Interspeech 2024
1-5 September 2024, Kos, Greece

Optical Flow Guided Tongue Trajectory Generation for Diffusion-based Acoustic to Articulatory Inversion

Yudong Yang[1], *Rongfeng Su*[1,2,*], *Rukiye Ruzi*[1], *Manwa Ng*[3], *Shaofeng Zhao*[4], *Nan Yan*[1,2], *Lan Wang*[1,2,*]

[1]Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, China
[2]Guangdong-Hong Kong-Macao Joint Laboratory of Human-Machine Intelligence-Synergy Systems, China
[3]Division of Speech and Hearing Sciences, University of Hong Kong, China
[4]Department of Rehabilitation Medicine, The Eighth Affiliated Hospital of Sun Yat-sen University, China

{yd.yang2,rf.su,rkym.rouzi,nan.yan,lan.wang}@siat.ac.cn, manwa@hku.hk, zhaosf1@163.com

Abstract

The diffusion-based Acoustic-to-Articulatory Inversion (AAI) approach has demonstrated impressive results in converting audio into Ultrasound Tongue Imaging (UTI) data with clear tongue contours. However, Mean Square Error (MSE) based diffusion models focus only on the pixel error between reference and generated UTI data, inherently omitting changes in tongue movements. This leads to the discrepancy in tongue trajectory between reference and generated UTI data. To address this issue, this paper presents an Optical Flow Guided tongue trajectory generation method for training the diffusion-based AAI model. The optical flow method calculates the displacement information of the tongue contours in consecutive frames, enabling the tongue trajectory similarity between reference and generated UTI data to be used as an additional constraint for Diffusion Model network optimization. Experimental results show that our proposed diffusion-based AAI system with additional tongue trajectory constraint outperformed the baseline system across various evaluation metrics.

Index Terms: Diffusion Model, Optical Flow, Ultrasound Tongue Image, Tongue Trajectory, Ultrasound Generation

1. Introduction

Acoustic-to-articulatory inversion (AAI) involves converting speech signals into articulatory movement representation, such as Ultrasound Tongue Imaging (UTI) data. Researchers use acoustic signals to reconstruct the motion trajectory of the tongue surface, which is crucial for exploring the articulator function during speech production [1, 2, 3]. The tongue contour, embodying the shape and movements of the tongue, is vital for UTI data analysis and interpretation [4, 5, 6].

To achieve high-definition tongue contours, various approaches have been proposed, including Gaussian Mixture Model (GMM) based and Deep Neural Network (DNN) based AAI models [7, 8]. However, the DNN-based AAI systems often generate UTI sequences with blurry tongue contours [8], making it challenging to obtain distinct tongue motion trajectories from a discriminative model perspective. In contrast, generative-based methods can learn the joint probability distribution of speech and articulatory motions, potentially generating results with better temporal coherence [9, 10]. For example, Yang et al. successfully applied diffusion models to AAI tasks [11]. They developed a diffusion-based AAI model that incorporates both acoustic information about individual tongue movements from the original speech signals and general information related to the universality of tongue motions from Automatic Speech Recognition (ASR) transcriptions in the textual space. This model is optimized using Mean Square Error (MSE) as the loss function and generates high-quality UTI data with explicit tongue contours.

The mechanism of speech production dictates that different tongue movement trajectories should correspond to different speech outputs [4, 5, 6]. Therefore, to establish a reliable mapping from speech to UTI data, AAI models must accurately reproduce these tongue movement trajectories [12, 13]. However, MSE based AAI models minimize the pixel error between reference and output images, inherently neglecting changes in tongue movements.This oversight leads to the discrepancies in tongue trajectories between real and generated UTI data. Incorporating knowledge of tongue trajectories into the modeling process and designing practical metrics to guide model optimization are potential solutions for accurately estimating tongue contour.

