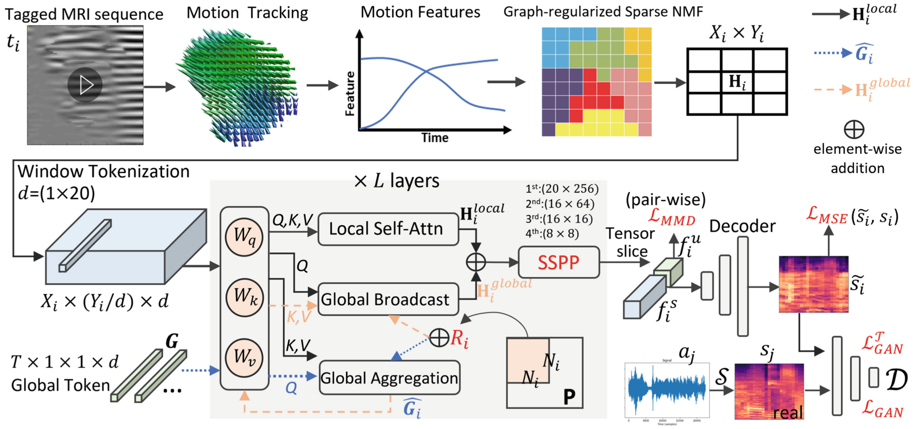


Fig. 1. Illustration of our translation framework. Only the NMF and translator with heterogeneous PLT encoder and 2D CNN decoder are used for testing.

- To our knowledge, this is the first attempt at relating functional units with audio waveforms by means of intermediate representations, including weighting maps and spectrograms.
- We developed a plastic light-transformer to achieve efficient global modeling of position sensitive weighting maps with variable sizes and long dimensions.
- We further explored the pair-wise utterance consistency constraint with MMD minimization and adversarial training as additional supervision signals to deal with relatively limited training samples.

Both quantitative and qualitative evaluation results demonstrate superior synthesis performance over comparison methods. Our framework has the potential to support clinicians and researchers in deepening their understanding of the interplay between tongue movements and speech waveforms, thereby improving treatment strategies for patients with speech-related disorders.

2 Methods

2.1 Preprocessing

During the training phase, we are given M pairs of synchronized tagged MRI sequences t_i and audio waveforms a_i, i.e., $\{t_i, a_i\}_{i=1}^{M}$. First, we apply a non-linear transformation using librosa to convert a_i into mel-spectrograms, denoted as s_i with the function $\mathcal{S} : a_i \rightarrow s_i$. This transformation uses a Hz-scale to emphasize human voice frequencies ranging from 40 to 1000 Hz Hz, while suppressing high-frequency instrument noise. Second, for each tagged MRI sequence t_i, we use a phase-based diffeomorphic registration method [29] to track the internal motion of the tongue. This allows us to generate corresponding weighting maps denoted as $\mathbf{H}_i$, which are based on input motion features $\mathbf{X}_i$, including the magnitude and angle of each track, by optimizing the following equation.

$$\mathcal{E} = \frac{1}{2}\|\mathbf{X}_i - \mathbf{W}_i\mathbf{H}_i\|_F^2 + \frac{1}{2}\lambda \mathrm{Tr}(\mathbf{H}_i\mathbf{L}_i\mathbf{H}_i^\top) + \eta\,\|\mathbf{H}_i\|_{1/2}, \tag{1}$$

where λ and η denote the weights associated with the manifold and sparse regularizations, respectively, and $\mathrm{Tr}(\cdot)$ represents the trace of a matrix. The graph Laplacian is denoted by $\mathbf{L}$.

