The GestureMaster entry to the GENEA Challenge 2022

ANONYMOUS AUTHOR(S)*

This paper describes the GestureMaster entry to the GENEA (Generation and Evaluation of Non-verbal Behaviour for Embodied Agents) Challenge 2022. Given speech audio and text transcriptions, GestureMaster can automatically generate a high-quality gesture sequence to accompany the input audio and text transcriptions in terms of style and rhythm. GestureMaster system is based on the recent ChoreoMaster publication[3]. ChoreoMaster can generate dance motion given a piece of music. We make some adjustments to ChoreoMaster to suit for the speech-driven gesture generation task. We are pleased to see that among the participating systems, our entry attained the highest median score in the human-likeness evaluation. In the appropriateness evaluation, we ranked first in upper-body study and second in full-body study.

CCS Concepts: • **Computing methodologies → Procedural animation**.

Additional Key Words and Phrases: Gesture Generation, Audio-driven Pose Estimation

ACM Reference Format:
Anonymous Author(s). 2022. The GestureMaster entry to the GENEA Challenge 2022. In . ACM, New York, NY, USA, 8 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Non-verbal-behaviour such as gestures are vital in human communication. Automatically generating high-quality gestures from audio and text transcription remains a challenging task. The GENEA Challenge 2022 [7] on speech-driven gesture generation aims to bring together researchers that use different methods for non-verbal-behaviour generation and evaluation.

Recent deep learning-based approaches like StyleGestures[1] have successfully been applied to synthesizing gesture poses. These methods grasp some deeper relationships between audios, text transcriptions and gestures than traditional techniques. However, these methods are limited by the representation power of proposed neural networks. Neural networks characterize data by projecting it into a low-dimensional latent space, while high-frequency motion details of gestures are considered to be noised and internally ignored. This lowers the quality of generated gestures, causing them to be "dull" and "blurred".

We have developed GestureMaster system. It is adjusted from recent music-to-dance system Choreomaster[3]. Given paired audio, text transcriptions and gestures, we first build a gesture database. This database consists of gesture phases split from training gesture motion by an automatically split algorithm. Then, we find a style embedding by StyleGestures-like network by mapping audio into a desired gesture feature, and a rhythm embedding for each phase of audio and gesture. The style embedding and rhythm embedding are then incorporated within a graph-based motion synthesis framework. It can generate high-quality gesture motions with high-level human-likeness and high appropriateness score for the associated held-out speech, in terms of timing or rhythm.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
© 2022 Association for Computing Machinery.
Manuscript submitted to ACM