The motion of objects in an image or a video can be effectively tracked with optical flow estimation [14]. It captures the changes in pixel values over time and analyzes the correlation between adjacent frames to determine the correspondence between consecutive frames. Optical flow estimation methods include sparse and dense optical flow. Dense optical flow, which captures the motion of every pixel's in an image sequence, is designed for processing high-quality, high-contrast data [15, 16]. However, for tongue ultrasound data that only requires the tongue contour, dense optical flow may introduce unnecessary details. In contrast, sparse optical flow can be calculated specifically for the region of interest, which is the tongue contour [17]. Consequently, this approach minimizes interference from irrelevant input and reveals essential tongue motion information. Therefore, this paper proposes a sparse optical flow-guided tongue trajectory estimation method to optimize the diffusion-based AAI model. This method utilizes sparse optical flow to calculate the displacement information of the tongue contours between consecutive frames and to measure the similarity between real and generated tongue trajectories. This similarity serves as an additional constraint for optimizing the diffusion model network. Experimental results show that the proposed diffusion-based AAI system outperforms baseline systems across various evaluation metrics. Notably, in subjective assessments, most participants found the tongue trajectories

*Corresponding Authors

10.21437/Interspeech.2024-1864

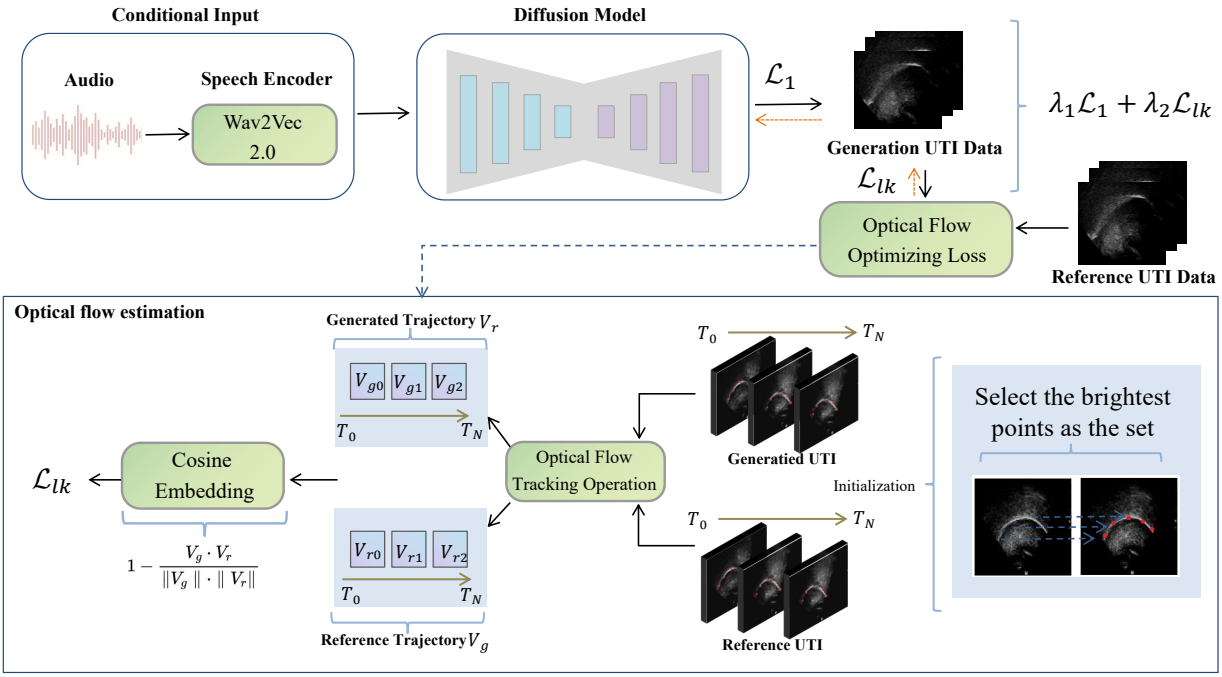

Figure 1: *The framework of the method. where $\mathcal{L}_{lk}$ represents the trajectory similarity loss, $\mathcal{L}_1$ represents the loss of the diffusion model, and V_g and V_r represent the vector field representation of the trajectory, respectively.*

of the generated UTI data to be more similar to the real ones.

2. Methods

In this section, we explore the use of Optical Flow Guided tongue trajectory estimation to optimize the diffusion model. As shown in Figure 1, the diffusion model takes audio-only features extracted from a speech encoder as conditional input to generate UTI data [11, 18]. Meanwhile, the core area of the tongue is tracked with optical flow as illustrated in Figure 2. This tracking is carried out for both generated and reference UTI data. The similarity between the two trajectories is incorporated into the diffusion model as a part of loss function to enhance the UTI generation process.