2.2 Encoding Variable Size $\mathbf{H_i}$ with Plastic Light-Transformer

Directly modeling correlations among any two elements in a given weighting map $\mathbf{H}_i \in \mathbb{R}^{X_i \times Y_i}$ can impose quadratic complexity of $\mathcal{O}(X_i^2 Y_i^2)$. The recent efficient vision transformers (ViTs) [5,14,20,20,32] usually adopt a local patch design to compute local self-attention and correlate patches with CNNs. Specifically, the input is divided into $N_i = \frac{X_i}{P_x} \times \frac{Y_i}{P_y}$ patches[1], each of which is flattened to a token vector with a length of $d = P_x \times P_y$ [7]. The local self-attention is then formulated with a complexity of $\mathcal{O}(N_i d^2 = X_i Y_i d)$ as follows:

$$\mathbf{H}_i^{\mathrm{local}} = \mathrm{Attn}(\mathbf{H}_i^q, \mathbf{H}_i^k, \mathbf{H}_i^v) = \mathrm{SoftMax}(\frac{\mathbf{H}_i^q \mathbf{H}_i^{k\top}}{\sqrt{d}})\mathbf{H}_i^v, \in \mathbb{R}^{X_i \times Y_i}, \tag{2}$$

where vectors $\mathbf{H}_i^q$, $\mathbf{H}_i^k$, $\mathbf{H}_i^v \in \mathbb{R}^{N_i \times d}$ are produced by the linear projections of query (W_q), key (W_k), and value (W_v) branches, respectively [5,7,32]. The global correlation of ViTs with CNN [5,31,32] or window shifting [20], however, may not be efficient for our wide matrix $\mathbf{H}_i$, which lacks explicit row-wise neighboring features and may have a width that is too long for hierarchical convolution modeling. To address these challenges, we follow the lightweight ViT design [14], which uses a global embedding $\mathbf{G} \in \mathbb{R}^{T \times d}$ with $T \ll N_i$ randomly generated global tokens as the anchor for global information aggregation $\hat{\mathbf{G}}_i$. The aggregation is performed with attention of $\mathbf{G}^q, \mathbf{H}_i^k, \mathbf{H}_i^v$, which is then broadcasted with attention of $\mathbf{H}_i^q, \hat{\mathbf{G}}_i^k, \hat{\mathbf{G}}_i^v$ to leverage global contextual information [14].

While LightViT backbones have been shown to achieve wide global modeling within each layer [14], they are not well-suited for our variable size input and fixed size output translation. Although the self-attention scheme used in ViTs does not constrain the number of tokens, the absolute patch-position encoding in conventional ViTs [7] can only be applied to a fixed N_i [32], and the attention module will keep the same size of input and output. Notably, the number of tokens N_i will change depending on the size of $X_i \times Y_i$. As such, in this work, we resort to the directional product relative position bias [28] to add $\mathbf{R}_i \in \mathbb{R}^{N_i \times N_i}$, where element $r_{a,b} = \mathbf{p}_{\delta^x_{a,b}, \delta^y_{a,b}}$ is a trainable scalar, indicating the relative position weight between the patches a and b[2]. We set the offset of patch position in

[1] The bottom-right boundary is padded with 0 to ensure $X_i \% P_x = 0$ and $Y_i \% P_y = 0$.

[2] a learnable matrix $\mathbf{p} \in \mathbb{R}^{(2P_x-1)\times(2P_y-1)}$ is initialized with trunc_normal_, where $P_x = 20$ and $P_y = \frac{12000}{20} = 600$ are the maximum patch dimensions in our task.

x and y directions $\delta^x_{a,b} = x_a - x_b + P_x, \delta^y_{a,b} = y_a - y_b + M_y$ as the index in $\mathbf{p}$. Furthermore, the product relative position bias utilized in this work can distinguish between vertical or horizontal offsets, whereas the popular cross relative position bias [28] in computer vision tasks does not need to differentiate between time and spatial neighboring relationships in two dimensions.

Therefore, for global attention, we can aggregate the information of local tokens by modeling their global dependencies with

$$\hat{\mathbf{G}}_i = \mathrm{Attn}(\mathbf{G}^q, \mathbf{H}^k_i, \mathbf{H}^v_i) = \mathrm{SoftMax}(\frac{\mathbf{G}^q\mathbf{H}^{\mathbf{k}\top}_{\mathbf{i}} + \mathbf{R}_i}{\sqrt{d}})\mathbf{H}^v_i, \in \mathbb{R}^{X_i \times Y_i}. \tag{3}$$

