# Gesture Generation using VAEs and LSTMs

**Somansh Budhwar**
HDSI
sbudhwar@ucsd.edu

## Abstract

Procedural generation in virtual reality, video games and animation has evolved exponentially recently. However, convincing gesture generation and a natural body language has been wanting. To bridge this gap, the goal of this project is to develop a speech-driven gesture generation model.

This paper begins with a literature review to gather information on previous research on speech-driven gesture generation. The GENEA Challenge, affiliated with Electronic Arts, is the most promising lead to work in this direction. It contains a well defined speech and gesture dataset, and a well laid out pipeline to render the gestures generated by a model to a video. The challenge led to multiple papers being published in ACM, including industry leaders such as Ubisoft. Based on the previous work three novel approaches were pursued, one using VAE for general body motion coupled with GAN for hand motion, second using VAE to generate 10 second gestures, and third using LSTM encoder-decoder architecture to generate sequence of gestures.

First approach was too unstable, while second approach yielded erratic results. The third approach was the smoothest but it degrades as input sequence gets longer. Second approach achieved a score of zero over all human evaluation metrics. LSTM based approach yielded a score of 0.3 out of 10 for human-likeness and 0.7 for appropriateness of gesture timing out of 10.

## 1 Background

Gesture generation using speech is an emerging field of research that focuses on developing computer systems capable of generating gestures from speech. The goal of this field is to create more natural and expressive human-computer interfaces that can enhance communication and interaction between humans and machines.

Gesture generation using speech is based on the idea that speech and gestures are closely linked in human communication. When people speak, they often use gestures to convey meaning and emphasize certain points. These gestures can be hand movements, facial expressions, or body posture changes. The aim of gesture generation using speech is to develop algorithms and models that can generate appropriate and contextually relevant gestures to complement spoken language.

There has been significant research in this field in recent years. Early work in this area focused on developing rule-based systems that generated gestures based on predefined rules and templates. However, these systems lacked the flexibility to generate natural and expressive gestures that could be customized to individual speakers or contexts. In other words, most of the models work on rule-based and supervised learning approaches have still not achieved a desirable performance.

One recent development has been in the emergence of GENEA challenge[5] hosted by Electronic Arts. This challenge saw the rise of developing machine learning models that can learn to generate gestures from speech data coupled with text and speakers identity. These models typically use deep learning algorithms and large datasets of audio and video recordings of people speaking to learn patterns in speech-gesture correlations.

This paper aims to explore novel approaches that may be used to generate gestures using only speech data. This paper used the framework built by GENEA challenge and the dataset while altering the core models to explore the possibilities.

Key Contributions of this paper GItHub Code:

1. The Variational Auto-Encoder approach to take 300 frames at once to generate gesture distribution. It did not work well and resulted in erratic behavior.
2. Gesture generation using LSTM with 3-inputs (audio, text, identity). LSTM based generation was smooth, indicated signs of natural movements, and captured excitement in speech and silence well. But with more frames it degenerated and resulted in unnatural motions.
3. All models were built from scratch and in file model2.py in the repository, while training, gesture generation and data pre-processing files were altered to work with new model architectures and reduce computational complexity.
4. One model for gesture generation of whole body seems to be the better approach rather than a separate models for hand motion.

Link to GItHub repository: https://github.com/somanshbudhwar/ece285_code/tree/main

## 2 Previous work

In the past GENEA challenges, various approaches were taken to generate smooth gestures. The general trend was to take a form of input such as text or audio, and convert it into an embedding, then combine that embedding with the agent type to generate an appropriate gesture. Various approaches were used such as, rule-based gesture generation by Gesture Mater, Variational Auto-encoders, Generative Adversarial Networks, and Normalizing flows. A brief overview is as follows.

Text2gestures[1] used only text and agent attributes to generate convincing gestures. It creates a word embedding from only the text input and concatenates it with the speaker's identity (gender, role). It then passes it through a transformer based encoding layer, and the output is combined with the past gesture frame to predict the next gesture. Text2gestures achieved a rating of "plausible" by human evaluators on the web. Gesticulator[3] and Nvidea's Audio2gesture use audio and text combination to generate gestures. Gesticulator concats and aligns the audio and text embedding, then passes it through a feed forward neeural network which auto-regressively outputs the sequence of gestures. Whereas Trimodal[4] used all three inputs and applied a Generative Adversarial Network to generate gestures. However, the results of above approaches were generic and not as expressive, as evaluated by humans.

So, IVI Labs[2] altered the Gesticulator's approach by changing two things. First, they used all 3 inputs instead, and secondly they used Tacotron2 as backbone model to synthesise gestures as shown in figure 3. This approach yielded the second highest appropriateness score for full body gesture. Hence taking three inputs namely, audio, text and speaker identity seem to be the way forward. Due to computational limitations, transformer based approaches have not been considered, thus VAE, LSTMs and GAN were chosen to create a novel method to generate gestures.