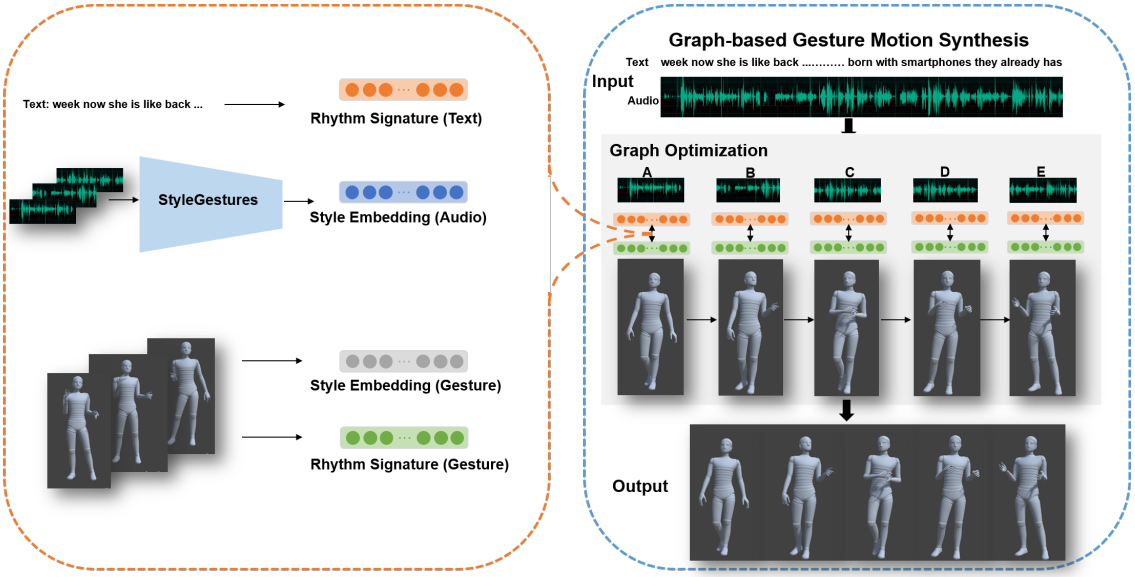


Fig. 1. Overview of our proposed system GestureMaster. Given an input of audio and text transcriptions, we split them into phases. For each audio phase, we calculate its rhythm signature (see Figure 2) and style embedding using StyleGestures. Then the graph-based gesture motion synthesis module searchs matching gesture motion nodes from database with lowest cost, in terms of rhythm, style and transition (see equation 2).

2 DATA PREPARATION

The dataset provided by challenge organizer is adapted from Talking With Hands 16.2M[5]. It comes from recording of dyadic interactions between different speakers. Each dyad has been separated into two independent sides with one speaker each. The training dataset inlcudes 293 recordings with an overall length of 18 hours. Each recording consists of audio, text transcripts and gesture motion.

First we segment each gesture motion in training dataset into phase-level gesture phases automatically by time interval of words larger than 0.4 seconds in text transcriptions. We manually remove gesture phases with low quality such as motion with jitter or wrong rotation. For motion capture without finger animation, we simply search and transfer the rotation from finger motion capture with lowest position distance. Then these gesture phases are semi-automatically annotated with rhythm signature and style signature for graph-based motion synthesis. Finally We mirror the gesture phases and build a gesture database with more than 6000 phases, range from 1 seconds to more than 10 seconds, determined by the length of each phase.

2.1 Rhythm Embedding

The term rhythm is often expressed in terms of beat. Beat corresponds to pulses of sound in audio, while gesture motion beat corresponds to pausing or sharp truning of gesture movements. The proposed rhythm signature consists of 32 bits in our system (see Figure 2). In each rhythm signature, bits denote the presence of beats (1 : present, 0 : not present) which correspond to the evenly-spaced beats indicated by the time signature. For rhythm signature of audio and text transcriptions, bits denote the presence of words. For rhythm signature of gesture motion, bits denote the presence of

pausing, sharp turning or stroke gesture. Obviously, a time of silence will result in a rhythm signature in which all bits are zeros. The distance between two rhythm signatures can be defined using Hamming distance, the number of bit positions in which the two bit patterns differ. Lower hamming distance indicates a better match of rhythm between audio and gesture motion.

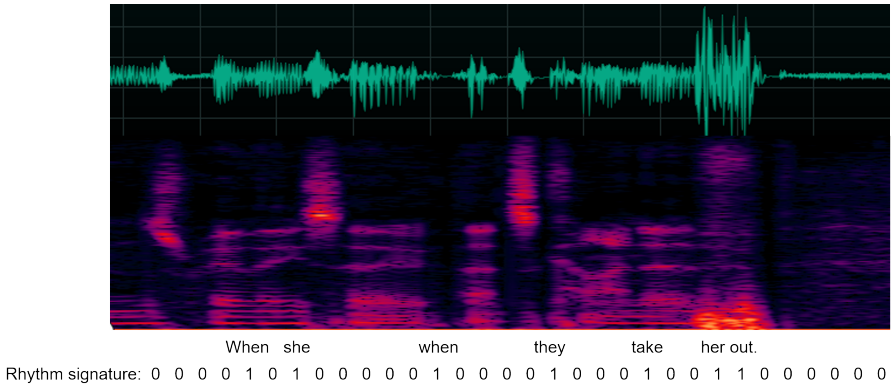


Fig. 2. Rhythm signature of audio examples. Bits denote the presence of words (1 : present, 0 : not present).