2.1. Diffusion Model for Ultrasound Tongue Imaging

The distribution of training ultrasound data samples is denoted as $q(x)$, and the distribution of generated data samples is denoted as $q(x')$. We aim for the distribution of the generated samples to closely resemble that of the training data, thereby achieving more authentic restoration results.

The diffusion model consists of a forward process and a reverse process. For the forward process, x represents tongue ultrasound data, while $q(x)$, σ_{utl} represent its distribution and corresponding standard deviation, respectively. Gaussian noise with a standard deviation of σ is added to the ultrasound data. When $\sigma_{max} \gg \sigma_{utl}$, distribution of $p(x;\sigma)$ can be obtained, where $p(x;\sigma_{max})$ approximates Gaussian noise.

The key procedure for the diffusion model is to randomly sample a starting point $x_0 \sim \mathcal{N}(0, \sigma_{\max}^2 I)$ that contains noise and then denoise it through progressively decreasing noise levels from $\sigma_{max} = \sigma_0$ to $\sigma_1 \cdots$ to $\sigma_N = 0$ as in ultrasound tongue data x_i, for $x_i \sim p(x;\sigma_i)$. Ultimately, the distribution of generated ultrasound data x_N closely approximate the real data distribution $q(x)$.

To reverse the forward process of the diffusion model, transitioning from a random state back to the real data distribution, we introduce a denoising function $D(x,\sigma)$ to minimize the $L2$ denoising error, where y represents a training data point and n represents noise. The loss function is defined as follows:

$$\mathcal{L}_1 = \mathbb{E}_{y\sim q}\mathbb{E}_{n\sim\mathcal{N}(0,\sigma_{\max}^2 I)}||D(y+n,\sigma) - y||_2^2 \quad (1)$$

2.2. Optical Flow Estimation for Tongue Trajectory

2.2.1. *initialization points selection*

The tongue movement trajectories during different articulation are crucial for generating high-quality UTI data. We use optical flow estimation to capture the tongue movement trajectories in both real and generated data, and then utilize the similarity between those two sets of trajectories to optimize the diffusion model. Specifically, we select high-brightness points in the UTI as initialization points, denoted as set I, for the optical flow estimation, since the tongue reflects higher brightness in UTI.

$$I = \{(i,j) \in x_i | TopN(x_i)\} \quad (2)$$

Where $TopN(x_i)$ denotes the tongue brightness areas in the x_i UTI data, and (i,j) refers to coordinates.

2.2.2. *Optical flow estimation*

We select the Lucas-Kanade(LK) algorithm [19, 20] for the optical flow estimation. The primary objective is to estimate the motion in frames by minimizing the sum of squared differences between two consecutive frames. The optimization function is defined as follows:

$$E(p) = \sum_{I\in\Omega} \left(F_{t-1}(W(I;p)) - F_t(I)\right)^2 \quad (3)$$

where $E(p)$ denotes the total sum of squared differences between the previous frame F_{t-1} after transformation W and the current frame F_t under the parameters p. The set Ω refers to a group of patch positions centered around I.

Additionally, we estimate the optical flow vector $v = (dx, dy)$, which represents the displacement of feature between two frames. We minimize the error$E(p)$ by adjusting the parameters p. The parameter update follows an iterative strategy, where the gradient of the error function is calculated. The parameter update is given by:

$$\Delta p = -(J^T J)^{-1} J^T \bigtriangledown E \tag{4}$$

Here, Δp is update amount for parameter p, J is the Jacobian matrix of $E(P)$ with respect to parameter p, and E is the gradient of the error function. Parameters are updated iteratively ($p \leftarrow p + \Delta p$) until convergence.

During this process, the Lucas-Kanade algorithm ensures the continuity and accuracy of motion trajectories, maintaining coherence in the trajectory of each feature point and conforming to the laws of motion. This method captures the motion trajectories of the tongue in different articulation states, providing reliable tongue motion estimation for UTI reconstruction.

2.2.3. Trajectory Similarity loss

Cosine embedding can better reflect the degree of similarity between two trajectories. We adopt cosine embedding [21], immediately after optical flow estimation, to optimize the similarity between the generated tongue trajectories and the real ones.

