CMRI2SPEC: CINE MRI SEQUENCE TO SPECTROGRAM SYNTHESIS VIA A PAIRWISE HETEROGENEOUS TRANSLATOR

Xiaofeng Liu[1], *Fangxu Xing*[1], *Maureen Stone*[3], *Jerry L. Prince*[2] *Fellow, IEEE, Jangwon Kim*[4], *Georges El Fakhri*[1], *Fellow, IEEE, Jonghye Woo*[1], *Senior Member, IEEE,*

[1]Dept. of Radiology, Massachusetts General Hospital and Harvard Medical School, Boston, MA, USA
[2]Dept. of Electrical and Computer Engineering, Johns Hopkins University, Baltimore, MD, USA
[3]Dept. of Neural and Pain Sciences, University of Maryland School of Dentistry, Baltimore, MD, USA
[4]Amazon.com, USA

ABSTRACT

Multimodal representation learning using visual movements from cine magnetic resonance imaging (MRI) and their acoustics has shown great potential to learn shared representation and to predict one modality from another. Here, we propose a new synthesis framework to translate from cine MRI sequences to spectrograms with a limited dataset size. Our framework hinges on a novel fully convolutional heterogeneous translator, with a 3D CNN encoder for efficient sequence encoding and a 2D transpose convolution decoder. In addition, a pairwise correlation of the samples with the same speech word is utilized with a latent space representation disentanglement scheme. Furthermore, an adversarial training approach with generative adversarial networks is incorporated to provide enhanced realism on our generated spectrograms. Our experimental results, carried out with a total of 63 cine MRI sequences alongside speech acoustics, show that our framework improves synthesis accuracy, compared with competing methods. Our framework thereby has shown the potential to aid in better understanding the relationship between the two modalities.

Index Terms— Video to Spectrogram Synthesis, Magnetic Resonance Imaging, Encoder and Decoder, GAN.

1. INTRODUCTION

Multimodal representation learning and translation from visual movements shown in cine magnetic resonance imaging (MRI) to intelligible speech is an important yet challenging problem, due to their highly disparate features inherent in the data structure [1, 2]. As shown in Fig. 1, a two-dimensional plus time (2D+t) MRI sequence has high spatial resolution with low frame rate, while a two-dimensional (2D) mel spectrogram converted from its one-dimensional (1D) audio waveform represents the short-term power spectrum of the audio signal. Since cross-modality speech models often lose pitch information [2, 3], recent studies make use of spectrograms as an intermediate means of covering both pitch and resonance

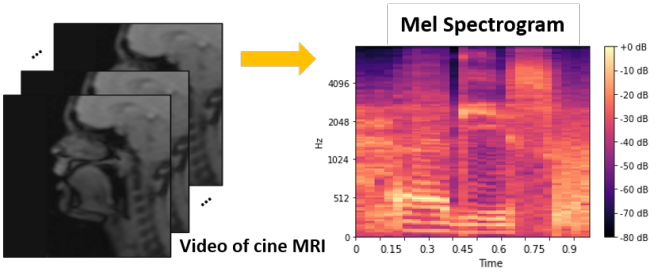

Fig. 1. Illustration of a cine MRI sequence and its spectrogram.

information of audio signals [2, 3, 4, 5]. In this work, we are thus interested in developing a method to carry out the multimodal translation from MRI sequences to spectrograms.

An early attempt in lip reading proposed in [3, 6] used convolutional neural networks (CNN) to extract face visual features and modeled audio features with linear predictive coding analysis and line spectrum pairs decomposition; however, that approach lost the fundamental frequency and periodicity. To bypass this issue, a new approach, termed Lip2AudSpec [2], was proposed, in which an end-to-end lip reading network with CNN for spatial modeling, recurrent neural networks (RNN) for temporal modeling, and fully connected layers was presented. However, RNN, in general, is difficult to train [7, 8] and is likely to yield a suboptimal solution, when limited training data are used [9]. In addition, the use of the fully connected layers in Lip2AudSpec unavoidably loses spatial and temporal correlations [10], and incur a large number of to-be-trained parameters [11].

