Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

PREPRINT

SPECIAL ISSUE PAPER

ExpressGesture: Expressive Gesture Generation from Speech through Database Matching

Ylva Ferstl*[1] | Michael Neff[2] | Rachel McDonnell[3]

[1]Trinity College Dublin, Ireland
[2]University of California, Davis, United States
[3]Trinity College Dublin, Ireland

Correspondence
*Ylva Ferstl, School of Computer Science and Statistics, Trinity College Dublin, Dublin 2, Ireland.
Email: yferstl@tcd.ie

Summary

Co-speech gestures are a vital ingredient in making virtual agents more human-like and engaging. Automatically generated gestures based on speech-input often lack realistic and defined gesture form. We present a database-driven approach guaranteeing defined gesture form. We built a large corpus of over 23,000 motion-captured co-speech gestures and select individual gestures based on expressive gesture characteristics that can be estimated from speech audio. The expressive parameters are gesture velocity and acceleration, gesture size, arm swivel, and finger extension. Individual, parameter-matched gestures are then combined into animated sequences. We evaluate our gesture generation system in two perceptual studies. The first study compares our method to the ground truth gestures as well as mismatched gestures. The second study compares our method to five current generative machine learning models. Our method outperformed mismatched gesture selection in the first study and showed competitive performance in the second.

KEYWORDS:
gesture generation, expressive agents, conversational agents, computer animation, perception, motion matching

Co-speech gestures are a vital ingredient in making virtual agents more human-like and engaging. Automatically generated gestures based on speech-input often lack realistic and defined gesture form. We present a database-driven approach guaranteeing defined gesture form. We built a large corpus of over 23,000 motion-captured co-speech gestures and select individual gestures based on expressive gesture characteristics that can be estimated from speech audio. The expressive parameters are gesture velocity and acceleration, gesture size, arm swivel, and finger extension. Individual, parameter-matched gestures are then combined into animated sequences. We evaluate our gesture generation system in two perceptual studies. The first study compares our method to the ground truth gestures as well as mismatched gestures. The second study compares our method to five current generative machine learning models. Our method outperformed mismatched gesture selection in the first study and showed competitive performance in the second.

1 | INTRODUCTION

Much previous research has explored methods for automatically generating gesture motion from speech or text input. Solely relying on speech prosody is attractive as this is a readily obtainable and automatically analyzable input signal. Speech-based generative models attempt to capture the prosodic variation in the input speech and its relationship to the gesture motion. However, a common result is averaged motion lacking realistic and defined gesture form. One problem driving this may be the modelling of gestures by means of exact joint positions or angles. Natural gesture behavior is variable, with many different gesture motions co-occurring with the same or similar utterance(s) at different times or across speakers. Joint position or angle representations of gesture impose a strong constraint for a generative model; even when a valid-looking, novel gesture is produced it may be penalized heavily during training if it is numerically far from the ground truth pose sequence. This can result in lethargic motion close to the mean pose that lacks definition.

Instead of modelling an implicit relationship of exact, high–dimensional joint poses and the speech signal, we use a higher level representation of the gesture motion through expressive parameters shown in previous work to be associated with the speech signal as well as being perceptually important to the quality of the speech-gesture match[1]. These gesture parameters are gesture velocity, acceleration, size, arm swivel angle and extent of hand opening.

We propose a full end-to-end speech-to-motion pipeline. Our method uses a merged approach of machine learning and database sampling to produce realistic gesture form, building on recent preliminary work by Ferstl et al.[2]. Offline, a machine learned model first establishes the speech-gesture relationship by encoding the relationship between acoustic speech features and expressive gesture parameters. Online, the speech signal is then used to automatically extract gesture timing; given the speech signal and gesture timing as input, the model estimates gesture parameters that are then used to search a gesture database for the gestures that best match the predicted expressive parameters. We evaluate the method with two perceptual studies, including a comparison against the state-of-the-art machine learning approaches.

Our final contribution is making our dataset of 6 hours openly available, which we believe is the largest natural conversational dataset of synchronized motion capture and speech recordings to be published open-source, as well as our gesture-by-gesture segmentation of 10 hours (~23,700 gestures) of conversational data.