Firstly, due to the innate individual differences in tongue structure, different speakers lead to variation in the intial points of optical flow in UTI. We align the generated trajectory vector V_g with the real trajectory vector V_r at their initial positions by calculating the difference. Then, we feed the aligned vectors V_r' and V_g' into the cosine embedding loss function, which allows us to compare the similarity between trajectories while maintaining consistency in their corresponding motion patterns. We define optimization function for this section as $\mathcal{L}_{lk}$:

$$\mathcal{L}_{lk} = \mathbb{E}_{y\sim q}\mathbb{E}_{n\sim\mathcal{N}\left(0,\sigma^2_{\max} I\right)}\left(1 - \frac{V_g' \cdot V_r'}{\parallel V_g' \parallel \cdot \parallel V_r' \parallel}\right) \tag{5}$$

Finally, the tongue motion trajectory similarity is used to optimize the diffusion-based AAI model, thereby improving the reconstruction quality of tongue movements. This methods further facilitates the understanding and analysis of the articulation mechanism. The overall loss function of the model is defined as follows:

$$\mathcal{L} = \lambda_1 \mathcal{L}_1 + \lambda_2 \mathcal{L}_{lk} \tag{6}$$

where λ_1 and λ_2 are adjustable weights for two optimization functions, respectively.

Algorithm 1 Description of the optical flow estimation

Input: F_{t-1}, F_t, $\mathbf{I}_{t-1}$, $\mathbf{p} = [0, 0]^T$
1.Select high brightness areas as initialization points $\mathbf{I}$
2.Calculate feature from F_{t-1} centered at $\mathbf{I}_{t-1}$
3.Calculate the gradient of the template feature
4.Pre-compute the Jacobian $\mathbf{J}$ and Hessian matrices $\mathbf{H}$
for $step = 1; step \leq max; step++$ **do**
    5. Extract target feature from F_t centered at $\mathbf{I}_{t-1} + \mathbf{p}$
    6. Compute the error of the template and target features
    7. Update the motion model: $p \leftarrow p + \Delta p$
end for
8.Obtain generated and real optical flow output v_g and v_r
9.Calculate cosine embedding loss of $v_g$and v_r
Output: $\mathcal{L}_{lk}$

3. EXPERIMENT

3.1. Dataset

A Mandarin speech-ultrasound dataset, comprising 6.85 hours of data was collected from 44 healthy individuals engaged in three distinct speech tasks: vowel, word, and sentence.The training set and test set consisted of approximately 5.48 and 1.37 hours of data, respectively, with no overlap between the sets. The UTI data were acquired using a commercial ultrasound system in a midsagittal orientation, with a sampling rate of 60 frames per second and a resolution of 920×700. Synchronization between the speech signals and UTI data was achieved using an external sound card. Due to computing resource limitations, we focused only on the vowels and words tasks.

3.2. Implementation Details

The model was trained on 4 × NVIDIA A6000 GPUs with a batch size of 4. The experimental setup included the following hyperparameters: minimum noise level of 0.002, maximum noise level of 160, data distribution standard deviation of 0.25, resolution size is of 112, and learning rate of 5×10^{-4}. The entire network underwent training for 5×10^6 iterations, with LK iterations set at 1 times. The training data was divided into patches with a window size of 15, and the model was sampled every 1000 iterations. Initial number of optical flow points N is 0.0005 of pixel value , and the optical flow loss weight started at 5×10^{-6}, increasing as the number of training iterations grew. The audio encoder is a Wav2Vec2.0 pretrained model obtained from a 56k-hour corpus, with all its parameters frozen during our training process [22].

3.3. Evaluation Metrics

Objective and subjective evaluation metrics are used to assess the proposed method. Objective evaluation includes Root Mean Square Error (RMSE), Peak Signal-to-Noise Ratio (PSNR), Learned Perceptual Image Patch Similarity (LPIPS) [23], and Fréchet Inception Distance (FID) [24]. These objective metrics offer a quantified assessment of image quality based on pixel-level comparisons. However, they are not capable of evaluating the dynamics and subtle differences of articulation captured in ultrasound tongue trajectories. This limitation is notable in the context of AAI, especially when using generated images for phonetic and clinical analysis. [25, 26, 27, 28]

To address this limitation, we developed a more detailed and well-designed evaluation method for the generated ultrasound tongue images. We invited 21 participants to score the generated ultrasound tongue images data on three dimensions using a 1-5 Likert scale, and their Mean Opinion Score (MOS) was calculated for analysis. Three dimensions related to tongue motion are Fluency (F-MOS), Intelligibility (I-MOS), and Trajectory Similarity (T-MOS). This evaluation designed to focus reviewers' attention on the trajectories of UTI, which compensates for the limitation of tradition evaluation metrics. More examples used are uploaded to https://github.com/Ytimed2020/Optical_UTI.