For audio, rhythm signature is annotated automatically using word-level timing information in text transcriptions (see Figure 2). For gestures, rhythm signature is annotated automatically using speed curve of two hands (see Figure 3).

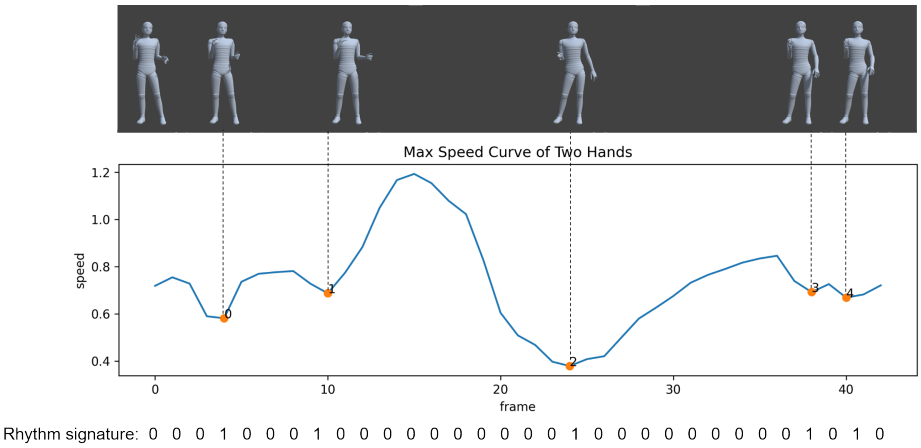


Fig. 3. Rhythm signature of gestures examples. Bits denote the presence of pausing, sharp turning or stroke gesture (1 : present, 0 : not present).

2.2 Style Embedding

StyleGestures is a probabilistic model which could generate gestures with different style, such as gesture speed, radius and height. We splice these features together as style embedding. We calculate mean speed, mean radius and mean

height of each gesture phase in database offline. And we adopt StyleGestures as a backbone of style embedding network and train the model on training dataset. In synthesis period, we could feed audio into StyleGestures to generate desired gesture and style embedding for graph-based optimization (see Figure 1).

3 SYSTEM OVERVIEW

In this section, we discuss the pipeline of proposed system (see Figure 1), including motion graph construction and graph-based optimization. We explain how the rhythm embedding and style embedding are incorporated into our graph-based motion synthesis framework.

3.1 Motion Graph Construction

A motion graph is a directed graph where each node denotes a motion segment in the database while each edge depicts the cost of transition between two adjacent nodes.

In our system, each node in our motion graph corresponds to a gesture phase. In our motion graph, the edge transition cost $T(D_p, D_q)$ between two nodes D_p and D_q is defined as:

$$T(D_p, D_q) = \lambda_1 T_p + \lambda_2 T_r \tag{1}$$

Where T_p, T_r is summed distance of positions, rotations between joints in transitional frames of two adjacent nodes, respectively. λ_1 and λ_2 are the corresponding weights. An edge is created in the graph if the transition cost between adjacent nodes is below a threshold δ_T. A higher δ_T results in more edges in the graph but may also cause artifacts as bad transition edges may also be included in the graph.

We build motion graph for upper body and lower body separately. For upper body motion graph, the style embedding vector and rhythm signature are also attached to each graph node.

