Motion Diffusion From Speech

Kushal Khangaonkar
Sanjay Subramanian
Daniel Klein
Trevor Darrell

Electrical Engineering and Computer Sciences
University of California, Berkeley

Technical Report No. UCB/EECS-2024-113
http://www2.eecs.berkeley.edu/Pubs/TechRpts/2024/EECS-2024-113.html

May 16, 2024

Acknowledgement

Special thanks to Sanjay Subramanian, Professor Daniel Klein, and Professor Trevor Darrell.

Motion Diffusion From Speech

by Kushal Khangaonkar

Research Project

Submitted to the Department of Electrical Engineering and Computer Sciences, University of California at Berkeley, in partial satisfaction of the requirements for the degree of **Master of Science, Plan II**.

Approval for the Report and Comprehensive Examination:

Committee:

Professor Daniel Klein
Research Advisor

Dan Klein

(May 15th, 2024)

* * * * * * *

Professor Trevor Darrell
Second Reader

(May 15th 2024)

Motion Diffusion from Speech

5th Year M.S. Project Report by Kushal Khangaonkar
Special Thanks to Sanjay Subramanian, Professor Daniel Klein, and Professor Trevor Darrell

May 16, 2024

Abstract :

This project explores methods for motion synthesis from speech. Given a recorded speech sample we aim to generate joint angles for body and hand motion that is realistic and corresponds to the input speech. We propose a diffusion-based method that uses Prosody Embeddings as conditioning for a transformer encoder diffusion model. Our work emphasizes the importance of classifier-free guidance during generation as a key factor in improving accuracy and realism of generated motion. We also find that using a velocity loss term is a crucial aspect of learning motion patterns. Our results show that there is potential for Prosody Embeddings as conditioning for realistic Motion synthesis, but additional conditioning may be necessary to generate motion with semantic connection to the input speech.

1 Introduction

Body and hand motion is one of the most important forms of non-verbal communication and is used to convey emotion at a more complex level than speech on its own. Motion can often provide additional cues and context to the spoken words, by emphasizing key points or adding nuance to the speech content. In AI, there is significant successful research work on building systems from speech [13] and also from text [11] which explore a variety of methods for generating realistic motion. Our work aims to explore how conditioning from speech models can be used to potentially improve motion synthesis.

The data sets we utilize for this work are the TalkSHOW [13] data and a subset of the Beats v1.3 dataset [7]. This data comes from speakers, talk show hosts, and lecturers, and captures expressive communication of speech and gestures. We work with a train dataset of 13.5 thousand samples (most of which are 10 seconds) and an evaluation set of 2k samples. We additionally train a model on the TalkSHOW data of 11k train samples for evaluation comparison with the TalkSHOW results.

Gesture generation for body and hand motion is a non-deterministic problem when generating from speech or text, as there are a vast amount of possible generations for given speech that capture the intended emotion and expression. While model training is achieved through Mean Squared Error, evaluating in this style would not capture whether or not the model is generalizing successfully. We evaluate for realism using Frechet distance between ground truth and predicted latents (calculated using a pretrained body motion auto encoder) to measure difference in latent space probability distribution. Additionally, we will evaluate diversity by calculating variance of outputted samples.

Through our experiments, we found that a variety of factors influenced the success of our model. Two key findings were the impact of classifier-free guidance and velocity loss. In this paper, we will explore how these factors enable diffusion to successfully generate realistic motion.

2 Related Work

2.1 Recent Methods for motion synthesis from speech

TalkSHOW: Generating Holistic 3D Human Motion from Speech [13] provides the main dataset we utilize and approaches the body/hand part of the problem through a compositional vector-quantized variational autoencoder (VQ-VAE). This involves training encoders and decoders to map gestures to/from a discrete body token codebook and hand token codebook. Then an autoregressive model is trained to output these discrete tokens given audio (post feature extraction). The TalkSHOW dataset is composed of four main speakers (Conan, Chemistry, Oliver, Seth), three talk-show hosts, and one lecturer who provide strong examples of expressive motion. The VQ-VAE method provides strong diversity of generation and the cross-conditional autoregressive model should provide both qualitative and quantitative evidence of realistic output.