2 | RELATED WORK

Efforts in automatic gesture generation can largely be divided into three groups; rule based, statistical models, and generative machine learning models. Rule based approaches use explicit phrase-to-gesture mappings[3–5], strongly limiting the amount of gesture variety. Statistical models use estimated conditional probabilities of specific speech features co-occurring with a set of motion features and can produce realistic gesture form as gesture sequences are built from parts of true motion data. For example, Neff et al.[6] annotated video corpora with gestural lexemes and semantic tags, and generate new gesture sequences procedurally generating the likely lexemes given a set of input semantics. Bergmann and Kopp[7] compute likelihoods of more detailed, hand-coded gesture descriptors, such as handshape and movement direction. Fernández-Baena et al.[8] annotate 6 minutes of performed beat gestures (gestures without a specific meaning) to analyze speech-gesture correlations of synchrony and intensity. A motion graph is used for gesture synthesis, searching for the gesture with the best transition, intensity, and timing match. Recent work proposed a motion graph approach for dyadic conversation behavior. Yang et al.[9] annotate timings of 30 minutes of speech and gesture. Gesture is synthesized by finding a path through the graph given a set of associated speech constraints. Motion graphs have limited expressivity and require significant computational power for large state spaces.

Machine learning models for gesture generation, specifically neural networks, can model large datasets of unstructured data, implicitly learning relationships between high-dimensional speech input and target motion output. While they can generate new, previously unseen motion, in practice, the output motion is often a poor imitation of the variety of the training data. The standard training paradigm using a mean squared error loss (as in[10–12]) tends to produce overly smooth and averaged motion. Alternatives have been proposed with generative adversarial networks[13, 14] as well as probabilistic generation through normalizing flows[15].

One problem for machine learning approaches is the nondeterministic relationship of speech and gesture; a plethora of valid gestures can and do occur for a given utterance, hence modelling co-speech gestures as exact joint positions or angles may be too constrictive to capture natural variety. Recent work instead proposes modelling the speech-gesture relation with higher-level parameters; Ferstl et al.[1] propose a set expressive gesture parameters, namely velocity, initial acceleration, gesture size, arm swivel, and hand opening, and show they can both be estimated from speech as well as perceptually impacting speech-gesture match. In a preliminary study, replacing ground truth gestures with parameter-matched gestures of similar length was shown to outperform unmatched gestures[2]. The approach, however, relied on extracting gesture timing and duration from the motion signal, making it impractical for real applications. In this work, we extend this approach and remove this need for motion analysis by designing a speech-based gesture phase extraction.

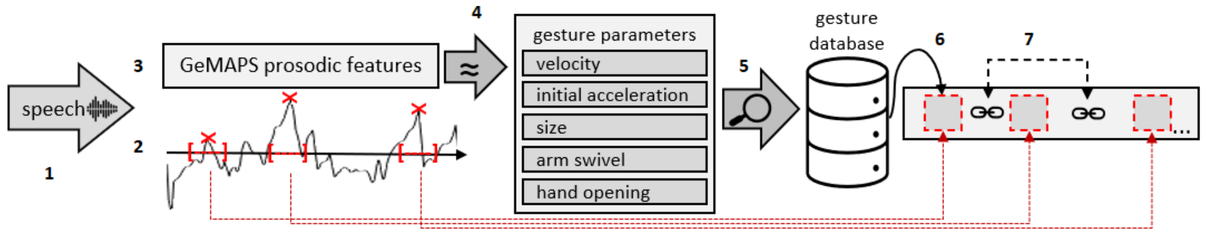

FIGURE 1 Overview of our gesture generation system. (1) The system receives as input the speech audio for a gesture segment. (2) Gesture timings are predicted by analyzing speech pitch peaks. (3) GeMAPS prosodic speech features are extracted automatically. (4) The desired values for the 5 gesture parameters are estimated from the GeMAPS. (5) The database is searched for the gesture with the closest matching parameter values. (6) The best matching gesture is inserted at the desired gesture position. (7) Synthetic preparation and retraction phases are generated to link the gestures in the sequence.

