Speech Audio Generation from Dynamic MRI via a Knowledge Enhanced Conditional Variational Autoencoder

Yaxuan Li[1⋆‡], Han Jiang[2⋆], Yifei Ma[1], Shihua Qin[3], Jonghye Woo[4], and Fangxu Xing[4†]

[1] Department of Computer Science, The University of Hong Kong, Pokfulam, Hong Kong
[2] School of Software Engineering, Xi'an Jiaotong University, Xi'an, Shaanxi, China
[3] Wake Forest University School of Medicine, Medical Center Boulevard, Winston-Salem, NC, USA
[4] Department of Radiology, Harvard Medical School, Boston 08544, USA
fxing1@mgh.harvard.edu

Abstract. Dynamic Magnetic Resonance Imaging (MRI) of the vocal tract has become an increasingly adopted imaging modality for speech motor studies. Beyond image signals, systematic data loss, noise pollution, and audio file corruption can occur due to the unpredictability of the MRI acquisition environment. In such cases, generating audio from images is critical for data recovery in both clinical and research applications. However, this remains challenging due to hardware constraints, acoustic interference, and data corruption. Existing solutions, such as denoising and multi-stage synthesis methods, face limitations in audio fidelity and generalizability. To address these challenges, we propose a Knowledge Enhanced Conditional Variational Autoencoder (KE-CVAE), a novel two-step "knowledge enhancement + variational inference" framework for generating speech audio signals from cine dynamic MRI sequences. This approach introduces two key innovations: (1) integration of unlabeled MRI data for knowledge enhancement, and (2) a variational inference architecture to improve generative modeling capacity. To the best of our knowledge, this is one of the first attempts at synthesizing speech audio directly from dynamic MRI video sequences. The proposed method was trained and evaluated on an open-source dynamic vocal tract MRI dataset recorded during speech. Experimental results demonstrate its effectiveness in generating natural speech waveforms while addressing MRI-specific acoustic challenges, outperforming conventional deep learning-based synthesis approaches[1].

Keywords: Speech audio generation · dynamic MRI · vocal tract · knowledge enhancement · variational inference.

⋆ Equal contribution. †Corresponding author.
‡ This work was done by Yaxuan Li as a research assistant at Harvard Medical School
[1] https://github.com/YaxuanLi-cn/KE-CVAE

1 Introduction

Capturing real-time deformations of the vocal tract during human speech through medical imaging is essential for various speech research applications [29,17], as accurately characterizing the functional behavior of vocal articulators has multiple clinical implications. Several imaging modalities are currently used for this task, including electromagnetic articulography (EMA), ultrasound, and magnetic resonance imaging (MRI). However, EMA requires attaching multiple sensors to the articulators, which can potentially disrupt natural speech patterns, and it can only track a limited number of points on the tongue's surface [25]. Although ultrasound is non-invasive, it has a restricted field of view and struggles to clearly visualize deeper structures such as the palate and pharyngeal walls [21]. In contrast, MRI offers a more effective solution by providing high-contrast anatomical imaging with high resolution while remaining non-invasive. Advances in MRI technology have led to the development of high-speed dynamic cine imaging, enabling rapid and real-time imaging for in vivo speech [16,9].

However, simultaneously recording the subjects' speech audio signals during image acquisition remains challenging due to several factors. (1) Hardware constraints require specialized MRI-compatible microphones that function within strong magnetic fields without interference from the MRI scanner or to the image quality [8]. (2) Acoustic interference from MRI scanners (the loud pulse sequence sounds) generates substantial and unpredictable noise, varying with scanning protocols/parameters and making it difficult to capture clean recordings of human voice [2]. (3) Data integrity issues arise as the process of dynamic MRI reconstruction can affect sound recordings, with synchronization difficulties, MRI-induced noise, missing data or artifacts leading to corrupted or incomplete audio segments [23]. Despite these challenges, capturing speech waveforms remains a critical requirement in many speech imaging studies [18,30].