To alleviate the concerns above, in this work, we design an efficient end-to-end network, aimed at the cine MRI sequence-to-spectrogram synthesis. In particular, we resort to a fully convolutional translator with a 3D CNN encoder and a 2D transpose decoder. Notably, we do not rely on the RNN nor fully connected layers, and therefore our network has 10× fewer parameters than Lip2AudSpec [2].

In addition, we impose an additional optimization constraint, following a prior knowledge that feature representations of the same word content are similar from one subject to another. Accordingly, following this observation, we split the latent space of translator into the word content (i.e., "ageese" or "asouk" in this work) and subject-specific (i.e., style of an articulation) parts with a tensor slice operation. For the pairs with the same word content, we explicitly enforce the similarity of their word content part w.r.t. the Kullback-Leibler (KL) divergence. Then, the decoder takes both the word content and the style of the articulation information for the spectrogram synthesis. Furthermore, an adversarial training scheme as in generative adversarial networks (GAN) [12] is incorporated into our framework to yield enhanced realism on the synthesized spectrograms. Taken together, the proposed approach can potentially benefit clinicians and researchers in understanding the relationship between the two disparate modalities and in improving treatment strategies for patients with speech-related disorders.

2. METHODOLOGY

Given the paired cine MRI sequence x and its quasi-synchronous audio spectrogram y, we propose learning a parameterized mapping $f : x \rightarrow \tilde{y}$ from the cine MRI sequence x to the generated spectrogram $\tilde{y}$ so as to closely resemble the real spectrogram y. Toward this goal, our framework is based on a pairwise disentangled fully convolutional heterogeneous translator, and a generative adversarial network (GAN) loss is simply added on to further boost the performance. In addition to a vanilla encoder-decoder, we further propose to exploit the similarity of the same word content in a latent space and to enforce an additional constraint.

Specifically, our backbone model is based on a heterogeneous translator, which takes a MRI sequence as input and outputs a spectrogram. Since we have the fixed number of MR images for each subject, i.e., 26 frames, we propose to adopt the 3D CNN [13] for fast encoding. The output of a 3D convolutional layer has the same dimension in the temporal direction, and the dimension is halved by each 3D MaxPooling operation. We use five 3D convolutional layers to generate a feature representation, followed by a reshape operation and a decoder network with four deconvolutional layers. The detailed network structure is provided in Table 1. We use the rectified linear unit (ReLU) as the non-linear activation function for all intermediate layers, and use the sigmoid function to normalize the final output. The reconstruction loss takes the most important place in an encoder-decoder translator-based framework. We adopt the mean square-error (MSE) loss, which can be formulated as:

$$\mathcal{L}_{MSE} = \frac{1}{2}\sum_{n=1}^{N} ||y_n - \tilde{y}_n||_2^2, \quad (1)$$

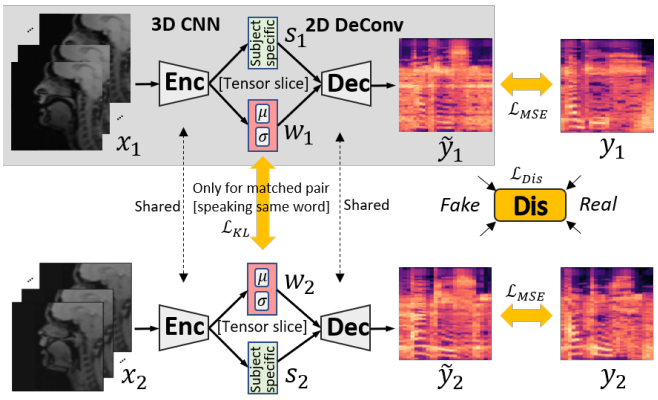

Fig. 2. Illustration of our proposed framework. Note that only the gray masked part is used in testing.

where $N = 64 \times 64$, which indicates the number of pixels in y.

To impose an additional constraint with the prior of the word content similarity in the latent space, we first propose disentangling the word content w from subject-specific factors s. Similar to deep metric learning [14], we require that the samples with the same speech word have a similar word content representation. Specifically, we take two samples x_1 and x_2 as our input and expect to extract their word content features w_1 and w_2, as well as the subject-specific factors s_1 and s_2. In order to separate those two components, a typical solution would be to use multiple branches, but it is inefficient in a fully convolutional network. Instead, we opt to differentiate the specific channels in the feature representation with a tensor slice operation [10][1].