Then, these global dependencies are broadcasted to every local token:

$$\mathbf{H}^{\mathrm{global}}_i = \mathrm{Attn}(\mathbf{H}^q_i, \hat{\mathbf{G}}^k_i, \hat{\mathbf{G}}^v_i) = \mathrm{SoftMax}(\frac{\mathbf{H}^q_i\hat{\mathbf{G}}^{\mathbf{k}\top}_{\mathbf{i}} + \mathbf{R}_i}{\sqrt{d}})\hat{\mathbf{G}}^v_i, \in \mathbb{R}^{X_i \times Y_i}. \tag{4}$$

By adding $\mathbf{H}^{\mathrm{local}}_i$ and $\mathbf{H}^{\mathrm{global}}_i$, each token can benefit from both local and global features, while maintaining linear complexity with respect to the input size. This brings noticeable improvements with negligible FLOPs increment. However, the sequentially proportional patch merging used in [5,14,32] still generates output sizes that vary with input sizes. Therefore, we utilize the single-level Spatial Pyramid Pooling (SSPP) [13] to extract a fixed-size feature for arbitrary input sizes. As illustrated in Fig. 1, the output of our channel-wise SSPP module with 20×256 bins has the size of $20 \times 256 \times d$, which can be a token merging scheme that adapts to the input size. Therefore, the final output of a layer is given by

$$\mathbf{H}'_i = \mathrm{SSPP}(\mathbf{H}^{\mathrm{local}}_i + \mathbf{H}^{\mathrm{global}}_i) \in \mathbb{R}^{X_i \times Y_i}. \tag{5}$$

We cascade four PLT layers with SSPP as our encoder to extract the feature representation $f_i \in \mathbb{R}^{8\times 8\times d}$. For the decoder, we adopt a simple 2D CNN with three deconvolutional layers to synthesize the spectrogram $\tilde{s}_i$.

2.3 Overall Training Protocol

We utilize the intermediate pairs of $\{\mathbf{H}_i, s_i\}^M_{i=1}$ to train our translator $\mathcal{T}$, which consists of a PLT encoder and a 2D CNN decoder. The quality of the generated spectrograms $\tilde{s}_i$ is evaluated using the mean square error (MSE) with respect to the ground truth spectrograms s_i:

$$\mathcal{L}_{\mathrm{MSE}} = ||\tilde{s}_i - s_i||^2_2 = ||\mathcal{T}(\mathbf{H}_i) - \mathcal{S}(a_i)||^2_2. \tag{6}$$

Additionally, we utilize the utterance consistency in the latent feature space as an additional optimization constraint. Specifically, we propose to disentangle f_i into two parts, i.e., utterance-related f^u_i and subject-related f^s_i. In practice, we split the utterance/subject-related parts channel-wise using tensor slicing method. Following the idea of deep metric learning [19], we aim to minimize the discrepancy between the latent features f^u_i and f^u_j of two samples

t_i and t_j that belong to the same utterance. Therefore, we use MMD [8] as an efficient discrepancy loss $\mathcal{L}_{\text{MMD}} = \gamma\text{MMD}(f_i^u, f_j^u)$, where $\gamma = 1$ or 0 for same or different utterance pairs, respectively.

Of note, the f_i^s is implicitly encouraged to incorporate the subject-related style of the articulation other than f_i^u with a complementary constraint [16,18] for reconstruction. Therefore, the decoder, which takes f_i^s conditioned on f_i^u can be considered as the utterance-conditioned spectrogram distribution modeling. This approach follows a divide-and-conquer strategy [3,17] for each utterance and can be particularly efficient for relatively few utterance tasks.