The EMAGE model [6] improves on the TalkSHOW method through masked gesture generation as previously used in language and vision models [3]. Similarly to TalkSHOW they leverage a VQ-VAE discrete codebook for gesture reconstruction. Their method shows that masked reconstruction can improve gesture generation, and is evaluated on a new dataset they compiled (BEAT 2.0). While the TalkSHOW and EMAGE methods are quite successful, neither utilize conditioning from existing Speech models and rely on processing audio with their newly trained models. Thus, there is potentially room for improvement through conditioned generation from successful Speech models.

2.2 Diffusion for Generative AI

One of the most successful recent methods for diverse conditioned generation is Diffusion Probabilistic models, initially used for text conditioned image generation [4]. Diffusion involves training a model to "denoise" step by step to generate output. This is accomplished by adding noise to the training data and training the model to learn the distribution difference between each diffusion step. Then at generation the model can start from noise and through several model steps denoise the data back to the ground-truth distribution. A key aspect of this is conditioning the diffusion model to guide the denoising process to the specific data the model should generate.

In Human Motion Diffusion Model [11], diffusion is used to generate motion from text-based task descriptions, for example, "A person walking in a straight line," or "A person tying their shoe." This is achieved through a language embedding generated from CLIP [9] used as a conditioning input for the diffusion model. MDM's success in using text conditioning inspired our method as we attempt to apply diffusion for speech to motion synthesis.

3 Method

For a given audio sample, we aim to generate motion that corresponds to the sample and appears realistic and accurate to the speech. Gestures are measured in joint angles of labelled body parts at 30 frames per second. For this work we focus on body and hand joints, which are 39-dimensional vectors representing joint angles of the upper body, and 90-dimensional vectors representing angles of the hand joints. We explored deep learning methods and decided to approach this problem through a diffusion framework that is conditioned on a high-level representation of the input audio.

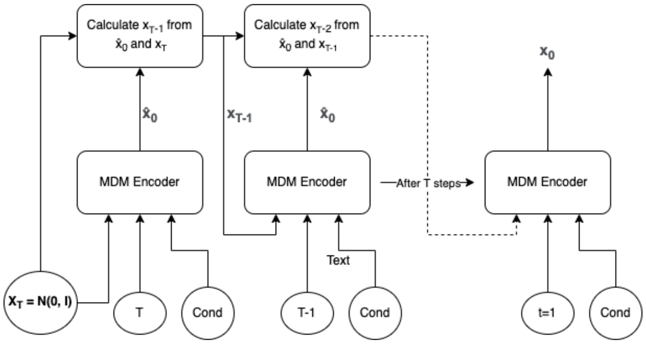

Figure 1. The MDM Diffusion process, starting from a noise distribution and with each step outputting a prediction of the gestures (as opposed to noise). This enables geometric loss terms that use the ground truth gestures.

3.1 Motion Generation via Diffusion

Motion is both high dimensional and temporal, and thus it is difficult to train a single model to output Motion directly from audio. TalkSHOW and Emage circumvent this by learning a discrete codebook representation of gestures. Experiments with transformer systems (both autoregressive and non-autoreggresive) were not successful at directly outputting the 129 dimensional gestures, but Human Motion Diffusion Model [11] inspired us to approach this problem through denoising. Their model (MDM) is a transformer encoder that can be conditioned through a single CLIP language embedding vector that is summed with the diffusion timestep conditioning and given as the first input to the transformer encoder. A key aspect of this system is it predicts the sample at each diffusion time step [10] instead of the noise as seen in Figure 1 (contrary to how typical diffusion models output). The input for the next step of denoising is calculated by diffusing the predicted gestures up to the following timestep. Calculating the raw gestures at each step enables calculating geometric loss terms, such as using the ground truth for Mean Squared Error and using the generated motion for velocity loss. Additionally, this work used **classifier-free guidance** through unconditioned training in order to guide output generation to match the input conditioning more closely. This is an important aspect of their method that was pivotal to the realism of our speech-to-motion model generations. [5].

3.2 Prosody Embeddings from Seamless Expressive

As our problem is from speech audio to motion, conditioning a diffusion transformer is not easily solved in the same way MDM uses CLIP Language Embeddings. This is where leveraging work in Speech generation has potential to lead to strong motion synthesis. Seamless Expressive [1] is a model system from Facebook that translate speech to speech across several languages. This system first encodes input speech represented as 80-dim MEL filterbarks to high level XLSR units. Additionally they use a separate model to encode the MEL filterbanks into a single vector prosody embedding. Then, these units are converted to the translated language using transformers conditioned on the target language and these prosody embeddings.