We explicitly enforce that w has information about the word content, since the word content is the shared part for the pairs across subjects, when speaking the same word [15]. In parallel, we rely on an implicit complementary constraint to achieve that s has subject-specific information [16, 17]. By enforcing an information bottleneck, i.e., compact latent feature [17], s should incorporate all of the complementary information (e.g., subject-specific style of an articulation) other than w to achieve accurate spectrogram reconstruction. The decoder can also be regarded as taking s conditioned on w for generation, which models a word-conditioned distribution of the spectrogram in a divide-and-conquer manner and thereby is easier than vanilla encoder-decoder modeling [18, 19].

To efficiently measure the distribution-wise discrepancy of the word content feature, we adopt the KL divergence with the reparameterization trick. In practice, we adopt the Gaussian prior and select a few channels in the feature representation to denote the mean μ and variance σ. We then utilize the reparametrization trick $w = \mu + \sigma \odot \epsilon$, where $\epsilon \in \mathcal{N}(0, I)$ [20]. The word content feature w_1, therefore, can be repre-

[1] Link: Slicing the tensor in PyTorch.

Table 1. Structure of the proposed networks.

Encoder (3D CNN)		Decoder	
Layers	Size	Layers	Size
Input	(1, 128, 128, 26)	Reshape	(128, 4, 4)
Conv3D (32) ReLU MaxPooling3D (2,2,2)	(32, 64, 64, 13)	Conv2DTrans(96) ReLU	(96, 8, 8)
Conv3D (32) ReLU MaxPooling3D (2,2,2)	(32, 32 32, 7)	Conv2DTrans(24) ReLU	(24, 16, 16)
Conv3D(64) ReLU MaxPooling3D (2,2,2)	(64, 16 16, 4)	Conv2DTrans(4) ReLU	(4, 32, 32)
Conv3D (64) ReLU MaxPooling3D (2,2,2)		Conv2DTrans(1) sigmoid	(1, 64, 64)
Conv3D (128) ReLU MaxPooling3D (2,2,2)	(128, 4, 4, 1) →	100:s/14:c_μ/14:c_σ	

sented by μ_1 and σ_1. We note that μ_1 and σ_1 should have the same size. Specifically, we split the encoded feature with 128 channels into three parts, i.e., 14 channels for the mean of the word content, 14 channels for the variance of the word content, and the remaining 100 channels for the subject-specific factors. The detailed KL divergence between w_1 and w_2 can be formulated as:

$$\mathcal{L}_{KL} = -\frac{1}{2}\sum_{m=1}^{M}[1 + \log\frac{\sigma_{1m}^2}{\sigma_{2m}^2} - \frac{\sigma_{1m}^2}{\sigma_{2m}^2} - \frac{(\mu_{1m} - \mu_{2m})^2}{\sigma_{2m}^2}], \tag{2}$$

where M indicates the number of channels of the mean or variance (14 in our implementation.) We note that $\mathcal{L}_{KL}$ is only applied to the sample pair with the same speech word.

To further enrich realism on our generated spectrograms, we added the GAN module. The discriminator takes both the generated spectrogram and the real spectrogram as input and detects which is generated or real. To this end, we use three convolutional layers and two fully connected layers with a sigmoid output as our discriminator. The binary cross-entropy loss of the discriminator can be formulated as:

$$\mathcal{L}_{Dis} = \mathbb{E}_y\{\log(Dis(y))\} + \mathbb{E}_{\tilde{y}}\{\log(1 - Dis(\tilde{y}))\}. \tag{3}$$

In an adversarial game [20], the translator should try to confuse the discriminator, by outputting realistic spectrograms. Therefore, the translator parts are trained by minimizing the following objective:

$$\mathcal{L}_T = \mathbb{E}_{\tilde{y}}\{-\log(1 - Dis(\tilde{y}))\}. \tag{4}$$

We note that the translator parts do not involve the real image in $\log(Dis(y))$ [12].

In summary, we jointly optimize the following objectives for the encoder, decoder, and discriminator:

$$\min_{Enc,Dec} \mathcal{L}_{MSE} + \beta\mathcal{L}_{KL} + \lambda\mathcal{L}_T, \tag{5}$$

$$\min_{Dis} \quad \mathcal{L}_{Dis}, \tag{6}$$

where β and λ are the weighting parameters. We note that $\beta = 0$ for the pairs with different speech words.