A GAN model can be further utilized to boost the realism of $\tilde{s}_i$. A discriminator $\mathcal{D}$ is employed to differentiate whether the mel-spectrogram is real $s_i = \mathcal{S}(a_i)$ or generated $\tilde{s}_i = \mathcal{T}(\mathbf{H}_i)$ with the following binary cross-entropy loss:

$$\mathcal{L}_{\text{GAN}} = \mathbb{E}_{s_i}\{\log(\mathcal{D}(s_i))\} + \mathbb{E}_{\tilde{s}_i}\{\log(1 - \mathcal{D}(\tilde{s}_i))\}. \tag{7}$$

In adversarial training, the translator $\mathcal{T}$ attempts to confuse $\mathcal{D}$ by optimizing $\mathcal{L}_{GAN}^{\mathcal{T}} = \mathbb{E}_{\tilde{s}_i}\{-\log(1 - \mathcal{D}(\tilde{s}_i))\}$. Of note, $\mathcal{T}$ does not involve real spectrograms in $\log(\mathcal{D}(s_i'))$ [24]. Therefore, the overall optimization objectives of our translator $\mathcal{T}$ and discriminator $\mathcal{D}$ are expressed as:

$$\min_{\mathcal{T}} \mathcal{L}_{\text{MSE}} + \beta\mathcal{L}_{\text{MMD}} + \lambda\mathcal{L}_{\text{GAN}}^{\mathcal{T}}; \quad \min_{\mathcal{D}} \mathcal{L}_{\text{GAN}}, \tag{8}$$

where β and λ represent the weighting parameters. Notably, only $\mathcal{T}$ is utilized in testing, and we do not need pairwise inputs for utterance consistency. Recovering audio waveform from mel-spectrogram can be achieved by the well-established Griffin-Lim algorithm [11] in the Librosa toolbox.

3 Experiments and Results

For evaluation, we collected paired 3D tagged MRI sequences and audio waveforms from a total of 29 subjects, while performing the speech words "a souk" or "a geese," with a periodic metronome-like sound as guidance [15,30]. The tagged-MRI sequences consisted of 26 frames, which were resized to 128×128. The resulting $\mathbf{H}$ matrix varied in size from $20 \times 5{,}745$ to $20 \times 11{,}938$ (we set one dimension to a constant value of 20.) The audio waveforms had varying lengths between 21,832 to 24,175. To augment the dataset, we employed a sliding window technique on each audio, allowing us to crop sections with 21,000 time points, resulting in 100 audio waveforms. Then, we utilized the Librosa library to convert all audio waveforms into mel-spectrograms with a size of 64×64. For our evaluation, we utilized a subject-independent leave-one-out approach. For the data augmentation of the $\mathbf{H}$ matrix, we randomly drop the column to round Y_i to the nearest hundred, e.g., 9,882 to 9,800, generating 100 versions of $\mathbf{H}$. We utilized the leave-one-out evaluation, following a subject-independent manner.

In our implementation, we set $P_x = 1$ and $P_y = 20$, i.e., $d = 20$. Our encoder consisted of four PLT encoder layers with SSPP, to extract a feature f_i with the size of $8 \times 8 \times 20$. Specifically, the first $8 \times 8 \times 4$ component was

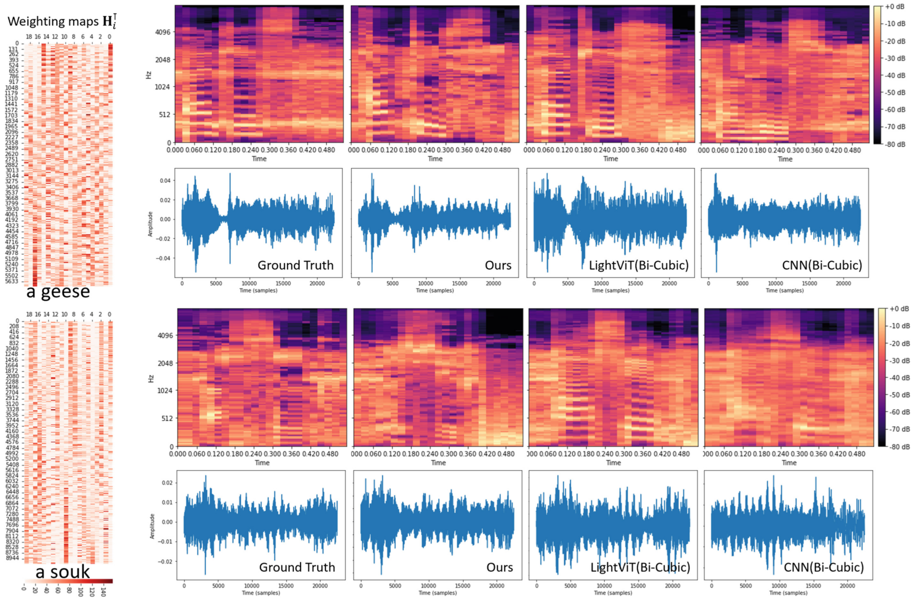


Fig. 2. Comparisons of our PLT with CNN and LightViT using bi-cubic interpolation. We show $\mathbf{H}_i^\top$ for compact layout. Audios are attached in supplementary.