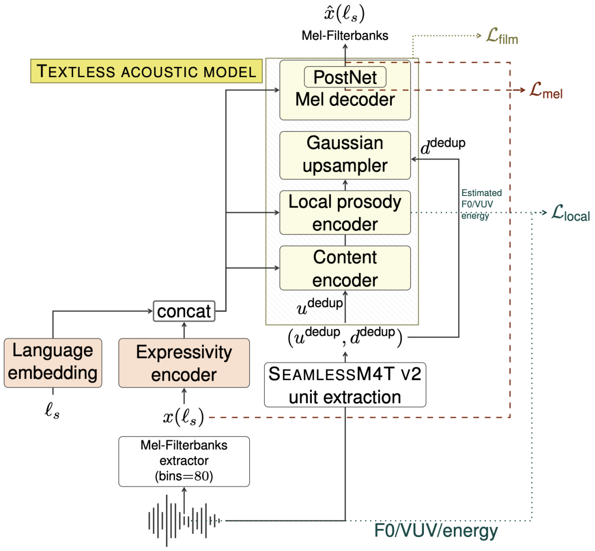

Figure 2. Visual from [1]. The PRETSSEL Textless acoustic Model utilizes Expressivity encodings as conditioning for the audio synthesis pipeline to maintain style and paralinguistics from source language during translation. We can use these encodings for a similar effect in motion synthesis.

Prosody in this context refers to paralinguistic communication, pertaining not non-semantic aspects of the input speech including style, emphasis and rhythm. In the context of language translation this is crucial in maintaining the same vocal style in the output speech. These embeddings are calculated using a expressivity encoder based on the ECAPA-TDNN architecture (strong at capturing acoustic representations [2]) that extracts a 512-dimensional vector directly from the input speech MEL filterbanks. This expressivity encoder is jointly trained with the PRETSSEL [1] textless acoustic model which outputs raw audio, and thus learns temporally aligned prosody. During audio synthesis PRETSSEL uses the expressivity embedding as conditioning at each step of the generation process as seen in Figure 2. We propose that this prosody vector can be used for a similar effect to the CLIP embeddings in body/hand motion generation.

We additionally experimented with using the Seamless PRETSSEL Model to encode XLSR unit representations of the input audio for additional semantic conditioning, but experiments thus far have been unsuccessful in consistently outputting reasonable motion (see Figure 6. This will be explored more in future work.

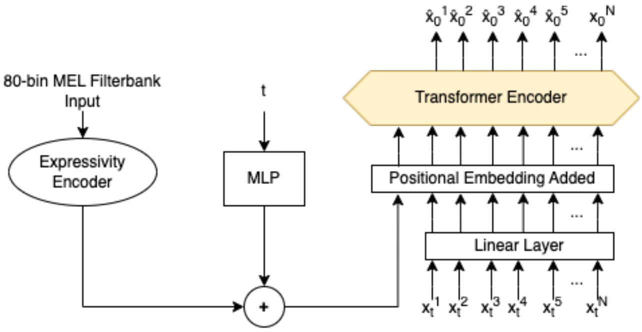
$\hat{x}_0^1$ $\hat{x}_0^2$ $\hat{x}_0^3$ $\hat{x}_0^4$ $\hat{x}_0^5$ ... x_0^N
80-bin MEL Filterbank Input
t
Transformer Encoder
Expressivity Encoder
MLP
Positional Embedding Added
Linear Layer
+
x_t^1 x_t^2 x_t^3 x_t^4 x_t^5 ... x_t^N

Figure 3. We utilize prosody embeddings as conditioning for the MDM model. The model input is a single conditioning vector of timestep + prosody (sometimes masked out for unconditional generation), followed by the gesture sequence at diffusion timestep t. The output is the corresponding gesture sequence at t=0, which will be diffused back to timestep t-1 for the next diffusion step.

3.3 *Model and Training Experiments*

Our diffusion model architecture is visualized in **Figure 3**, and is centered around a Transformer encoder following the standard architecture of Vaswani et al [12] and following hyperparameters similar to Motion Diffusion Model ($d_{model} = 512$). The additional model components include:

- **Expressivity Encoder**: Calculates Prosody embeddings as a single vector of size d_{model} from input Audio (given as a 80-bin MEL filterbank).
- **Timestep MLP**: Encodes the input diffusion timestep to a single vector of size d_{model}. As seen in **Figure 3** we sum this conditioning with the Prosody embeddings
- **Positional Encoder**: We use an unweighted positional encoder following [12] to encode the Timestep and in the input to the transformer
- **Gesture Upscale**: Linear layer to project input gestures from 129-dim vectors (39 body + 90 hand) to d_{model}
- **Gesture Projection**: Linear layer to downscale Transformer outputs from d_{model} to gesture dim.