Our proposed method relates to Levine et al.[16], who learn a hidden state representation of a small motion database based on 6 kinematic parameters. A Conditional Random Field then maximizes the probability of the hidden state distribution given prosodic input speech features. An animation sequence is generated by finding a lowest-cost path given the input signal. Similarly, Stone et al.[17] annotate both speech and motion features (such as points of speech prominence and gesture content) of a small dataset and jointly synthesize both modalities by recombining speech and motion segments.

Our approach further relates to the Motion Matching method[18], which draws animations from a database based on a set of specified motion properties, such positions of the end effectors, combining them with simple blending and inverse kinematics. The method became popular with gaming studios due to its reliable motion quality and its suitability for real-time animation through the use of efficient search algorithms. Memory usage for larger datasets can be a problem; by combining Motion Matching with neural network controllers, including compression of motion data to a low-dimensional representation, memory usage was reduced significantly by Holden et al.[19]. A number of other works have proposed methods for motion database retrieval based on sets of motion parameter keys (e.g.[20, 21]).

Relating to both statistical and machine-learning models, our method aims to combine the power of neural networks to capture the relationship of speech features with higher level motion features with the advantage of realistic gesture motion, sampled from a large database.

3 | GESTURE MATCHING METHOD

Briefly, our method of gesture generation consists of four steps: (1) Determining gesture timings through prosodic processing, (2) estimating gesture parameters from speech, (3) selecting appropriate gestures from the database, and (4) combining the selected gestures into a coherent animation sequence. We detail step 1 in Sec. 3.2 below, followed by steps 2 and 3 in Sec. 3.3. An illustrative overview is presented in Fig. 1.

3.1 | Gesture database

At the heart of our gesture generation system is our database of over 23,000 individual gestures, automatically tagged with their five respective expressive parameter values. The database represents gestures extracted from over 10 hours of motion captured data, from two multimodal datasets of speech and gesture. Dataset A has previously been used in Ferstl et al.[14] but has not been released until now. Dataset B is the open-source Trinity Speech-Gesture dataset[10]. Each dataset contains recordings of a different male native English speaker, producing natural speech with spontaneous gesture motion. The two speakers exhibit a distinctly different gesture style. Motion was captured with a 59 marker setup and 20 Vicon cameras at 120 Frames per Second (fps) in dataset A, and at 59.94 fps in dataset B.

We segment motion data into individual stroke phases, the expressive phase of a gesture[22]. 4 hours of dataset A have previously been hand-annotated, the rest is automatically annotated using a classifier taking as input the motion data as well as the speech pitch, a method presented in Ferstl et al.[14]. Stroke labelling results in close to 23,700 gesture strokes for our gesture database.

Accompanying this work, we will release our full database of gestures, together with their values for the five expressive parameters, and corresponding speech.[1] We also release the 5 trained speech-to-parameter models for reproducibility (step 4 in Fig. 1). We think this database will be a valuable resource for gesture research. The segmentation of continuous motion into individual gesture strokes also allows for easier integration with a number of existing systems, such as Motion Matching. To our knowledge, this will be the largest released dataset of synchronized motion capture and speech recordings; with this we hope to make a key contribution to future gesture generation system development.

3.2 | Gesture timing from speech

To determine when to generate a gesture, we analyze the speech pitch. We first extract the pitch contour using Praat. From this, we determine pitch peaks by setting the desired local prominence[2] (set to 0.5) and minimum peak distance (set to 1.1 seconds). These values can be tuned for the desired gesture frequency. As we can determine the actual gesture frequency for the used datasets through motion analysis, we set the pitch sensitivity to result in a similar gesture frequency. Both speakers (Dataset A and B) presents a gesture frequency averaging approximately one gesture every 1.5 seconds. Using the true gesture frequency allows for better comparison to ground truth gesture performance in this study. For other applications, this can be tuned as desired, for example, a higher gesture frequency can be used to increase perceptions of extraversion for a robot[23].

Following the rule that gesture peaks either precede or coincide with the associated speech peak[24], we set gesture timing such that gesture strokes are 55% complete at the pitch peak. We define the maximum time for a predicted stroke as twice the time between this and the nearest other peak. This window defines the speech segment to be analyzed in the next step (Sec. 3.3) that estimates gesture parameters. If a selected gesture is shorter than the time window, we re-align the gesture forward to again be 55% complete at the pitch peak. We choose this gesture timing approach for its simplicity and ease of reproducibility, however, our system may be combined with any other method of determining stroke timing.