## 3 Data

The data provided in GENEA challenge 2023 is a private unlabelled dataset containing about 18 hours of audio and motion dataset, including transcript. The data was processed using the IVI baseline code. It takes 3 inputs namely: Speaker identity(17 speakers), audio, and the text transcript. The audio and text are aligned for each frame. Then, Librosa and Parselmouth library were used to extract features such as mel-spectrograms, Mel-frequency cepstral coefficients (MFCCs), and prosody features from the audio. Mel-spectrograms and MFCC have been established as important for gesture generation, and the authors of IVI baseline model also decided to extract prosody features to extract emotion intensity and pitch from the audio. These features account for 110 long vector for each frame. The text is converted to word embedding of size 300 using FastText for each frame. Although this results in repetition of embeddings, it is used since it aligns with each frame for gesture generation. Finally, the speaker identity is one-hot encoded. Thus, the overall embedding for each frame comes to be 427 features.

For the motion data, since finger data was not reliable, it was excluded and only 25 joints were chosen to represent full body, as shown in figure. The representation of joint motion is a 78 dimensional embedding. Thus, the training data takes a batch of 300x427 tensor as input, 300 being the number of frames or sequence length, and the output is of the form 78x300.

# 4 Method

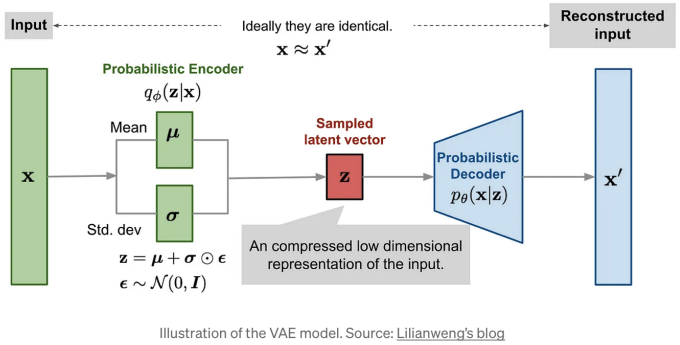



Figure 1: Variable Autoencoder for Generalist generation

$$\mathcal{L}(\phi, \theta, \mathbf{x^{(i)}}) = KL\big(q(\mathbf{z}|\mathbf{x}^{(i)}; \phi) \;||\; p(\mathbf{z}; \theta)\big) - \frac{1}{L}\sum_{l=1}^{L}[\log p(\mathbf{x}^{(i)}|\mathbf{z}^{(i\ l)}; \theta)]$$

Figure 2: Variable Autoencoder Loss function where $x_i$ the data

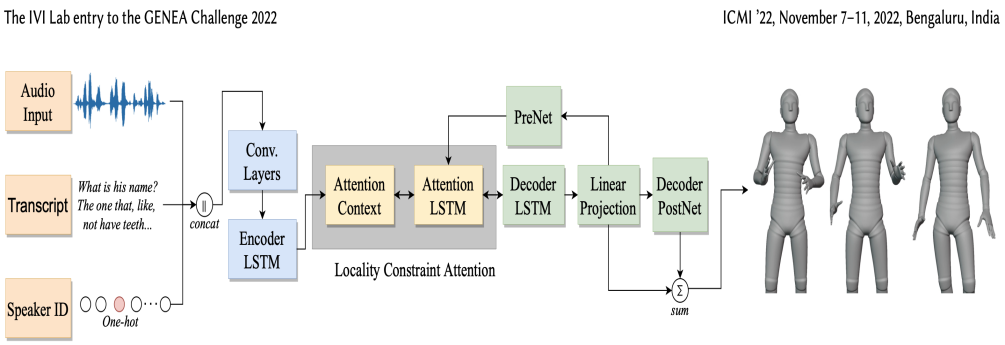



Figure 3: Tracton2 based IVI lab model

# 5 Approach and Architectures

The first approach was to use VAE for generating general body motion and GAN to generate hands and finger motion since it is more stochastic. However, due to data quality issues and highly unstable results this approach was promptly discarded. The idea of using two different models for different areas of the body was also discarded.

The second approach was to create a VAE model (figure 1) (128 latent dimensions) that takes all the 300 frames and generates a series of gestures at once. The assumption was that during 10 seconds

there is enough continuity of context and conversation topics that a probability distribution of a snippet can be created. It is akin to drawing a painting or creating music where there are high and low intensity notes bit there is also a continuity of context. The final model would break down the validation and test videos of varying lengths to 10 second segments and generate gestures block by block. As we see later this approach led to constrained but highly erratic motion.