Multiple solutions have been proposed to obtain high-quality speech sound files in MRI environments. Denoising techniques enhance audio clarity but can only reduce noise rather than fully eliminate it and cannot reconstruct missing speech segments [7,10]. More recently, an innovative approach has emerged that synthesizes speech directly from MRI data. Liu et al. [18] developed a system that converts 4D tagged-MRI data into audio using Non-negative Matrix Factorization (NMF). While effective for processing short phrases, this method relies on a precomputed sequence of deformation fields to synthesize audio, limiting its ability to directly process dynamic cine images. Additionally, its multi-step process is computationally intensive and may not generalize well to longer or more varied speech patterns. Although these methods represent significant progress, they still fall short of producing fully accurate and complete speech sound files for comprehensive research applications.

In non-medical imaging settings, several methods have been proposed for audio generation [12,15] that utilize an end-to-end variational inference framework, producing more natural-sounding signals than earlier two-stage models. Inspired by these approaches, we propose a Knowledge Enhanced Conditional Variational Autoencoder (KE-CVAE) with two key innovations: (1) A novel two-

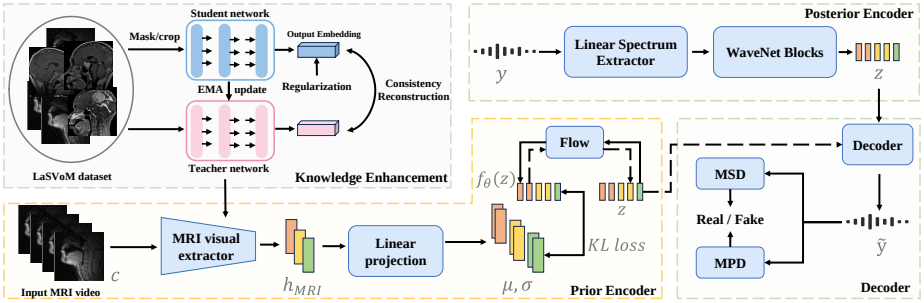

Fig. 1: Illustration of our KE-CVAE model. The dashed lines in the figure represent the inference process.

step "knowledge enhancement + variational inference" framework for speech audio generation from dynamic MRI sequences. Specifically, we have compiled a collection of thirteen public head and neck MRI image/video datasets for the training process using a self-supervised knowledge enhancement strategy. (2) Adoption of a variational inference framework to enhance the expressive power of generative modeling (Figure 1). We conducted comprehensive experiments to generate speech sound using the open-access speech dynamic MRI dataset described in [16], demonstrating the effectiveness of KE-CVAE over conventional CNN and transformers in multiple metrics such as correlation, PESQ, and a subjective MOS score (detailed in the results section).

2 Methods

2.1 Knowledge Enhancement with Domain-Specific Data

Recent advances in self-supervised pre-training on images [4,24] in computer vision have demonstrated that it is possible to learn robust and meaningful low-dimensional features without explicitly labeled supervision. Inspired by this concept, we propose a self-supervised knowledge enhancement strategy to learn latent variables from large-scale MRI data. Specifically, our model comprises a teacher network and a student network (Figure 1), both based on vision transformers (ViTs) [19], denoted as $F_{teacher}$ and $F_{student}$, respectively. By leveraging a large-scale curated set of vocal tract MRI, our strategy enables the visual extractor to focus on articulatory-relevant regions and capture subtle distinctions in vocal tract configurations during speech production. To effectively learn domain-specific knowledge without annotations, we introduce three complementary loss functions: consistency loss for representation alignment, reconstruction loss for masked image modeling, and KoLeo regularization for feature distribution optimization.

Consistency Loss. The consistency loss is designed to maximize the alignment of the output classification ($[CLS]$) token embeddings between $F_{teacher}$

and $F_{student}$, facilitating better representation learning. Specifically, we pass the $[CLS]$ token through a standard multilayer perceptron (MLP) model to generate a vector of scores, followed by a softmax function to obtain the pseudo-class probability p. We compute consistency loss at both the global and local levels, corresponding to image-level and patch-level augmentations, respectively. The global consistency loss is defined as the cross-entropy loss between the teacher output p_t and the student output p_s. The local consistency loss is computed as the cross-entropy loss between the $[CLS]$ embeddings of both networks, p_{sl} and p_{tl}. We apply Sinkhorn-Knopp centering [3] to the teacher network output to enhance distribution alignment. The final consistency loss integrates both global and local components, formulated as follows:

$$\mathcal{L}_{con} = -\sum_{i=1}^{N} p_t^{(i)} \log p_s^{(i)} - \sum_{i=1}^{N}\sum_{j=1}^{M} p_{tl}^{(i,j)} \log p_{sl}^{(i,j)}, \tag{1}$$

where N is the batch size and M is the number of patches per sample.