3.2 Graph-based Optimization

In the graph-based framework, each synthesized motion corresponds to a path in the motion graph. In our system, gesture generation can be viewd as finding optimal paths. Given audio and text transcriptions, we first split transcriptions into several phases with time interval threshold 0.4 seconds and we obtain an audio sequence $M = \{M_i | i = 1, \ldots, n\}$, where M_i represents phase i of input audio. Then we calculate rhythm signature R_{M_i} of M_i (see Figure 2). The goal of our system is to assign a gesture motion node D_i in the motion graph to each M_i and to minimize the following cost:

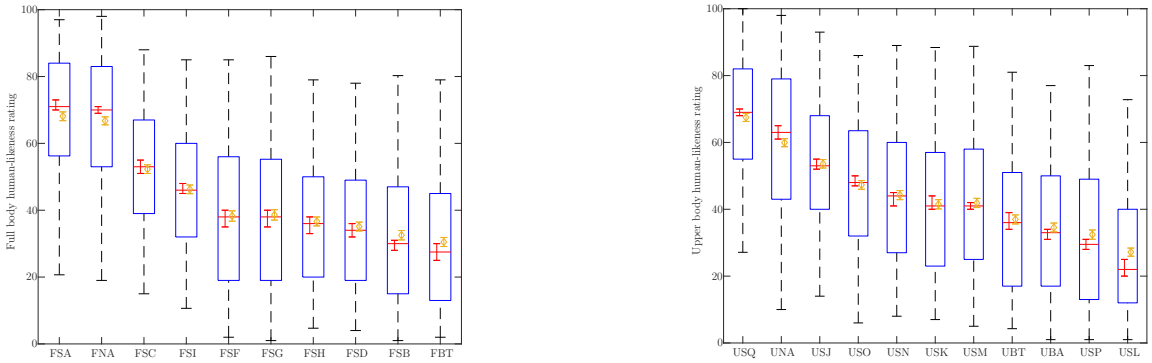
$$C = \lambda_3 \Sigma_{i=1}^{n} C_d(i) + \lambda_4 \Sigma_{i=1}^{n-1} C_t(i, i+1) \tag{2}$$

where C_d, C_t are the data term and transition term, respectively. λ_3, λ_4 are the corresponding weights.

Data term. $C_d(i)$ is the sum of rhythm embedding and style embedding matching cost between audio phase M_i and motion node D_i:

$$C_d(i) = \lambda_5 G_z(Z_{M_i}, Z_{D_i}) + \lambda_6 G_r(R_{M_i}, R_{D_i}) \tag{3}$$

where G_z, G_r are style embedding L2 distance and rhythm signature Hamming distance between audio and gestures. Style embedding of audio phase Z_{M_i} is calculated using StyleGestures while style embedding of gesture phase Z_{D_i} is calculated offline (see Section 2.2). λ_5, λ_6 are weights. Specifically, for optimization of lower body, we set $\lambda_5 = \lambda_6 = 0$ and data term $C_d(i) = 0$.



(a) Full-body study

(b) Upper-body study

Fig. 4. Box plots visualising the ratings distribution in the two studies. Red bars are the median ratings (each with a 0.05 confidence interval); yellow diamonds are mean ratings (also with a 0.05 confidence interval). Box edges are at 25 and 75 percentiles, while whiskers cover 95% of all ratings for each condition. Conditions are ordered descending by sample median for each tier.

Transition term. C_t ensures a smooth transition between adjacent motion phases in the synthesized motion.

$$C_t(i, i+1) = T(D_i, D_{i+1}) \tag{4}$$

The optimal gesture motion sequences are synthesized using a dynamic programming algorithm[2]. After generating upper body motion and lower body motion separately, we blend the two motion to create a full body motion. We smooth all synthesized gestures using Savitzky-Golay filter. For arm, hand and head joints, the length of the filter window and the order of the polynomial are 7 and 2 while for other joints are 7 and 1, respectively.