Initially our training experiments resulted in visually incorrect outputs, even though MSELoss was decreasing the output would be completely different from the ground truth (even training set examples). Even when overfitting on a tiny subset of data (e.g. 16 data points) a shakiness would appear, when the loss would indicate that the output should match the ground truth almost exactly. Concretely, the output would match the pattern of the ground truth but the hands and body would be jittering at a high frequency that made the output unrealistic. This was solved at generation time through classifier-free guidance [5]. In other diffusion models guidance (specifically image generation) is applied using image classifiers to ensure the model outputs what the conditioning dictates. In Motion Diffusion Model they apply a classifier free guidance by weighing unconditioned output with conditioned output which in the context of the text-to-motion model effectively trades diversity in output for accuracy and adherence to the text conditioning:

$$\hat{x} = \hat{x}_{uncond} + \lambda_{guide}(\hat{x}_{cond} - \hat{x}_{uncond}) \tag{1}$$

In our application, adding this unconditioned output term with a guidance scale of 2.5 seems to crucially improve the realism of the outputs . This required training the model for unconditioned examples as well, where we would mask out the prosody conditioning randomly for a subset of each batch during training so that the model learns to generate good motion without needing prosody guidance.

When scaling up to the full dataset, another problem we encountered was the model biasing towards less motion even after hundreds of epochs. One key finding we found experimentally was the importance of velocity loss in synthesis from speech. Velocity loss refers to adding a loss term equal to the average delta on output gestures. With $\hat{x_0}^{0:N}$ as output generation where the sequence length is N, and x_0 as ground truth we can define our training loss function.

$$L_{mse} = (x_0 - \hat{x_0})^2 \tag{2}$$

$$L_{vel} = \lambda_{vel} * (\hat{x_0}^{1:N} - x_0^{0:N-1}) \tag{3}$$

$$L_{tot} = L_{mse} + L_{vel} \tag{4}$$

This loss term being added to the MSELoss should encourage the model to match the velocity and frame changes of the ground truth.

Our best model was trained with the following parameters:

Parameter	Value
Encoder Dim	512
Attn Heads	4
Encoder Lyrs	8
Feedforward dim	1024
Batch Size	64
Learning Rate	0.0001
λ_{vel}	5.0
Uncond Training Prob	0.1
Num Diffusion Steps	1000
Total Train steps	500000

Table 1: Transformer Encoder and Training Hyperparameters

4 Results

4.1 *Our Quantitative Metrics*

- **Realism**: We define **FGD** as the Frechet distance between latent representations of ground truth gestures and predicted gestures, and use FGD as our measure of realism.
  Frechet distance is a metric for calculating the distance between two distributions, and is commonly used in Motion synthesis [8][6]. Using TalkSHOW's pretrained body autoencoder we could map our generated gestures and the ground truth motion into a motion latent space to measure distribution difference between generation and ground truth. Define μ_p, Σ_p as the mean and covariance of the prediction latents and μ_g, Σ_g as the mean and covariance of the ground truth latents.

$$\mathbf{FGD} = d^2 = |\mu_p - \mu_g| + \mathbf{tr}(\Sigma_p + \Sigma_g - 2(\Sigma_p\Sigma_g)^{1/2}) \tag{5}$$

The closer the generated distribution is to the motion latent space, the likelihood that the generations are visually realistic and coherent increases in accordance. This is a way to measure the realism of our generations. In the TalkSHOW paper they train a classifier for real and AI motion, but as their classifier is unreleased, we shall use the Frechet distance as our primary metric of realism.