set as the utterance-related factors, and the remaining 16 channels were for the subject-specific factors. Then, the three 2D de-convolutional layers were applied as our decoder to generate the 64 × 64 mel-spectrogram. The activation units in our model were rectified linear units (ReLU), and we normalized the final output of each pixel using the sigmoid function. The discriminator in our model consisted of three convolutional layers and two fully connected layers, and had a sigmoid output. A detailed description of the network structure is provided in the supplementary material, due to space limitations.

Our model was implemented using PyTorch and trained 200 epochs for approximately 6 h on a server equipped with an NVIDIA V100 GPU. Notably, the inference from a $\mathbf{H}$ matrix to audio took less than 1 s, depending on the size of $\mathbf{H}$. Also, the pairwise utterance consistency and GAN training were only applied during the training phase and did not affect inference. For our method and its ablation studies, we consistently set the learning rates of our heterogeneous translator and discriminator to $lr^{\mathcal{T}} = 10^{-3}$ and $lr^{\mathcal{D}} = 10^{-4}$, respectively, with a momentum of 0.5. The loss trade-off hyperparameters were set as $\beta = 0.75$, and we set $\lambda = 1$.

It is important to note that without NMF, generating intelligible audio with a small number of subjects using video-based audio translation models, such as Lip2AudSpect [1], is not feasible. As an alternative, we pre-processed the input by cropping, padding with zeros, or using bi-cubic interpolation to obtain

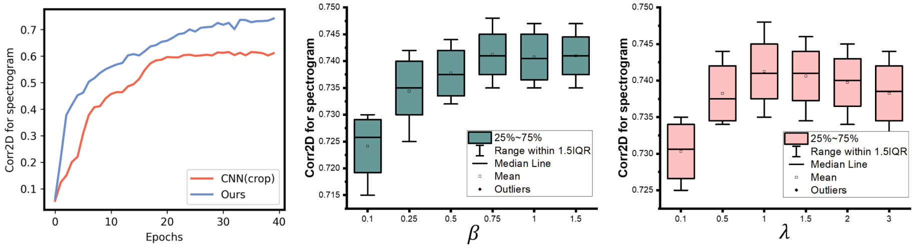
Table 1. Numerical comparisons during testing using leave-one-out evaluation

Encoder Models	Corr2D for spectrogram ↑	PESQ for waveform ↑
CNN (Crop)	0.614±0.013	1.126±0.021
CNN (Padding 0)	0.684±0.010	1.437±0.018
CNN (Bi-Cubic)	0.689±0.012	1.451±0.020
CNN+SSPP	0.692±0.017	1.455±0.022
LightViT (Crop)	0.635±0.015	1.208±0.022
LightViT (Padding 0)	0.708±0.011	1.475±0.015
LightViT (Bi-Cubic)	0.702±0.012	1.492±0.018
Ours	**0.742**±0.012	**1.581**±0.020
Ours with cross embedding	0.720±0.013	1.550±0.021
Ours w/o Pair-wise Disentangle	0.724±0.010	1.548±0.019
Ours w/o GAN	0.729±0.011	1.546±0.020

a fixed-size input **H**. We then compared the performance of our encoder module with conventional CNN or LightViT [14].

Figure 2 shows a qualitative comparison of our PLT framework with CNN and LightViT [14] using bi-cubic interpolation. We can observe that our generated spectrogram and the corresponding audio waveforms demonstrate superior alignment with the ground truth. It is worth noting that the CNN model or the CNN-based global modeling ViTs [5,31] require deep models to achieve large receptive fields [2,21]. Moreover, the interpolation process adds significant computational complexity for both CNN and LightViT, making it difficult to train on a limited dataset. In Fig. 3(a), we show that our proposed PLT framework achieves a stable performance gain along with the training and outperforms CNN with the crop, which lost the information of some functional units.