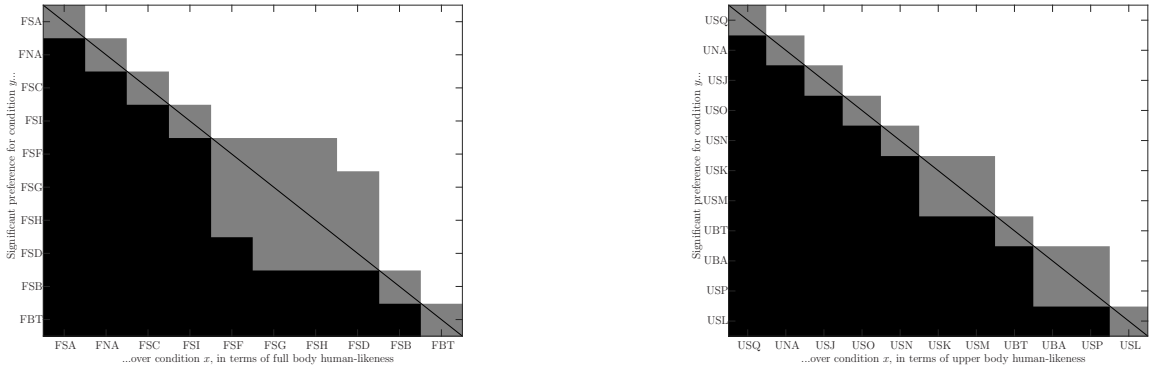
4 EVALUATION

For upper body graph, we set the hyperparameters to: $\lambda_1, ...\lambda_6 = 0.7, 0.3, 1.0, 3.0, 1.0, 0.1, \delta_T = 8, \zeta = 10000$. For lower body graph, we set the hyperparameters to: $\lambda_1, ...\lambda_6 = 0.7, 0.3, 0.0, 0.0, 0.0, 0.0, \delta_T = 8, \zeta = 0$. Our gesture synthesis system was tested on a desktop with a 3.70GHz i7-8700K CPU, 32GB RAM and a GTX 3070 GPU.

The evaluation of the submitted gesture motion will likely consider two aspects such as its perceived **human-likeness**, without accounting for the speech and its **appropriateness** for the associated held-out speech, in terms of timing and semantic content. Study participants were recruited through the crowdsourcing platform Prolific. The groundtruth natural motion was labelled **FNA** in the fullbody study and **UNA** in the upper-body study. Our condition ID in the upper-body evaluation was **USQ** and our condition ID in the full-body evaluation was **FSA**. The evaluations also included two baseline systems, one based on text-input only [6], and one based on audio-input only [4].

4.1 Human likeness study

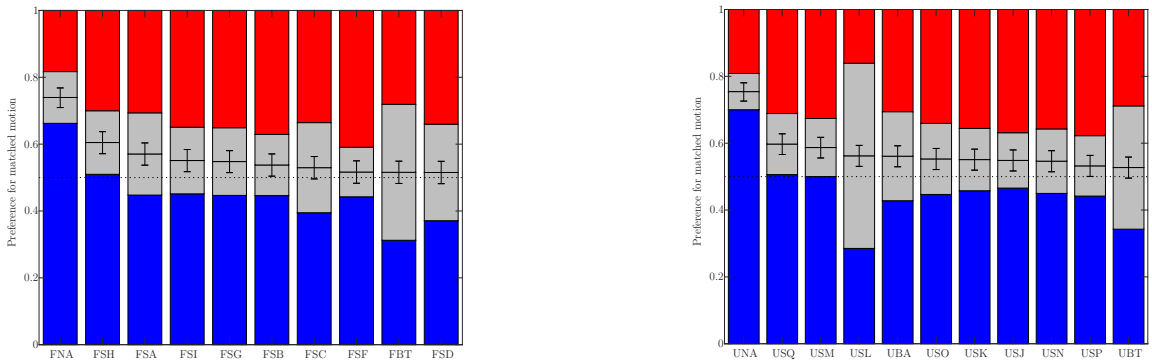
In human likeness study, study participants were asked "How human-likeness does the gesture motion appear?" then gave their ratings in response to this question on a scale from 0 (worst) to 100 (best). GestureMaster (**FSA**, **USQ**) ranked first and even above the groundtruth motion from the motion-capture recordings in both full-body and upper-body tiers. Bar plots and significance comparisons are shown in Figure 4 and Firure 5. Summary statistics (sample median and sample mean) for the ratings of all conditions in each of the two studies are shown in Table 1.