Reconstruction Loss. We randomly replace the patches in the input image with masks using the $[MASK]$ token or random patch features with a certain probability. The masked input MRI image $\mathbf{V}_i$ can be represented as $\mathbf{V}_{i\_mask}$ where the embeddings of some patches $\{v_i\}_{i\in B}$ in $\mathbf{V}_{i\_mask}$ are replaced by trainable $[MASK]$ token embeddings. The reconstruction loss is computed by comparing the output embeddings of the $[MASK]$ tokens from both the teacher and student models:

$$\mathcal{L}_{rec} = -\sum_{i=1}^{N}\sum_{j=1}^{M} h_j \cdot F_{teacher}(\mathbf{V}_i) \log F_{student}(\mathbf{V}_{i\_mask}), \tag{2}$$

where $h_j = 1$ indicates that the token at position j has been masked, and $h_j = 0$ indicates otherwise. Incorporating this reconstruction loss complements the consistency loss, facilitating the learning of robust MRI image embeddings at multiple levels.

KoLeo Regularization Loss. The KoLeo regularization loss [6] has been designed to encourage a uniform span of the features within a batch. Given a batch of N output $[CLS]$ token embeddings from $\{F_{student}(\mathbf{V}_i)[CLS]\}_{i=1}^{N}$ of the student network, KoLeo regularization loss is defined as:

$$\mathcal{L}_{KoLeo} = -\frac{1}{N}\sum_{i=1}^{N} \log(d_{N,i}), \tag{3}$$

where $d_{N,i} = \min_{j\neq i} \|F_{student}(\mathbf{V}_i)[CLS] - F_{student}(\mathbf{V}_j)[CLS]\|$ is the minimal distance between $F_{student}(\mathbf{V}_i)[CLS]$ and any other student network $[CLS]$ embedding within the batch. $F_{student}(\mathbf{V}_i)[CLS]$ is also ℓ_2-normalized before computing the KoLeo regularization loss.

2.2 Variational Inference Framework on Dynamic Speech MRI

We formulate the proposed model KE-CVAE as a conditional variational autoencoder. In this framework, $p(y|z)$ represents the likelihood function for gen-

erating an audio waveform data point y given the latent variable z, while $q(z|y)$ denotes the approximate posterior distribution that infers z from y. Additionally, $p(z|c)$ describes the prior distribution of the latent variables z, conditioned on the MRI sequence c. The objective of the CVAE is to maximize the variational lower bound, also known as the evidence lower bound (ELBO), of the intractable marginal log-likelihood of the data $\log p(y|c)$. This objective can be decomposed into two terms: the reconstruction loss and the KL divergence, expressed as follows:

$$\mathcal{L}_{ELBO} = \mathcal{L}_{recon} + KL(q(z|y) \,||\, p(z|c)). \tag{4}$$

The reconstruction loss, $\mathcal{L}_{recon}$, is defined as $\mathcal{L}_{recon} = \|y_{mel} - \tilde{y}_{mel}\|_1$, where y_{mel} denotes the mel-spectrogram of the input audio, and $\tilde{y}_{mel}$ represents the mel-spectrogram reconstructed by the decoder. The $\|\cdot\|_1$ notation refers to the ℓ_1 norm.