- **Diversity**: We measure diversity in output generation with variance. The variance **Var** is calculated from 8 samples generated for each input. For a particular input, **Var** is calculated across the sample dim, then averaged over the gesture and time dimensions. This gives us a metric for diversity, as a model that has high output Variance and low Frechet distance is both realistic and diverse.
- **Activity**: To report the effect of velocity training, we can calculate the max generated velocity **MaxV** (by magnitude) for each sequence in the evaluation set. With $\hat{x}^{0:N}_{(i)}$ as the i'th output generation of length N for a evaluation set of size K, we calculate MaxV as follows:

$$\mathbf{MaxV} = \frac{1}{K}\sum_{i=0}^{K} \max|\hat{x}^{1:N}_{(i)} - \hat{x}^{0:N-1}_{(i)}| \tag{6}$$

4.2 Quantitative Evaluation

To evaluate against the TalkSHOW model we use a version of the model trained on our train/val/test split of the data with a codebook size of 512 for hand and 512 body (original paper uses 1024 for both). The best TalkSHOW model from the original paper was also cross conditioned with the face motion, which our problem does not include. The main version of our Diffusion model was trained on about 2k more datapoints from the Beats v1.3 dataset. We additionally compare with a version of our model only trained on the TalkSHOW data for direct comparison.

Model	FGD ↓	Var ↑	MaxV
TalkSHOW	22.94	.969	0.58
Ours (w/o beats) $\lambda_{vel} = 0.5$	23.03	.487	0.53
Ours, $\lambda_{vel} = 0.0$	135.90	.669	5.58e-05
Ours, $\lambda_{vel} = 0.2$	25.92	.583	0.48
Ours (w/o CF guidance) $\lambda_{vel} = 0.5$	137.32	.619	0.52
Ours, $\lambda_{vel} = 0.5$	24.78	.562	0.55

Table 2: Quantitative realism and diversity evaluation on TalkSHOW test set. ↑ and ↓ indicate whether a higher number is better or worse respectively.

Looking at the realism metric, our model with velocity = 0.5 and classifier-free guidance trained on the TalkSHOW data exclusively led to the best result of **23.03**, only 0.07 less then the TalkSHOW model we trained. We can see the impact of both guided generation and velocity training as models without these additional elements were generating output very far away from the ground truth latent space.

Our model was unable to match the same level of diversity as the TalkSHOW model, in fact our best score on this metric came from our model trained without velocity loss. This leads us to conclude that **1.** Diversity is an important metric only if the model is generating realistic results, and **2.** Prosody Embeddings do not give enough context for the model to learn a variety of gesture patterns which results in our outputs following similar patterns each generation.

We additionally see that without a velocity component in the loss during training, Our method is unable to learn motion and essentially learns to stand still as a way to minimize loss.

4.3 Qualitative Analysis

When visually inspecting the outputs from our trained model, 3 main takeaways stand out:

- Prosody embeddings preserve realistic motion and joint angles are cohesive.
- Our generations are connected with the emotion of the speaker but not with the words.
- Our model lacks semantically meaningful generations.

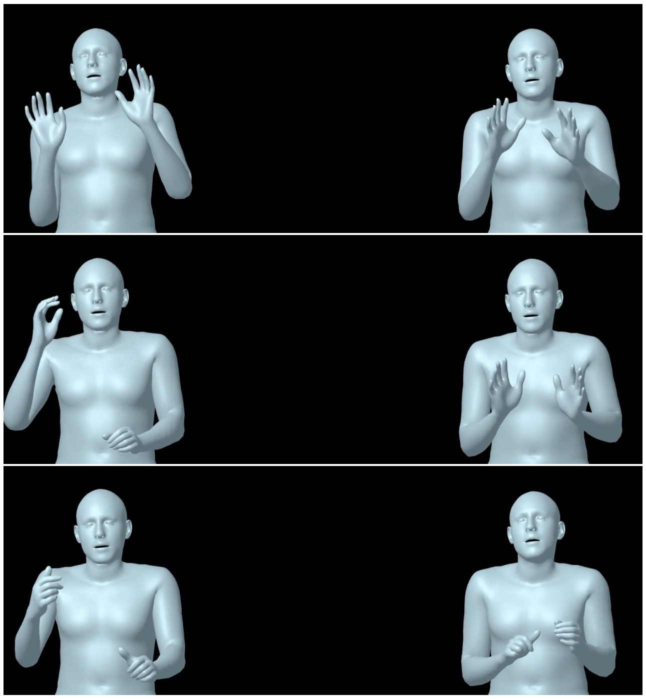

Figure 4. 3 snapshots in order from top to bottom from a John Oliver Motion generation. Our generation is on the left and the ground truth is on the right. The speech containing these frames is: "Maduro, The **big** banana fan comes in because because he was Chavez's hand picked successor but unfortunately had neither his **booming** economy nor his charisma, although he is **dead**." Each **bolded** word corresponds with each frame in order.