(a) Full-body study

(b) Upper-body study

Fig. 5. Significance of pairwise differences between conditions. White means that the condition listed on the x-axis rated significantly above the condition on the y-axis, black means the opposite (y rated below x), and grey means no statistically significant difference at the level $\alpha = 0.05$ after Holm-Bonferroni correction. Conditions are listed in the same order as in Figure 4, which is different for each of the two studies.



(a) Full-body study

(b) Upper-body study

Fig. 6. Bar plots visualising the response distribution in the appropriateness studies. The blue bar (bottom) represents responses where subjects preferred the matched motion, the light grey bar (middle) represents tied ("They are equal") responses, and the red bar (top) represents responses preferring mismatched motion, with the height of each bar being proportional to the fraction of responses in each category. The black horizontal line bisecting the light grey bar shows the proportion of matched responses after splitting ties, each with a 0.05 confidence interval. The dashed black line indicates chance-level performance. Conditions are ordered by descending preference for matched after splitting ties.

4.2 Appropriateness study

In appropriateness study, participants were given pair of videos – both from the same condition and thus having the same motion quality, but one matched to the speech and the other mismatched, coming from unrelated speech. Participants were then asked to pick the one video from the pair that best matched the speech. GestureMaster (**FSA**, **USQ**) ranked first in upper-body tier and second in full-body tier. Bar plots are shown in Figure 6.

ID	Human-likeness		Appropriateness			
			Number of responses			Percent matched
	Median	Mean	Match.	Equal	Mismatch.	(splitting ties)
FNA	70 $\in [69, 71]$	66.7 ± 1.2	590	138	163	$74.0 \in [70.9, 76.9]$
FBT	27.5 $\in [25, 30]$	30.5 ± 1.4	278	362	250	$51.6 \in [48.2, 55.0]$
FSA	71 $\in [70, 73]$	68.1 ± 1.4	393	216	269	$57.1 \in [53.7, 60.4]$
FSB	30 $\in [28, 31]$	32.5 ± 1.5	397	163	330	$53.8 \in [50.4, 57.1]$
FSC	53 $\in [51, 55]$	52.3 ± 1.4	347	237	295	$53.0 \in [49.5, 56.3]$
FSD	34 $\in [32, 36]$	35.1 ± 1.4	329	256	302	$51.5 \in [48.1, 54.9]$
FSF	38 $\in [35, 40]$	38.3 ± 1.6	388	130	359	$51.7 \in [48.2, 55.1]$
FSG	38 $\in [35, 40]$	38.6 ± 1.6	406	184	319	$54.8 \in [51.4, 58.1]$
FSH	36 $\in [33, 38]$	36.6 ± 1.4	445	166	262	$60.5 \in [57.1, 63.8]$
FSI	46 $\in [45, 48]$	46.2 ± 1.3	403	178	312	$55.1 \in [51.7, 58.4]$