Following [1], we used 2D Pearson's correlation coefficient (Corr2D) [4], and Perceptual Evaluation of Speech Quality (PESQ) [22] as our evaluation metrics to measure the synthesis quality of spectrograms in the frequency domain, and waveforms in the time domain, respectively. The numerical comparisons of different encoder structures with conventional CNN or LightViT with different crop or padding strategies and our PLT framework are provided in Table 1. The standard deviation was obtained from three independent random trials. Our framework outperformed CNN and lightViT consistently. In addition, the synthesis performance was improved by pair-wise disentangled utterance consistency MMD loss and GAN loss, as demonstrated in our ablation studies. Furthermore, it outperformed the in-directional cross relative position bias [28], since two dimensions in the weighting map indicate time and spatial relationship, respectively. Notably, even though CNN with SSPP can process varied size inputs, it suffers from limited long-term modeling capacity [2,21] and unstable performance [23]. The sensitivity analysis of our loss weights are given in Fig. 3(b) and (c), where the performance was relatively stable for $\beta \in [0.75, 1.5]$ and $\lambda \in [1, 2]$.



(a) (b) (c)

Fig. 3. (a) Comparison of Corr2D using our plastic light transformer and CNN with crop. Sensitivity analysis of β (b) and λ (c).

4 Conclusion

This work aimed to explore the relationship between tongue movements and speech acoustics by translating weighting maps, which represent the functional units of the tongue, to their corresponding audio waveforms. To achieve this, we proposed a deep PLT framework that can handle variable-sized weighting maps and generated fixed-sized spectrograms, without information loss or dimension expansion. Our framework efficiently modeled global correlations in wide matrix input. To improve the realism of the generated spectrograms, we applied pairwise utterance consistency with MMD constraint and adversarial training. Our experimental results demonstrated the potential of our framework to synthesize audio waveforms from weighting maps, which can aid clinicians and researchers in better understanding the relationship between the two modalities.

Acknowledgements. This work is supported by NIH R01DC014717, R01DC018511, R01CA133015, and P41EB022544.

References

1. Akbari, H., Arora, H., Cao, L., Mesgarani, N.: Lip2audspec: speech reconstruction from silent lip movements video. In: ICASSP, pp. 2516–2520. IEEE (2018)
2. Araujo, A., Norris, W., Sim, J.: Computing receptive fields of convolutional neural networks. Distill **4**(11), e21 (2019)
3. Che, T., et al.: Deep verifier networks: verification of deep discriminative models with deep generative models. In: AAAI (2021)
4. Chi, T., Ru, P., Shamma, S.A.: Multiresolution spectrotemporal analysis of complex sounds. J. Acoust. Soc. Am. **118**(2), 887–906 (2005)
5. Chu, X., et al.: Twins: revisiting the design of spatial attention in vision transformers. Adv. Neural. Inf. Process. Syst. **34**, 9355–9366 (2021)
6. Chung, J.S., Zisserman, A.: Lip reading in the wild. In: Lai, S.-H., Lepetit, V., Nishino, K., Sato, Y. (eds.) ACCV 2016. LNCS, vol. 10112, pp. 87–103. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-54184-6_6