The KE-CVAE model primarily comprises three components: a posterior encoder, a prior encoder, and a decoder. We detail the three modules as follows:
Posterior Encoder. The posterior encoder extracts the latent representation $z \sim q(z|y)$ from the input waveform y. We first transform the raw waveform into its linear spectrum. Several non-causal WaveNet residual blocks [11] are applied to extract an embedding sequence. Then we employ a linear layer to project the mean and variance from the normal posterior distribution $p(z|y)$.
Prior Encoder. In our proposed setting, dynamic MRI sequences serve as the condition for speech generation. The prior encoder models the conditional prior distribution $p(z|c)$ given the MRI sequence c. As described in Section 2.1, the pre-trained MRI visual extractor $F_{teacher}$ is applied in this module to obtain the hidden representation h_{MRI}. The linear projection layer following the blocks produces the mean μ and variance σ of the normal posterior distribution. To further improve the scalability of the approximated posterior distributions, we use a normalizing flow f_θ to apply a sequence of invertible transformations [28]. Therefore, the KL divergence is calculated by:

$$\begin{aligned} KL(q(z|y) \,||\, p(z|c)) &= \log q(z|y) - \log p(z|c), \\ p(z|c) &= \mathcal{N}(f_\theta(z); \mu(c), \sigma(c)) \left| \det \frac{\partial f_\theta(z)}{\partial z} \right|. \end{aligned} \tag{5}$$

where det denotes the determinants.
Decoder. The decoder uses the latent variable z to reconstruct the waveform $\tilde{y} \sim p(y|z)$. In addition, we adopt the adversarial training strategy. The discriminator D follows HiFi-GAN's multi-period discriminator (MPD) and multi-scale discriminator (MSD) architecture [13]. Specifically, the adversarial losses [22] for the generator G and the discriminator D are defined as:

$$\mathcal{L}_{adv}(D) = \mathbb{E}_{(y,z)} \left[(D(y) - 1)^2 + (D(G(z)))^2 \right], \tag{6}$$

$$\mathcal{L}_{adv}(G) = \mathbb{E}_z \left[(D(G(z)) - 1)^2 \right]. \tag{7}$$

where $\mathbb{E}$ denotes the expectation. Specifically, we utilize the feature matching loss [14] as an element-wise reconstruction loss for more stable training procedure.

3 Experiments and Results

3.1 Datasets and Evaluation Metrics

LaSVoM Dataset. To address the scarcity of large-scale head and neck MRI resources, we constructed a "large-scale vocal track MRI" data collection (LaSVoM) by aggregating and curating data from thirteen public MRI image and video datasets in the head and neck region. Through systematic frame extraction and resizing, we collected over 30,000 high-quality mid-sagittal head and neck MRI images, each resized to focus on the vocal tract region with a size of 84×84 pixels. These images encompass a wide range of phonetic contexts, speakers, and articulation patterns, providing a comprehensive resource for the proposed knowledge enhancement process.

Variational Inference Dataset. We utilized the open dynamic MRI in speech dataset in [16], which consists of cine MRI capturing dynamic vocal tract movements in the sagittal plane, accompanied by synchronized audio recordings. The dataset includes cine sequences from 75 participants performing various speech tasks. For training and testing, we split the dataset into an 8:2 ratio.

Evaluation Metrics. We perform all experiments in a speaker-independent manner so that the metrics we used are not affected by the change of speakers. We calculated conventional objective metrics for audio waveform evaluation, including 2D Pearson's correlation coefficient (Corr2D) [5] and perceptual evaluation of speech quality (PESQ) [27]. We also performed a subjective MOS (mean opinion score) test to evaluate all compared methods. Specifically, we randomly selected 30 from 1662 test audios for subjective listening and asked 10 human raters to assess their quality. The raters evaluated the audio quality by comparing each reconstructed audio sample with its corresponding ground truth, rating them on a scale from 1 to 5 based on their similarity and overall quality.

3.2 Implementation and Training Protocols

Due to significant imaging differences across scanner platforms, we normalized each image in the LaSVoM dataset to a mean of 0.5 and a standard deviation of 0.5. For both the training and test datasets, we applied a noise reduction algorithm[2] to remove background noise from the audio, followed by audio magnitude normalization. Since the dataset contained a substantial number of silent segments (audio volume below -60 dB for more than 0.2 seconds), we used FFmpeg[3] to detect and segment audio sequences, retaining only the voiced segments. Finally, we resampled the audio to 16,000 Hz and the video to 80 fps.