3.3 | Synthesizing a gesture sequence

The gesture timing (Sec. 3.2) provides a sequence of empty motion slots with associated speech data, each to be filled with a gesture stroke from the database. The first step in selecting a gesture computing a set of desired gesture parameters. Theoretically, we can use any parameter automatically computable from a motion segment, but choose the five gesture parameters that were shown previously to be associated with the speech signal as well as impacting perceptions of the quality of the speech-gesture match[1]: (1) Gesture velocity, (2) the mean acceleration to the first major velocity peak, (3) gesture size measured by the total path length, (4) arm swivel (bringing the elbow closer or further from the body), and (5) hand opening (calculated as the mean distance of the finger tips from the base of the hand). The 5 parameters are estimated from the GeMAPS prosodic speech features[25], computed automatically using openSMILE[26], a process described in detail in Ferstl et al.[1].

Given the gesture parameter values, the database is searched for the best match, as determined by a match rank. A gesture's match rank for one parameter is given by the relative similarity to the desired parameter value. For example, the gesture with the most similar velocity receives velocity rank 1, and the gesture with the least similar velocity receives rank 23,700. Each gesture receives 5 rank values, one for each parameter. Rank values are weighted based on the findings of Ferstl et al.[1] of how well a parameter can be predicted from speech and its perceptual importance for speech-gesture match. For example, gesture swivel and size are predicted well from speech, and acceleration is predicted better than velocity; hand shape has a strong perceptual impact on the perceived gesture match[1]. We define the following parameter weights: $weight_{velocity} = 0.6$, $weight_{acceleration} = 0.8$, $weight_{size} = 1.1$, $weight_{swivel} = 1.3$, $weight_{hand} = 1$. The 5 weighted ranks are combined into a total match rank.

The complete gesture synthesis process is visualized in Fig. 1, the gesture matching algorithm is given below in pseudo-code:

```
Offline, before first use:
    for each gesture g in gesture_database:
        for each parameter p in gesture_parameters:
            calculate p(g)
Online, to synthesize a new sequence:
    for each gesture slot s in gesture_sequence:
        return max(rowsum(W*(P(s)-P(G))) )
```

[1] Database release: https://trinityspeechgesture.scss.tcd.ie/

[2] mathworks.com/help/signal/ug/prominence.html

Where P(G) denotes the array containing all parameter values of all gestures G in the database, W is the array of the 5 parameter weights, and P(s) are the estimated parameters of a speech segment s. All parameter values are normalized to the range of 0-1. The online computation component has a time complexity of $O(s * (3np + n))$, with s as the number of gesture slots in an input speech segment, n the number of gestures in the database, and p the number of gesture parameters.

To improve the smoothness of gesture transitions, we return the top 10 gestures in the match rank algorithm and choose a gesture that allows a reasonable transition by calculating the distances of candidates' starting wrist positions from the previous gesture's end positions ($O(s)$). Taking into account the time available for transition, we select a gesture resulting in realistic transition speed. We constrain gesture selection to gestures with a maximum duration of the speech window defined in the previous Sec. 3.2. In our experiments, for 470 gesture slots (corresponding to about 12 minutes of speech), 455 unique gestures were selected, indicating low repetitiveness, but if desired, gesture selection could be further restricted to unused gestures.

Selected gestures are combined using software based on the open-source animation environment DANCE[27], taking as input motion data and corresponding stroke labels and synthesizing preparations and retractions for the strokes using splines. The preparation brings the hands into position for the stroke, and the retraction returns the hands to a rest position. Preparation and retraction are proportionally matched to the stroke speed. If not enough time is available for a retraction before the next gesture, a transitional preparation is synthesized instead. Fig. 2 illustrates such a synthesized sequence.

4 | EVALUATION

We evaluate the performance of our gesture generation method with two perceptual studies. In the first study, we compare our method of gesture placement and selection to randomized gestures as well as to the ground truth placement and selection. In our second study, we compare our method to state-of-the-art machine learning models, namely the five entries of the recent GENEA gesture generation challenge[28].