3.4. Quantitative Result

The performance of the diffusion model optimized with optical flow constraints is shown in Table 1. Our model has achieved improvements in RMSE, PSNR, and LPIPS. On the FID index, which reflects the overall distribution quality of the generated data, our model showed an improvement of 7.96% over the baseline model. This result indicates that our method reconstructs tongue images excellently at the pixel level. However, these objective metrics assess pixel-level similarity and cannot

account for tongue movement trajectories. The assessments focus on a micro-scale (pixel in image) comparison, while tongue movement trajectories are presented on a macro-scale (curve in image). Since tongue movement trajectories is essential for analyzing the mapping mechanisms from acoustics to articulation, additional metrics are needed to comprehensively analyze the model's performance.

Table 1: *Performance comparison of various ultrasound-based AAI systems on the objective evaluation*

Methods	RMSE↓	PSNR↑	LPIPS↓	FID↓
DNN	5.7674	32.9245	0.3201	322.06
Diffusion Model	5.6738	33.0813	0.1958	28.89
Ours	**5.6689**	**33.0882**	**0.1882**	**26.76**

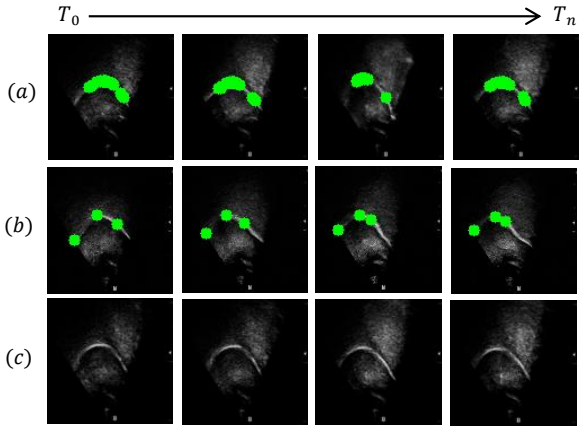

Figure 2: *The (a) and (b) demonstrate tracking trajectories of optical flow points in varying quantities over time T, against UTI, while the (c) displays the original image sequence.*

3.5. Subjective Analysis Result

Table 2 presents the scoring results from 21 participants using our evaluation method. For F-MOS, our method achieved a score of 3.5474, significantly higher than the baseline diffusion model score of 3.3526, indicating that the UTI data generated by our method exhibits more fluency in articulatory trajectory movements. In terms of ntelligibility, our method also achieved the highest score of 3.4684, surpassing the baseline score of 3.0526. The high scores for fluency and intelligibility, which involve smoothness in tongue motion and authenticity in tongue contour and position, indicate that the tongue and its movement details are well represented in the generated data.

For the core dimension of our evaluation method, trajectory similarity, the optical flow-guided diffusion model scored the highest with MOS of 3.4526. This highest score validates the effectiveness of optical flow constraints in optimizing tongue motion trajectory reconstruction and highlights the importance of accurate tongue motion trajectories for the mapping relationship between acoustic signals and speech articulation. Examples and subjective files can be found on the website.

To explore whether the different initial point setups in optical flow will influence UTI generation , we conducted 3 experiments and measured the results using subjective evaluation metrics. As shown in Table 3, when there is only one initial point (1N) in optical flow, the scores for all three metrics are relatively low. This is because a single point cannot fully describe overall tongue positional and movement information in three-dimensional space. Increasing the number of initial points

Table 2: *Subjective scores of various ultrasound-based AAI systems*

Methods	F-MOS↑	I-MOS↑	T-MOS↑
DNN	1.0190	1.0143	1.0619
Diffusion Model	3.3762	3.1095	3.0047
Ours	**3.5810**	**3.5190**	**3.5000**

to three (3N) significantly improves the scores for all three metrics, but the trajectory similarity score of 3.2631 is not the highest. This suggests that while three points can capture the basic tongue trajectories, they may still be inadequate to capture detailed trajectories. When increased to five initial points (5N), intelligibility and trajectory similarity improved, with a slight reduction in fluency score compared to three initial points. These changes confirm that moderately increasing the number of initial points helps enhance the detail of tongue trajectory in UTI generation. However, too many points can introduce additional computational complexity and interference, negatively affecting other metrics. Therefore, the number of initial points needs to balance between trajectory accuracy and other metrics.