Our model generates plausible motion that can often pass for reasonable gestures, but fails to be expressive enough to be useful yet. Example outputs appear robotic on occasion but usually follow reasonable patterns and realistic motions.

However many outputs seem to be less expressive as the ground truth, and often we can observe this as a lack of connection to the specific words being spoken. For example, during a clip where the speakers pitch changes to a more aggressive emotional tone, the hand gestures will match the change in energy, but they are unlikely to match the emphasis on syllables that are in the ground truth.

Looking at the example in Figure 4, the beginning of the speech matches the pose pretty well, but our generation switches to an alternate position earlier than the ground truth, By itself, this isn't a huge issue as long as the movement was realistic. Once in the new position, the speaker (John Oliver) begins to emphasis words in a way that the hands are forcefully motioning downwards with each down-syllable. Specifically, He says "H**an**d P**ick**ed Succ**ess**or", which corresponds to the rhythm of the hand motion. Our generation has similar hand motion but doesn't match the rhythm of the ground truth.

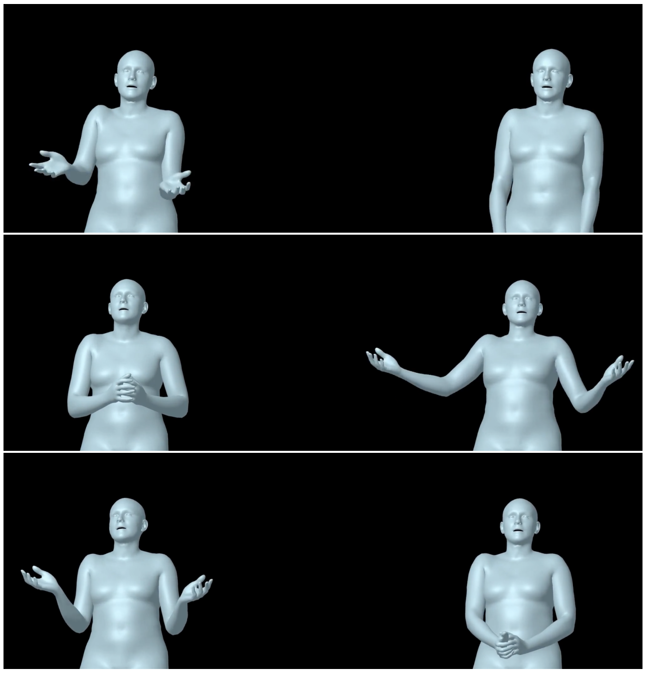

Figure 5. 3 snapshots in order from top to bottom from a Conan motion generation. Our generation is on the left and the ground truth is on the right. The speech corresponding to the motion: "In a new **interview**, President Trump reveals that he tweets in bed...**(pause)**...yeah. When **asked** about this..."

In addition to sometimes missing on rhythm, the generation will occasionally output the correct gesture at the wrong time. For example the Conan generation in Figure 5 includes Conan stating a headline and pausing ironically. During the pause, he puts his arms up in exasperation. In our generation, this particular motion happens both before the pause and after the pause. Its certainly difficult to learn a specific motion during speaker silence, but a more realistic output would not have this motion during the Conan's recitation of the headline. This suggests that the Prosody embeddings do not contain quite enough temporal information to consistently output gestures at the precise time, especially during pauses.

Our main attempt at solving this was through PRETSSEL [1] encoded sequences, which are prosody conditioned semantic sequences, and after decoding can be up-sampled to audio. This intermediate sequence seemed plausible to be used as in-context inputted right before the gestures in our transformer encoder. We conducted several experiments with this concept and it appeared that this occasionally imparted the context we were looking for, specifically gestures that were more connected with what the speaker was saying. However, realistic gestures were only achieved a fraction of the time, and more commonly the motion would be random and unrealistic. Looking at Figure 6 the motion does occasionally match the rhythm and cadence of the words being spoken but repeatedly enters unrealistic positions and wrangled motion: in this case the hands splay through each other unnaturally on the word "Chavez" before returning to more normal motion. While this method did not work we are still led to believe that additional conditioning above prosody embeddings is necessary to output better gestures.