At the testing stage, we only use the translator part to make inference as shown in Fig. 1, and do not need the pairwise inputs and the discriminator. Therefore, the inference speed in implementation is not affected by the pairwise framework and additional adversarial training.

3. EXPERIMENTS AND RESULTS

3.1. Data Acquisition

For the experiments carried out in this study, cine MR images were acquired with acoustics, while speaking "asouk," and "ageese" following a periodic metronome-like sound. The data were acquired on a Siemens 3.0T TIM Trio system with a 12-channel head coil and a 4-channel neck coil using a segmented gradient echo sequence [21, 22]. The imaging parameters are as follows: the field of view was 240×240 mm with a resolution of 1.87mm×1.87mm×6mm. The MRI sequence was acquired at the rate of 26 fps.

3.2. Experimental Setup

In our experiments, we used a total of 63 MRI sequence and audio pairs, in which 43 subjects performed "ageese" and 20 subjects performed "asouk." Each paired dataset consists of 26 MRI slices and the corresponding audio waveform (the length varies from 21,832 to 24,175). For data augmentation and normalization, we used a sliding window for the audio waveform to crop 21,000 time points, generating 100× audio data. Then, the cropped audio waveforms were converted into their mel-spectrograms as a pre-processing[2]. We resized the MRI slices to the size of 128×128, and the corresponding mel-spectrogram to the size of 64×64. We used leave-one-out evaluation in a subject independent manner.

To control the weights of our optimization objectives, we set $\beta = 0.5$ for the pair with the same speech word, while $\beta = 0$ for the pair with different speech words. In addition, λ is not sensitive for a relatively large range, e.g., 0.2 to 0.7, and we simply set $\beta = 0.5$ consistently.

We implemented our framework using the PyTorch toolbox and used the librosa library for audio processing. The training was implemented in a server with an NVIDIA V100 GPU, which took about 3 hours. In testing, the inference took only 0.5s. The learning rate was set at $lr_{Enc,Dec} = 1\text{e}{-3}$ and $lr_{Dis} = 1\text{e}{-4}$ and the momentum was set at 0.5.

3.3. Qualitative Evaluations

The synthesis results using our proposed framework and Lip2AudSpect [2] are shown in Fig. 3. We can clearly observe that the result of our framework is more consistent with the ground truth than the competing method and ablation

[2]Link: The librosa for audio to mel-spectrogram.