We rendered gesture sequences on the GENEA model using Unity3D. Participants first read the study instructions and completed training (detailed in the respective sections below). Following this, each experiment trial consisted of watching a 15 second video clip followed by the question, "How appropriate were the gestures for the speech?", presented with a 7-point Likert scale ranging from "Very bad match" (1) to "Very good match" (7). Phrasing of the rating question was taken from the GENEA challenge.

Study completion time was approximately 15 minutes. Participants were recruited via Prolific; fluency in English was required. Attention was assessed through content questions: At a random trial number within each quartile of the study, a multiple choice question about what the speaker said was presented instead of the gesture rating question. 9 participants (out of a total of 64) with less than 50% correct were rejected.

4.1 | Baseline evaluation

Baseline evaluation consisted of three gesture conditions: (1) Ground truth gesture strokes with synthesized preparations, retractions, and transitions, preserving gesture timing (GT-S), (2) random gesture selection with the same overall gesture frequency but not timed to the speech, (3) our method of gesture placement and selection.

In addition to the arm gesture motion, the character displayed some idle torso body swaying and head movements.

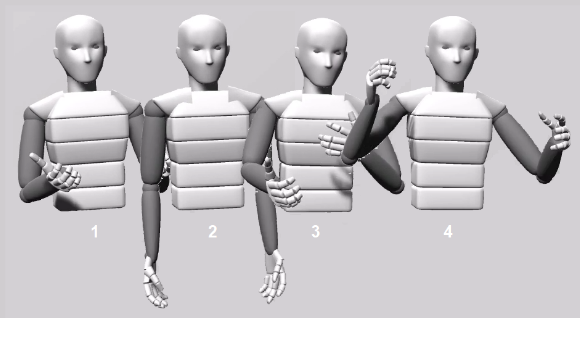

FIGURE 2 Example of a generated gesture sequence. Between gestures (1) and (3), the arms and hands are moved to/from a rest position through spline interpolation (2). If not enough time is available between gestures for a rest pose, a trajectory is instead created from a stroke end position (3) to the next stroke's start position (4).

Participants first completed a guided training. First, participants were informed that they will watch 2 examples of well-matched speech and gesture, and they were instructed to rate these as such. They were then presented with two ground truth training trials (one for each speaker). Next, participants were informed they would see 2 examples of badly matched speech and gesture, and were instructed to rate these accordingly. They were then presented with two unmatched (random) training clips (one for each speaker). Finally, participants were informed about attention checks and presented with one example.

There were 30 experiment trials (5 clips each for the 2 speakers, for the 3 conditions), presented in random order. Each clip contained a different speech segment. For each participant and condition, the 5 clips for each speaker were selected randomly from a pool of 10-15 clips in order to get a representative sample of gesture sequences while minimizing participant fatigue.

A total of 25 participants completed the experiment (12 females, aged 18-61 years, $M = 29.9$, $SD = 10.1$), all of whom gave informed consent regarding their participation. Participants represented a wide population sample: 13 different residence countries were reported.

All stimuli can be viewed at https://youtube.com/playlist?list=PLLrShDUC_FZx8bbo4SYiQWVsQqDylWf7O

4.1.1 | Results

For statistical analysis of the results of the perceptual experiment, Likert rating scores were treated as ordinal data and a cumulative link model was fitted using clm from the R ordinal package. A one-way repeated measures ANOVA of the estimated model showed a main effect of condition ($p < .001$), with an effect size measured by Wald Chi Square $\chi^2 = 133.0$. The ground truth gesture condition was rated significantly higher than both other conditions (both $p < .001$), as expected with a mean rating score of 5.20. Gesture sequences generated with our method were rated significantly higher than random gesture sequences ($p < .01$), with a mean rating score of 3.94 (random: $mean = 3.58$). Ratings are visualized in Fig. 3.

4.1.2 | Discussion

Our method showed better performance than the mismatched *random* condition. This is notable as such a baseline is notoriously hard to beat for automatic gesture generation (see GENEA[28], as well as Sec. 4.2 below).