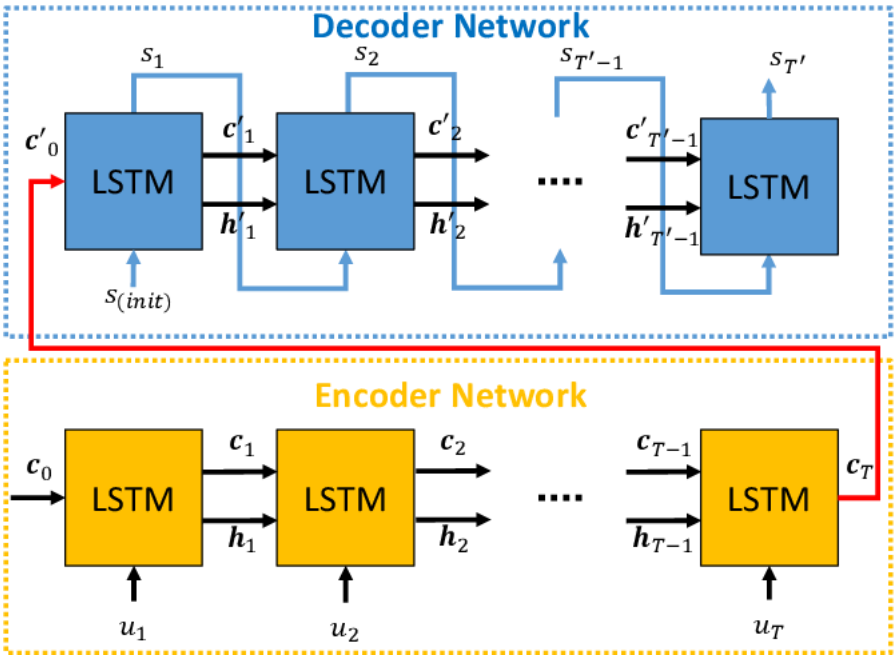



Figure 4: LSTM Encoder-Decoder architecture

The third approach treats it as a sequence-to-sequence problem rather than taking the whole 10 second input at once. So, an LSTM encode-decoder architecture (figure 4) was employed. The LSTM cell has just one layer. The LSTM encoder is fed a sequence of 300 frames with embedding size of 427 and each hidden unit outputs a 78 sized embedding. The decoder uses the last encoder output as input and decodes a 300 length sequence of 78 dimensions. However, since the output of the LSTM ranges between -1 to 1, a linear layer is also applied to the output to get the correct range of motion data. This approach led to smoother results however, they were not natural. Some notable aspects were that the gestures were intense during intense dialog and slower during silence. The limbic motion was of higher range when talking.

### 5.1 Advantages

The first approach in theory should allow us to parallelize the training process since both the processes can be dealt with separately. It can also allow us to fine-tune generalist and more expressive elements separately. However, this paper could not reach that far.

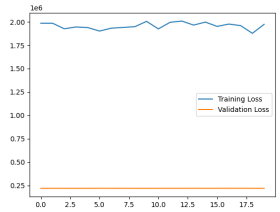
The second approach could be useful for generating gestures taking in the whole context of the scene. An attention based transformer architecture would be better suited for this approach such as the Tactron2 model But its limitation is that it cannot be used in real-time environments such as VR where the model cannot look ahead.

The third approach is good for real-time generation of gestures, and it was the approach that led to the smoothest results. However, it degenerates as the input sequence gets longer, so one solution can be to apply sliding window where past inputs are discarded and new inputs are added to the input.

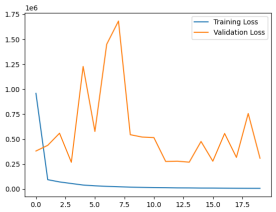
## 6 Experiment Setup and Results

Since the first approach was dropped, experiments were done using only second and third approach. For the VAE the effect of learning rate was negligible, while for the LSTM teacher forcing, learning rate, and batch size had noticeable impact. Following are the results from the experiment as seen in figures 5 and 6.

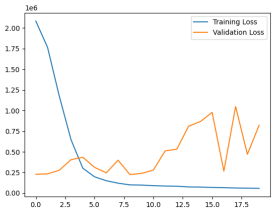
The experiments were done using UCSD Datahub resources on 2080Ti GPU. Hyper-parameters were chosen based on 2000 iterations and final models were trained for 20,000 iterations. The render pipeline used Blender 3.5 scripts provided by Electronic Arts.