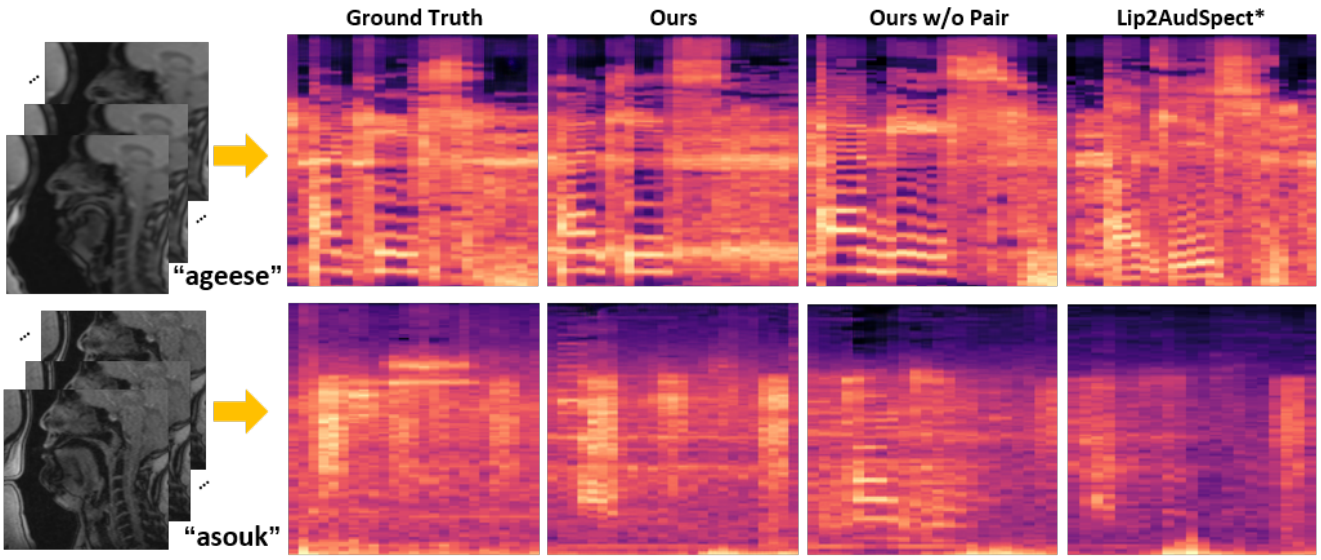

Fig. 3. Comparison of different MRI sequence-to-spectrogram generation approaches, including Lip2AudSpect and a series of our ablation studies. *Note that we reimplemented Lip2AudSpect suitable for MRI data.

Table 2. Numerical comparisons in testing with leave-one-out evaluation. The best and the second best results are **bold** and underlined, respectively.

Methods	L1 ↓
Lip2AudSpect [2]	0.315±0.004
Ours w/o PairDisen&GAN	0.273±0.007
Ours w/o PairDisen	0.260±0.005
Ours w/o GAN	0.252±0.005
Ours	**0.246**±0.006

studies. We note that Lip2AudSpect used a single translator structure and thus did not consider the pairwise constraint for disentangled representation learning. In addition, the 3D CNN of Lip2AudSpect with the temporal window size of 5 only extracted local temporal correlation, thus relying on the RNN for long-term exploration; however, Lip2AudSpect has a difficulty in training the network with a limited number of datasets. To further demonstrate the effectiveness of our pairwise disentanglement, we showed that the result of using a single translator alongside the GAN model without pairwise interaction.

3.4. Quantitative Evaluations

Table 2 lists numerical comparisons between the proposed framework, its ablation studies, and Lip2AudSpect [2]. The standard deviation is reported by three random trials. Actually, our proposed framework without the pairwise disentanglement learning and GAN adopted a similar translator network as in Lip2AudSpect [2], but the network backbone is different. Our heterogeneous fully convolutional translator does not rely on the RNN nor fully connected layers, which thus has much fewer parameters and is easy to train. In the ablation studies, we can clearly see that both the pairwise learning and GAN loss contribute to the superior performance.

4. CONCLUSION

In this work, we proposed a novel synthesis framework to generate mel-spectrograms, given cine MRI sequences. Our proposed framework is based on a novel fully convolutional heterogeneous translator, comprising a 3D CNN encoder for efficient sequence encoding and a 2D transpose convolution decoder. A pairwise correlation of the samples with the same speech word is exploited with a latent space representation disentanglement scheme. Specifically, we adopted the tensor slice to facilitate the disentanglement in our fully convolutional framework, and enforced the same word content feature similarity using the KL divergence with the reparameterization trick. Furthermore, an adversarial training with GAN was incorporated to provide improved realism on our generated spectrograms. Our experimental results demonstrated that the proposed approach improved the synthesis accuracy by a large margin, compared with the competing methods. Our approach thereby has shown effective at training and deploying the framework with a limited data size, potentially helping clinicians improve treatment strategies for patients with speech-related disorders.

5. ACKNOWLEDGMENT

This work is supported by NIH R01DC018511.

6. REFERENCES

[1] Joon Son Chung and Andrew Zisserman, "Lip reading in the wild," in *ACCV*. Springer, 2016, pp. 87–103.

[2] Hassan Akbari, Himani Arora, Liangliang Cao, and Nima Mesgarani, "Lip2audspec: Speech reconstruction from silent lip movements video," in *ICASSP*. IEEE, 2018, pp. 2516–2520.

[3] Ariel Ephrat and Shmuel Peleg, "Vid2speech: speech reconstruction from silent video," in *ICASSP*. IEEE, 2017, pp. 5095–5099.

[4] Gewen He, Xiaofeng Liu, Fangfang Fan, and Jane You, "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*, 2020, pp. 912–913.

[5] Gewen He, Xiaofeng Liu, Fangfang Fan, and Jane You, "Classification-aware semi-supervised domain adaptation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 964–965.