[2] `https://github.com/timsainb/noisereduce`
[3] `https://github.com/FFmpeg/FFmpeg`

Table 1: Quantitative evaluation results of the audio quality using different synthesis methods on various metrics

Method	#Param	Corr2D ↑	PESQ ↑	MOS ↑
Vanilla (CNN) *w/* KE	124.0 M	0.672	1.135	3.48 (±0.14)
Vanilla (CNN)	124.0 M	0.641	1.116	3.16 (±0.09)
Vanilla (Transformer) *w/* KE	146.5 M	0.688	1.140	3.57 (±0.15)
Vanilla (Transformer)	146.5 M	0.649	1.131	3.22 (±0.12)
Ours	121.1 M	**0.818**	**1.251**	**4.13 (±0.08)**
Ours *w/o* adversarial training	60.0 M	0.759	1.190	3.80 (±0.13)
Ours *w/o* Flow	120.6 M	0.798	1.237	4.02 (±0.10)

Table 2: Ablation study in our knowledge enhancement strategy

Method	Corr2D ↑	PESQ ↑	MOS ↑
Baseline	0.716	1.152	3.74 (± 0.12)
$+\mathcal{L}_{con}$	0.758	1.170	3.88 (± 0.14)
$+\mathcal{L}_{con} + \mathcal{L}_{rec}$	0.802	1.226	4.06 (± 0.10)
$+\mathcal{L}_{con} + \mathcal{L}_{rec} + \mathcal{L}_{Koleo}$	0.818	1.251	4.13 (± 0.08)

In the proposed model, The MRI visual extractor was composed of 12 ViTs blocks. The dimension of the WaveNet was 512. The decoder and discriminator version of the HiFi-GAN we used was V1. The networks were trained using the AdamW optimizer [20] with $\beta_1 = 0.8$, $\beta_2 = 0.99$ and weight decay $\lambda = 0.01$. We implemented our approach using the Pytorch toolbox and trained for a total of 48 hours on an NVIDIA Tesla A100 GPU, where 40 epochs were used for the knowledge enhancement stage and 100 epochs were used for the variational inference stage.

3.3 Benchmarking

Quantitative Analysis. To validate the effectiveness of the proposed KE-CVAE framework, we compared it against two vanilla network variants following [1,26]. Specifically, we replaced the variational inference step with conventional CNN and transformer architectures with a similar number of parameters for the same speech audio generation task. In these vanilla variants, we directly used the MRI sequences as input and passed them through a series of encoders to obtain the audio output. Both vanilla models were optimized using the reconstruction loss $\mathcal{L}_{recon}$ as defined in our method. From Table 1, it is prominent that the performances of both architectures are evidently lower than that of KE-CVAE. The p-value obtained from the one-tailed Student's t-test is less than 0.05, indicating statistical significance. Notably, our knowledge enhancement step (KE in Table 1) also proves to be beneficial for the two vanilla methods, demonstrating its general applicability and effectiveness.

We further conducted an ablation study to assess the contributions of different components in the KE-CVAE framework. Results in Table 1 indicate that

removing adversarial training and Flow mechanisms leads to a noticeable performance drop, highlighting the critical role of either component in achieving optimal performance. Meanwhile, every component in Section 2.1 plays an important role in the whole framework. To evaluate the impact of each objective function, we performed additional ablation studies and presented the results in Table 2, where we used the CVAE model without knowledge enhancement as baseline. Results show that all the proposed loss functions in Section 2.1 are essential in improving the final performance.

Qualitative Analysis. Moreover, we visualized the spectrograms of the generated audio clips from the test stage, with one example shown in Figure 2. As can be seen, the generated spectrogram and its corresponding audio waveform from KE-CVAE exhibit superior alignment with the ground truth. In contrast, the CVAE framework without the knowledge enhancement step (*w/o* KE in Figure 2) and the vanilla transformer model's quality are visually lower than that of the proposed full KE-CVAE.