Table 3: *The ablation study of setting different numbers of optical flow initial points on generating UTI*

Methods	Points	F-MOS↑	I-MOS↑	T-MOS↑
Ours	$1N$	3.3333	2.9048	2.8476
	$3N$	**3.6810**	3.2905	3.2571
	$5N$	3.5810	**3.5190**	**3.5000**

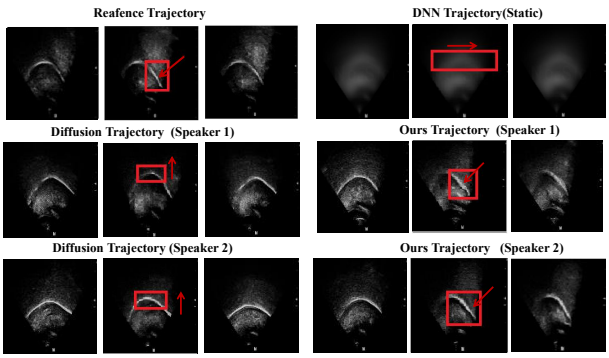

Figure 3: *The generation trajectory direction of different models*

4. Conclusion

This paper proposed an Optical Flow Guided tongue trajectory generation method for diffusion-based AAI model. By imposing constraints on the generative model, our approach addresses the limitations of UTI diffusion models that focus solely on pixel quality reconstruction without considering trajectory similarity. The generated UTI achieved highest scores on various evaluation metrics, especially in trajectory similarity. With more authentic trajectories, our results help expand the understanding of how tongue articulation mechanism works, which is particularly valuable for speech-language pathologists. Future work will be dedicated in investigating task-specific constraints on diffusion model for new generation problems.

5. Acknowledgment

This work is supported by National Natural Science Foundation of China (U23B2018, NSFC 62271477), Shenzhen Science and Technology Program (JCYJ20220818101411025, JCYJ20220818101217037, JCYJ20220818102800001), and Shenzhen Peacock Team Project (KQTD20200820113106007).

6. References

[1] A. S. Shahrebabaki and G. Salvi et al, "Acoustic-to-articulatory mapping with joint optimization of deep speech enhancement and articulatory inversion models," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 135–147, 2022.

[2] M. S. Ribeiro, J. Cleland, A. Eshky, K. Richmond, and S. Renals, "Exploiting ultrasound tongue imaging for the automatic detection of speech articulation errors," *Speech Communication*, vol. 128, pp. 24–34, 2021.

[3] T. G. Csapó, F. V. Arthur, P. Nagy, and Ádám Boncz, "Towards Ultrasound Tongue Image prediction from EEG during speech production," in *Proc. INTERSPEECH 2023*, 2023, pp. 1164–1168.

[4] P. J. L, B. T. M. A, B. S. E, and et al, "Ultrasound images of the tongue: A tutorial for assessment and remediation of speech sound errors," pp. 3672–3676, 2017.

[5] E. Karimi, L. Ménard, and C. Laporte, "Fully-automated tongue detection in ultrasound images," *Computers in Biology and Medicine*, vol. 111, p. 103335, 2019.

[6] L. C and M. L, "Robust tongue tracking in ultrasound images: a multi-hypothesis approach," in *Proc. INTERSPEECH 2015*, 2015, pp. 633–637.

[7] J. Wang, Y. Yang, J. Wei, and J. Zhang, "Continuous ultrasound based tongue movement video synthesis from speech," in *ICASSP 2016*, 2016, pp. 1716–1720.

[8] T. G. Csapó and T. G. et al, "DNN-Based Ultrasound-to-Speech Conversion for a Silent Speech Interface," in *Proc. Interspeech 2017*, 2017, pp. 3672–3676.

[9] J. Ho, T. Salimans, A. Gritsenko, W. Chan, M. Norouzi, and D. J. Fleet, "Video diffusion models," in *Advances in Neural Information Processing Systems(NeurIPS)*, 2022.

[10] L. Zhu, X. Liu, X. Liu, R. Qian, Z. Liu, and L. Yu, "Taming diffusion models for audio-driven co-speech gesture generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023, pp. 10 544–10 553.