[6] Daniel Michelsanti, Zheng-Hua Tan, Shi-Xiong Zhang, Yong Xu, Meng Yu, Dong Yu, and Jesper Jensen, "An overview of deep-learning-based audio-visual speech enhancement and separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2021.

[7] Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio, "On the difficulty of training recurrent neural networks," in *ICML*. PMLR, 2013, pp. 1310–1318.

[8] Wanqing Xie, Lizhong Liang, Yao Lu, Hui Luo, and Xiaofeng Liu, "Deep 3d-cnn for depression diagnosis with facial video recording of self-rating depression scale questionnaire," *JBHI*, 2021.

[9] Jing Wang, Xiaofeng Liu, Fangyun Wang, Lin Zheng, Fengqiao Gao, Hanwen Zhang, Xin Zhang, Wanqing Xie, and Binbin Wang, "Automated interpretation of congenital heart disease from multi-view echocardiograms," *Medical Image Analysis*, vol. 69, pp. 101942, 2021.

[10] Xiaofeng Liu, Tong Che, Yiqun Lu, Chao Yang, Site Li, and Jane You, "Auto3d: Novel view synthesis through unsupervisely learned variational viewpoint and global 3d representation," in *ECCV*. Springer, 2020, pp. 52–71.

[11] Ian Goodfellow, Yoshua Bengio, and Aaron Courville, "Deep learning (adaptive computation and machine learning series)," .

[12] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen, "Improved techniques for training gans," *NIPS*, vol. 29, pp. 2234–2242, 2016.

[13] Xiaofeng Liu, Zhenhua Guo, Jane You, and BVK Vijaya Kumar, "Dependency-aware attention control for image set-based face recognition," *IEEE TIFS*, vol. 15, pp. 1501–1512, 2019.

[14] Xiaofeng Liu, BVK Vijaya Kumar, Jane You, and Ping Jia, "Adaptive deep metric learning for identity-aware facial expression recognition," in *CVPR*, 2017, pp. 20–29.

[15] Xiaofeng Liu, Fangxu Xing, Georges El Fakhri, and Jonghye Woo, "A unified conditional disentanglement framework for multimodal brain mr image translation," in *ISBI*. IEEE, 2021, pp. 10–14.

[16] Xiaofeng Liu, Site Li, Lingsheng Kong, Wanqing Xie, Ping Jia, Jane You, and BVK Kumar, "Feature-level frankenstein: Eliminating variations for discriminative recognition," in *CVPR*, 2019, pp. 637–646.

[17] Xiaofeng Liu, Yang Chao, Jane J You, C-C Jay Kuo, and Bhagavatula Vijayakumar, "Mutual information regularized feature-level frankenstein for discriminative recognition," *IEEE TPAMI*, 2021.

[18] Tong Che, Xiaofeng Liu, Site Li, Yubin Ge, Ruixiang Zhang, Caiming Xiong, and Yoshua Bengio, "Deep verifier networks: Verification of deep discriminative models with deep generative models," *AAAI*, 2021.

[19] Xiaofeng Liu, Bo Hu, Linghao Jin, Xu Han, Fangxu Xing, Jinsong Ouyang, Jun Lu, Georges EL Fakhri, and Jonghye Woo, "Domain generalization under conditional and label shifts via variational bayesian inference," *IJCAI*, 2021.

[20] Xiaofeng Liu, Fangxu Xing, Jerry L Prince, Aaron Carass, Maureen Stone, Georges El Fakhri, and Jonghye Woo, "Dual-cycle constrained bijective vae-gan for tagged-to-cine magnetic resonance image synthesis," in *ISBI*. IEEE, 2021, pp. 1448–1452.

[21] Junghoon Lee, Jonghye Woo, Fangxu Xing, Emi Z Murano, Maureen Stone, and Jerry L Prince, "Semi-automatic segmentation of the tongue for 3D motion analysis with dynamic MRI," in *ISBI*. IEEE, 2013, pp. 1465–1468.

[22] Fangxu Xing, Jonghye Woo, Emi Z Murano, Junghoon Lee, Maureen Stone, and Jerry L Prince, "3D tongue motion from tagged and cine MR images," in *MICCAI*. Springer, 2013, pp. 41–48.