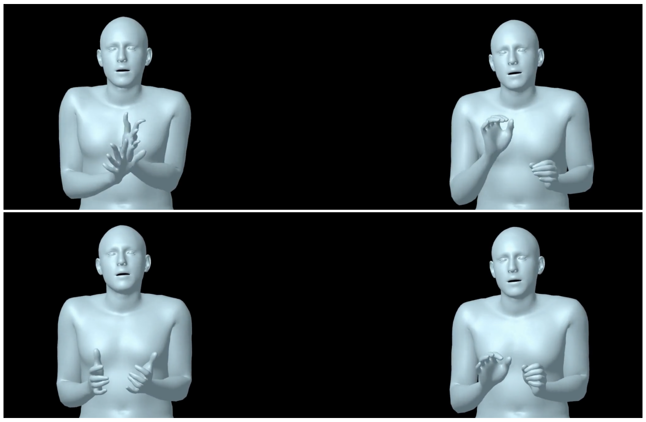

Figure 6. 2 snapshots in order from top to bottom from a Oliver motion generation using PRETSSEL encoder sequences as conditioning. Our generation is on the left and the ground truth is on the right. The sentence being spoken is "**Chavez** was massively popular in venezuela, **beloved** for both his generous social programs and his large...".

5 Conclusion

Motion Diffusion from from speech can be approached in a variety of ways, and we sought to use conditioning obtained from existing successful speech models. The Prosody Embeddings we used from SeamlessExpressive resulted in motion comparably realistic to the TalkSHOW model, achieved through our use of velocity loss and classifier-free guidance. However our generations were not as high in diversity of output and not as qualitatively connected to the input speech. A main reason for this is that Prosody embeddings were not enough to connect motion to syllable level emphasis and word level meaning. As we build on this work, it will be crucial to improve our conditioning to capture more information about the speech while preserving realism.

References

[1] Seamless Communication, Loïc Barrault, Yu-An Chung, Mariano Coria Meglioli, David Dale, Ning Dong, Mark Duppenthaler, Paul-Ambroise Duquenne, Brian Ellis, Hady Elsahar, Justin Haaheim, John Hoffman, Min-Jae Hwang, Hirofumi Inaguma, Christopher Klaiber, Ilia Kulikov, Pengwei Li, Daniel Licht, Jean Maillard, Ruslan Mavlyutov, Alice Rakotoarison, Kaushik Ram Sadagopan, Abinesh Ramakrishnan, Tuan Tran, Guillaume Wenzek, Yilin Yang, Ethan Ye, Ivan Evtimov, Pierre Fernandez, Cynthia Gao, Prangthip Hansanti, Elahe Kalbassi, Amanda Kallet, Artyom Kozhevnikov, Gabriel Mejia Gonzalez, Robin San Roman, Christophe Touret, Corinne Wong, Carleigh Wood, Bokai Yu, Pierre An-

drews, Can Balioglu, Peng-Jen Chen, Marta R. Costa-jussà, Maha Elbayad, Hongyu Gong, Francisco Guzmán, Kevin Heffernan, Somya Jain, Justine Kao, Ann Lee, Xutai Ma, Alex Mourachko, Benjamin Peloquin, Juan Pino, Sravya Popuri, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, Anna Sun, Paden Tomasello, Changhan Wang, Jeff Wang, Skyler Wang, and Mary Williamson. Seamless: Multilingual expressive and streaming speech translation, 2023.

[2] Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification, October 2020.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[4] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020.

[5] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance, 2022.

[6] Haiyang Liu, Zihao Zhu, Giorgio Becherini, Yichen Peng, Mingyang Su, You Zhou, Xuefei Zhe, Naoya Iwamoto, Bo Zheng, and Michael J. Black. Emage: Towards unified holistic co-speech gesture generation via expressive masked audio gesture modeling, 2024.

[7] Haiyang Liu, Zihao Zhu, Naoya Iwamoto, Yichen Peng, Zhengqing Li, You Zhou, Elif Bozkurt, and Bo Zheng. Beat: A large-scale semantic and emotional multi-modal dataset for conversational gestures synthesis, 2022.

[8] Evonne Ng, Hanbyul Joo, Liwen Hu, Hao Li, Trevor Darrell, Angjoo Kanazawa, and Shiry Ginosar. Learning to listen: Modeling non-deterministic dyadic facial motion, 2022.

[9] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.

[10] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents, 2022.

[11] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H. Bermano. Human motion diffusion model, 2022.

[12] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023.

[13] Hongwei Yi, Hualin Liang, Yifei Liu, Qiong Cao, Yandong Wen, Timo Bolkart, Dacheng Tao, and Michael J. Black. Generating holistic 3d human motion from speech, 2023.