[11] Y. Yang, R. Su, X. Liu, N. Yan, and L. Wang, "An audio-textual diffusion model for converting speech signals into ultrasound tongue imaging data," in *ICASSP 2024*, 2024, pp. 1–5.

[12] Y. Akgul, C. Kambhamettu, and M. Stone, "Extraction and tracking of the tongue surface from ultrasound image sequences," in *Proceedings. 1998 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. No.98CB36231)*, 1998, pp. 298–303.

[13] X. Liu, X. Tan, Y. Guo, Y. Chen, and Z. Zhang, "CSTRM: Contrastive Self-Supervised Trajectory Representation Model for trajectory similarity computation," *Computer Communications*, vol. 185, pp. 159–167, 2022.

[14] M. Zhai, X. Xiang, N. Lv, and X. Kong, "Optical flow and scene flow estimation: A survey," *Pattern Recognition*, vol. 114, p. 107861, 2021.

[15] E. Evain, K. Faraz, T. Grenier, D. Garcia, M. De Craene, and O. Bernard, "A pilot study on convolutional neural networks for motion estimation from ultrasound images," *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, vol. 67, no. 12, pp. 2565–2573, 2020.

[16] C.-H. Chang, C.-N. Chou, and E. Y. Chang, "Clkn: Cascaded lucas-kanade networks for image alignment," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[17] S. Sun, Z. Kuang, L. Sheng, W. Ouyang, and W. Zhang, "Optical flow guided feature: A fast and robust motion representation for video action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[18] T. Karras, M. Aittala, T. Aila, and S. Laine, "Elucidating the design space of diffusion-based generative models," *NeurIPS 2022*, vol. 35, pp. 26 565–26 577, 2022.

[19] S. Baker and I. Matthews, "Lucas-Kanade 20 Years On: A Unifying Framework," *International Journal of Computer Vision*, vol. 56, no. 3, pp. 221–255, Feb. 2004. [Online]. Available: https://doi.org/10.1023/B:VISI.0000011205.11775.fd

[20] X. Dong, S.-I. Yu, X. Weng, S.-E. Wei, Y. Yang, and Y. Sheikh, "Supervision-by-registration: An unsupervised approach to improve the precision of facial landmark detectors," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[21] G. Yang and D. Ramanan, "Volumetric Correspondence Networks for Optical Flow," in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d. Alché-Buc, E. Fox, and R. Garnett, Eds., vol. 32. Curran Associates, Inc., 2019. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2019/file/bbf94b34eb32268ada57a3be5062fe7d-Paper.pdf

[22] A. Baevski and Y. Zhou et al, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *NeurIPS 2020*, vol. 33, pp. 12 449–12 460, 2020.

[23] F. N. Iandola and S. Hanet al, "Squeezenet: Alexnet-level accuracy with 50x fewer parameters and ¡0.5mb model size," *arXiv preprint arXiv:1602.07360*, 2016.

[24] I. Skorokhodov and S. Tulyakov et al, "Stylegan-v: A continuous video generator with the price, image quality and perks of stylegan2," in *CVPR 2022*, 2022, pp. 3626–3636.

[25] J. L. Preston, T. McAllister Byun, S. E. Boyce, S. Hamilton, and E. R.-C. A. W. D. H. Tiede, M.and Phillips, "Ultrasound images of the tongue: A tutorial for assessment and remediation of speech sound errors," in *J. Vis. Exp*, 2017.

[26] C.-Y. Chien, J.-W. Chen, C.-H. Chang, and C.-C. Huang, "Tracking Dynamic Tongue Motion in Ultrasound Images for Obstructive Sleep Apnea," *Ultrasound in Medicine & Biology*, vol. 43, no. 12, pp. 2791–2805, 2017.

[27] J. Y. Song and F. Eckman, "Using ultrasound tongue imaging to study covert contrasts in second-language learners' acquisition of English vowels," *Language Acquisition*, vol. 28, no. 4, pp. 344–369, 2021, publisher: Routledge _eprint: https://doi.org/10.1080/10489223.2021.1910266.

[28] S. Hu, X. Xie, M. Geng, M. Cui, J. Deng, G. Li, T. Wang, H. Meng, and X. Liu, "Exploiting Cross-Domain And Cross-Lingual Ultrasound Tongue Imaging Features For Elderly And Dysarthric Speech Recognition," in *Proc. INTERSPEECH 2023*, 2023, pp. 2313–2317.