7. Dosovitskiy, A., et al.: An image is worth 16x16 words: transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
8. Dziugaite, G.K., Roy, D.M., Ghahramani, Z.: Training generative neural networks via maximum mean discrepancy optimization. arXiv preprint arXiv:1505.03906 (2015)
9. Ephrat, A., Peleg, S.: Vid2speech: speech reconstruction from silent video. In: ICASSP, pp. 5095–5099. IEEE (2017)
10. Goodfellow, I., et al.: Generative adversarial networks. Commun. ACM **63**(11), 139–144 (2020)
11. Griffin, D., Lim, J.: Signal estimation from modified short-time Fourier transform. IEEE Trans. Acoust. Speech Signal Process. **32**(2), 236–243 (1984)
12. He, G., Liu, X., Fan, F., You, J.: Image2audio: facilitating semi-supervised audio emotion recognition with facial expression image. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp. 912–913 (2020)
13. He, K., Zhang, X., Ren, S., Sun, J.: Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Trans. Pattern Anal. Mach. Intell. **37**(9), 1904–1916 (2015)
14. Huang, T., Huang, L., You, S., Wang, F., Qian, C., Xu, C.: Lightvit: towards light-weight convolution-free vision transformers. arXiv preprint arXiv:2207.05557 (2022)
15. Lee, J., Woo, J., Xing, F., Murano, E.Z., Stone, M., Prince, J.L.: Semi-automatic segmentation of the tongue for 3D motion analysis with dynamic MRI. In: ISBI, pp. 1465–1468. IEEE (2013)
16. Liu, X., Chao, Y., You, J.J., Kuo, C.C.J., Vijayakumar, B.: Mutual information regularized feature-level Frankenstein for discriminative recognition. IEEE TPAMI **44**, 5243–5260 (2021)
17. Liu, X., et al.: Domain generalization under conditional and label shifts via variational Bayesian inference. IJCAI (2021)
18. Liu, X., et al.: Feature-level Frankenstein: eliminating variations for discriminative recognition. In: CVPR, pp. 637–646 (2019)
19. Liu, X., Vijaya Kumar, B., You, J., Jia, P.: Adaptive deep metric learning for identity-aware facial expression recognition. In: CVPR, pp. 20–29 (2017)
20. Liu, Z., et al.: Swin transformer: hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10012–10022 (2021)
21. Luo, W., Li, Y., Urtasun, R., Zemel, R.: Understanding the effective receptive field in deep convolutional neural networks. In: Advances in Neural Information Processing Systems, vol. 29 (2016)
22. Recommendation, I.T.: Perceptual evaluation of speech quality PESQ): an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs. Rec. ITU-T P. 862 (2001)
23. Richter, M.L., Byttner, W., Krumnack, U., Wiedenroth, A., Schallner, L., Shenk, J.: (Input) size matters for CNN classifiers. In: Farkaš, I., Masulli, P., Otte, S., Wermter, S. (eds.) ICANN 2021. LNCS, vol. 12892, pp. 133–144. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-86340-1_11
24. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X.: Improved techniques for training GANs. NIPS **29**, 2234–2242 (2016)
25. Woo, J., et al.: A sparse non-negative matrix factorization framework for identifying functional units of tongue behavior from MRI. IEEE Trans. Med. Imaging **38**(3), 730–740 (2018)

26. Woo, J., et al.: A deep joint sparse non-negative matrix factorization framework for identifying the common and subject-specific functional units of tongue motion during speech. Med. Image Anal. **72**, 102131 (2021)
27. Woo, J., et al.: Identifying the common and subject-specific functional units of speech movements via a joint sparse non-negative matrix factorization framework. In: Medical Imaging 2020: Image Processing, vol. 11313, pp. 446–451. SPIE (2020)
28. Wu, K., Peng, H., Chen, M., Fu, J., Chao, H.: Rethinking and improving relative position encoding for vision transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10033–10041 (2021)
29. Xing, F., et al.: Phase vector incompressible registration algorithm for motion estimation from tagged magnetic resonance images. IEEE TMI **36**(10), 2116–2128 (2017)
30. Xing, F., Woo, J., Murano, E.Z., Lee, J., Stone, M., Prince, J.L.: 3D tongue motion from tagged and cine MR images. In: Mori, K., Sakuma, I., Sato, Y., Barillot, C., Navab, N. (eds.) MICCAI 2013. LNCS, vol. 8151, pp. 41–48. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-40760-4_6
31. Yang, J., et al.: Focal self-attention for local-global interactions in vision transformers. arXiv preprint arXiv:2107.00641 (2021)
32. Zhang, Q., Yang, Y.B.: Rest: an efficient transformer for visual recognition. Adv. Neural. Inf. Process. Syst. **34**, 15475–15485 (2021)