(a) Full-body study

ID	Human-likeness		Appropriateness			
			Number of responses			Percent matched
	Median	Mean	Match.	Equal	Mismatch.	(splitting ties)
UNA	63 $\in [61, 65]$	59.9 ± 1.3	691	107	189	$75.4 \in [72.5, 78.1]$
UBA	33 $\in [31, 34]$	34.6 ± 1.4	424	264	303	$56.1 \in [52.9, 59.3]$
UBT	36 $\in [34, 39]$	37.0 ± 1.4	341	367	287	$52.7 \in [49.5, 55.9]$
USJ	53 $\in [52, 55]$	53.6 ± 1.3	461	164	365	$54.8 \in [51.6, 58.0]$
USK	41 $\in [40, 44]$	41.5 ± 1.4	454	185	353	$55.1 \in [51.9, 58.3]$
USL	22 $\in [20, 25]$	27.2 ± 1.3	282	548	159	$56.2 \in [53.0, 59.4]$
USM	41 $\in [40, 42]$	41.9 ± 1.4	503	175	328	$58.7 \in [55.5, 61.8]$
USN	44 $\in [41, 45]$	44.2 ± 1.4	503	175	328	$58.7 \in [55.5, 61.8]$
USO	48 $\in [47, 50]$	47.3 ± 1.4	439	209	335	$55.3 \in [52.1, 58.5]$
USP	29.5 $\in [28, 31]$	32.4 ± 1.4	440	180	376	$53.2 \in [50.0, 56.4]$
USQ	69 $\in [68, 70]$	67.5 ± 1.2	504	182	310	$59.7 \in [56.6, 62.9]$

(b) Upper-body study

Table 1. Summary statistics of user-study ratings from all user studies, with confidence intervals at the level $\alpha = 0.05$. "Percent matched" identifies how often participants preferred matched over mismatched motion in terms of appropriateness.

5 CONCLUSION

We have proposed GestureMaster, a graph-based gestures synthesis system. We build a gesture database including more than 6000 gesture phases with style embedding and rhythm embedding. Given audio and text transcriptions, a graph-based optimization is adopted to generate high-quality gesture motion. The evaluation results demonstrate that GestureMaster can synthesize gestures with high human-likeness score as well as high appropriateness score for associated speech in terms of rhythm.

There is a gap between GestureMaster and ground truth motion in the appropriateness study. In future research, a better rhythm embedding could be used for better rhythm matching. Because of imbalanced data, we do not evaluate the appropriateness for the individual gesticulation style of the indicated test speaker in each segment. Howerer, GestureMaster could simply generate gestures for each speaker by building different graphs for different speakers, indicating the potential of GestureMaster to generate individual-related gestures.

REFERENCES

[1] Simon Alexanderson, Gustav Eje Henter, Taras Kucherenko, and Jonas Beskow. 2020. Style-controllable speech-driven gesture synthesis using normalising flows. *Computer Graphics Forum* 39, 2 (2020), 487–496. https://doi.org/10.1111/cgf.13946

[2] G David Forney. 1973. The viterbi algorithm. *Proc. IEEE* 61, 3 (1973), 268–278.

[3] Chen Kang, Zhipeng Tan, Jin Lei, Song-Hai Zhang, Yuan-Chen Guo, Weidong Zhang, and Shi-Min Hu. 2021. ChoreoMaster : Choreography-Oriented Music-Driven Dance Synthesis. *ACM Transactions on Graphics (TOG)* 40, 4 (2021).

[4] Taras Kucherenko, Dai Hasegawa, Gustav Eje Henter, Naoshi Kaneko, and Hedvig Kjellström. 2019. Analyzing input and output representations for speech-driven gesture generation. In *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*. 97–104.

[5] Gilwoo Lee, Zhiwei Deng, Shugao Ma, Takaaki Shiratori, Siddhartha S Srinivasa, and Yaser Sheikh. 2019. Talking with hands 16.2 m: A large-scale dataset of synchronized body-finger motion and audio for conversational motion analysis and synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 763–772.

[6] Youngwoo Yoon, Woo-Ri Ko, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee. 2019. Robots learn social skills: End-to-end learning of co-speech gesture generation for humanoid robots. In *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 4303–4309.

[7] Youngwoo Yoon, Pieter Wolfert, Taras Kucherenko, Carla Viegas, Teodor Nikolov, Mihail Tsakov, and Gustav Eje Henter. 2022. 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.