(a) LSTM training without teacher forcing

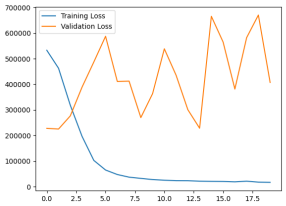



(b) LSTM training with teacher forcing

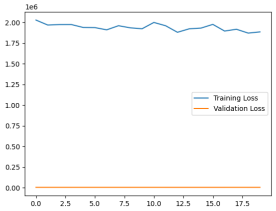



(c) LSTM training with learning rate 0.001

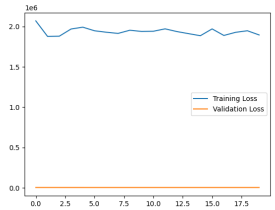
Figure 5: Model training Loss Curves




(a) LSTM training with learning rate 0.0001




(b) VAE training with learning rate 0.001 (Val loss approx 5900)




(c) VAE training with learning rate 0.0001 (Val loss approx 5800)

Figure 6: Model training Loss Curves

The best setting for the VAE model was a learning rate of 0.0001 and the best setting for the LSTM model was learning rate 0.001, teacher forcing ON and batch size of 64. Although the validation loss for the VAE indicates that it must be the better model, but its motions were repetitive and erratic as seen in the video posted on the repository. Whereas the LSTM model had increasing validation loss but had much more stable performance in the gesture generation. It had the flaw of being unnatural but was able to capture the intensity of the motion, smoothness of gestures and an overall coherence with the speech and silence.

### 6.1 Ratings by Humans

The motions were evaluated by 3 human evaluators (acquaintances). The average score is as follows

| | | | |
|---|---|---|---|
| Perceived human-likeness | 0 | 0 | 0 |
| Appropriateness (timing)l | 0 | 0 | 0 |
| Appropriateness (content) | 0 | 0 | 0 |

Table 1: VAE model evaluation scores out of 10

| | | | |
|---|---|---|---|
| Perceived human-likeness | 1 | 0 | 0 |
| Appropriateness (timing)l | 2 | 0 | 1 |
| Appropriateness (content) | 0 | 0 | 0 |

Table 2: LSTM Model Evaluation scores out of 10

In the Tables 1 and 2 above, the ratings are given out of 10 for the following criteria. Perceived human-likeness implies if the motion seemed to follow human body-language to the evaluator. Appropriateness of timing implies if the gestures were in tune with silence and speech, that is, was the rendered person moving when they were speaking and were they more stable during silence. Appropriateness of content means if the gestures were in line with the content of the speech. Table 11 got a zero score overall and understandably so since the model was too erratic. Table 22 shows, although its score is low, it was definitely an improvement and was on the right path to generate gestures.

## 7 Conclusion

In conclusion, the VAE model that learns at a 10 second gesture clip and takes the whole scene into context failed to perform well, which indicates that it is not a valid strategy to work with VAEs coupled with convolution layers, an approach where it is treated as a sequence is much better as shown by previous submissions in the GENEA challenge. The LSTM model's performance shows that motions are smoother and more natural if the problem is treat as a sequence to sequence problem. However, its performance is not satisfactory enough, and the next step would definitely be to explore attention based transformers.

Therefore, the novel approaches failed to perform as hypothesised and did not even beat the baseline model or rules based model provided in the GENEA challenge. However, it shows which approaches to avoid while generating gestures from speech.

## References


[1] Uttaran Bhattacharya et al. "Text2Gestures: A Transformer-Based Network for Generating Emotive Body Gestures for Virtual Agents". In: *2021 IEEE Virtual Reality and 3D User Interfaces (VR)*. IEEE, Mar. 2021. DOI: 10.1109/vr50410.2021.00037. URL: https://doi.org/10.1109%2Fvr50410.2021.00037.

[2] Che-Jui Chang, Sen Zhang, and Mubbasir Kapadia. "The IVI Lab entry to the GENEA Challenge 2022 – A Tacotron2 Based Method for Co-Speech Gesture Generation With Locality-Constraint Attention Mechanism". In: *GENEA: Generation and Evaluation of Non-verbal Behaviour for Embodied Agents Challenge & Workshop 2022*. 2022. URL: https://openreview.net/forum?id=gMTaia--AB2.

[3] Taras Kucherenko et al. "Gesticulator: A framework for semantically-aware speech-driven gesture generation". In: *Proceedings of the ACM International Conference on Multimodal Interaction*. 2020.

[4] Youngwoo Yoon et al. "Speech gesture generation from the trimodal context of text, audio, and speaker identity". In: *ACM Transactions on Graphics* 39.6 (Nov. 2020), pp. 1–16. DOI: 10.1145/3414685.3417838. URL: https://doi.org/10.1145%2F3414685.3417838.

[5] Youngwoo Yoon et al. "The GENEA Challenge 2022: A large evaluation of data-driven co-speech gesture generation". In: *Proceedings of the ACM International Conference on Multimodal Interaction*. ICMI '22. ACM, 2022, pp. 736–747. DOI: 10.1145/3536221.3558058.