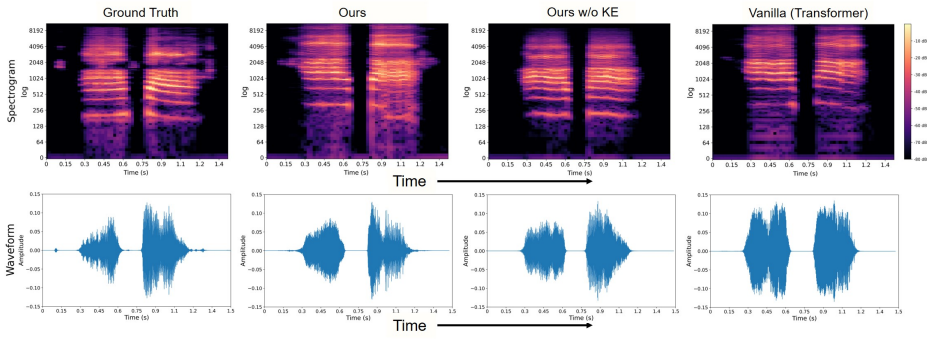

Fig. 2: An example of generated audio spectrograms and waveforms using different methods compared to the ground truth.

4 Conclusion

In this work, we developed a novel two-step "knowledge enhancement + variational inference" framework to synthesize high-quality speech waveforms from dynamic MRI sequences, addressing critical challenges in synchronizing speech audio with real-time MRI recordings. In the knowledge enhancement phase, we proposed a robust self-supervised training pipeline that leverages large-scale MRI data to learn domain-specific features without explicit annotations. Following this, we integrated a variational inference framework to enhance the model's expressiveness. Specifically, we employed a posterior encoder, a prior encoder, and a decoder to effectively map latent variables to speech waveforms, further augmenting them with normalizing flow and adversarial training to improve the

overall training process. Experimental results have demonstrated the efficacy of KE-CVAE in generating high-quality, temporally accurate speech from dynamic MRI data. This method has the potential to significantly enhance the accuracy of speech analysis in both clinical and research settings.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Akbari, H., Arora, H., Cao, L., Mesgarani, N.: Lip2audspec: Speech reconstruction from silent lip movements video. In: 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP). pp. 2516–2520. IEEE (2018)
2. Bresch, E., Nielsen, J., Nayak, K., Narayanan, S.: Synchronized and noise-robust audio recordings during realtime magnetic resonance imaging scans. The Journal of the Acoustical Society of America **120**(4), 1791–1794 (2006)
3. Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A.: Unsupervised learning of visual features by contrasting cluster assignments. Advances in neural information processing systems **33**, 9912–9924 (2020)
4. Chen, X., He, K.: Exploring simple siamese representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 15750–15758 (2021)
5. Chi, T., Ru, P., Shamma, S.A.: Multiresolution spectrotemporal analysis of complex sounds. The Journal of the Acoustical Society of America **118**(2), 887–906 (2005)
6. Delattre, S., Fournier, N.: On the kozachenko–leonenko entropy estimator. Journal of Statistical Planning and Inference **185**, 69–93 (2017)
7. Ertürk, M.A., Bottomley, P.A., El-Sharkawy, A.M.M.: Denoising mri using spectral subtraction. IEEE Transactions on biomedical engineering **60**(6), 1556–1562 (2013)
8. Jacobs, S.M., Versteeg, E., van der Kolk, A.G., Visser, L.N., Oliveira, Í.A., van Maren, E., Klomp, D.W., Siero, J.C.: Image quality and subject experience of quiet t1-weighted 7-t brain imaging using a silent gradient coil. European radiology experimental **6**(1), 36 (2022)
9. Jin, R., Li, Y., Shosted, R.K., Xing, F., Gilbert, I., Perry, J.L., Woo, J., Liang, Z.P., Sutton, B.P.: Optimization of 3d dynamic speech mri: Poisson-disc undersampling and locally higher-rank reconstruction through partial separability model with regional optimized temporal basis. Magnetic Resonance in Medicine **91**(1), 61–74 (2024)
10. Jin, R., Shosted, R.K., Xing, F., Gilbert, I.R., Perry, J.L., Woo, J., Liang, Z.P., Sutton, B.P.: Enhancing linguistic research through 2-mm isotropic 3d dynamic speech mri optimized by sparse temporal sampling and low-rank reconstruction. Magnetic Resonance in Medicine **89**(2), 652–664 (2023)
11. Kim, J., Kim, S., Kong, J., Yoon, S.: Glow-tts: A generative flow for text-to-speech via monotonic alignment search. Advances in Neural Information Processing Systems **33**, 8067–8077 (2020)
12. Kim, J., Kong, J., Son, J.: Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In: International Conference on Machine Learning. pp. 5530–5540. PMLR (2021)

13. Kong, J., Kim, J., Bae, J.: Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. Advances in neural information processing systems **33**, 17022–17033 (2020)
14. Larsen, A.B.L., Sønderby, S.K., Larochelle, H., Winther, O.: Autoencoding beyond pixels using a learned similarity metric. In: International conference on machine learning. pp. 1558–1566. PMLR (2016)
15. Lee, S.H., Kim, S.B., Lee, J.H., Song, E., Hwang, M.J., Lee, S.W.: Hierspeech: Bridging the gap between text and speech by hierarchical variational inference using self-supervised representations for speech synthesis. Advances in Neural Information Processing Systems **35**, 16624–16636 (2022)
16. Lim, Y., Toutios, A., Bliesener, Y., Tian, Y., Lingala, S.G., Vaz, C., Sorensen, T., Oh, M., Harper, S., Chen, W., et al.: A multispeaker dataset of raw and reconstructed speech production real-time mri video and 3d volumetric images. Scientific data **8**(1), 187 (2021)
17. Lim, Y., Zhu, Y., Lingala, S.G., Byrd, D., Narayanan, S., Nayak, K.S.: 3d dynamic mri of the vocal tract during natural speech. Magnetic resonance in medicine **81**(3), 1511–1520 (2019)
18. Liu, X., Xing, F., Stone, M., Zhuo, J., Fels, S., Prince, J.L., El Fakhri, G., Woo, J.: Speech audio synthesis from tagged mri and non-negative matrix factorization via plastic transformer. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 435–445. Springer (2023)
19. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10012–10022 (2021)
20. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
21. Lulich, S.M., Berkson, K.H., de Jong, K.: Acquiring and visualizing 3d/4d ultrasound recordings of tongue motion. Journal of phonetics **71**, 410–424 (2018)
22. Mao, X., Li, Q., Xie, H., Lau, R.Y., Wang, Z., Paul Smolley, S.: Least squares generative adversarial networks. In: Proceedings of the IEEE international conference on computer vision. pp. 2794–2802 (2017)
23. Menchón-Lara, R.M., Simmross-Wattenberg, F., Casaseca-de-la Higuera, P., Martín-Fernández, M., Alberola-López, C.: Reconstruction techniques for cardiac cine mri. Insights into imaging **10**, 1–16 (2019)
24. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
25. Perkell, J.S., Cohen, M.H., Svirsky, M.A., Matthies, M.L., Garabieta, I., Jackson, M.T.: Electromagnetic midsagittal articulometer systems for transducing speech articulatory movements. The Journal of the Acoustical Society of America **92**(6), 3078–3096 (1992)
26. Prajwal, K., Mukhopadhyay, R., Namboodiri, V.P., Jawahar, C.: Learning individual speaking styles for accurate lip to speech synthesis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 13796–13805 (2020)
27. 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)
28. Rezende, D., Mohamed, S.: Variational inference with normalizing flows. In: International conference on machine learning. pp. 1530–1538. PMLR (2015)

29. Toutios, A., Lingala, S.G., Vaz, C., Kim, J., Esling, J.H., Keating, P.A., Gordon, M., Byrd, D., Goldstein, L., Nayak, K.S., et al.: Illustrating the production of the international phonetic alphabet sounds using fast real-time magnetic resonance imaging. In: Interspeech. pp. 2428–2432 (2016)
30. Zheng, R.C., Ai, Y., Ling, Z.H.: Incorporating ultrasound tongue images for audio-visual speech enhancement. IEEE/ACM Transactions on Audio, Speech, and Language Processing (2024)