While directly comparing our method to the random condition conflates effects of gesture selection and of timing, preliminary work has compared two baselines, one preserving gesture timing but disregarding expressive parameter match, and one

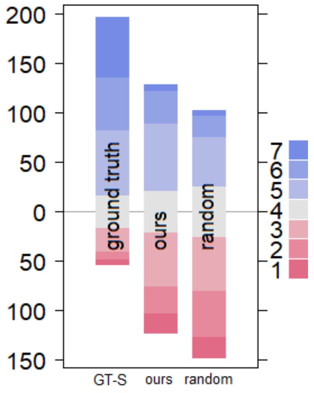

FIGURE 3 Frequency of perceptual rating scores in the baseline evaluation.

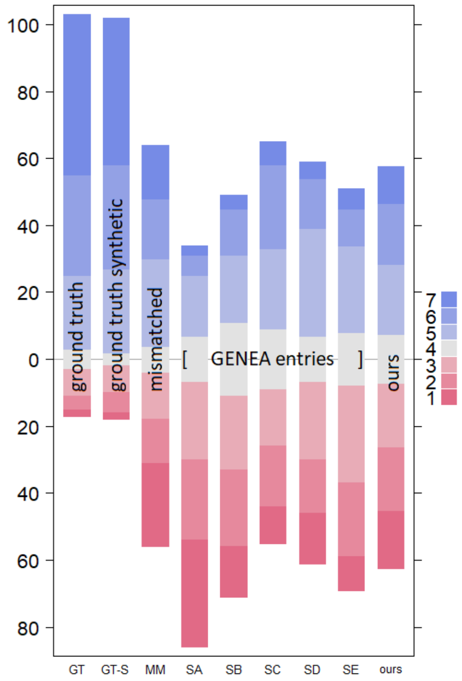

FIGURE 4 Frequency of perceptual rating scores for the 9 conditions in the comparative performance evaluation.

disregarding both timing and match. Preserving gesture timing showed no advantage, suggesting that timing alone is not enough to outperform the random condition[2]. We chose the animated character to match that of the large scale gesture evaluation study GENEA[28] and this character has a somewhat robotic appearance, however, the discussed preliminary work[2] has also evaluated the proposed method on the more realistic Brad character of the VHTK toolkit[29].

4.2 | Comparative performance evaluation

The comparative performance evaluation consisted of 9 gesture conditions: (1) Ground truth motion-captured motion (GT), (2) ground truth gesture strokes with synthesized preparations, retractions, and transitions (GT-S), (3) mismatched motion, belonging to another speech segment (MM), (4-8) motion generated by the GENEA gesture generation challenge entries (SA-SE) (9) our method. For conditions GT-S as well as for our method, the character displayed some idle torso body swaying and head movements in addition to the arm gesture motions. All other conditions already contained motion data for these joints.

Participants completed a guided training before starting the experiment. First, they were informed they would see an example of a very good speech and gesture match and instructed to rate it as such. They were then presented with a ground truth (GT) example trial. Next, they were informed they would see an example of badly matching speech and gesture and were instructed to rate it as such. An example of mismatched motion (MM) was presented. Participants were then informed that some motions may appear more synthetic while still matching the speech well, followed by an example of ground truth gestures with synthetic transitions (GT-S). Next, participants were informed that synthetic motions may show a very bad gesture-speech match, followed by an example clip of mismatched gestures with synthetic transitions.

There were 36 experiment trials (4 clips each for the 9 conditions), presented in random order. For each participant and condition, the 4 clips were selected randomly from a large pool of clips, ensuring adequate representation of the variation for each condition while maintaining a reasonable experiment duration and hence minimizing participant fatigue. Within participant, each clip contained a different speech segment. The GENEA challenge included only dataset B, so all experiment stimuli were therefore restricted to speech from this dataset.

30 participants from 10 different residence countries completed the experiment (15 females, ages 18-43 years, $M = 26.5$, $SD = 6.9$), all of whom gave informed consent regarding their participation.

All stimuli are at https://youtube.com/playlist?list=PLLrShDUC_FZyNS6-Wd7Yba-1wTgRAbqCj

4.2.1 | Results

An one-way repeated measures ANOVA of the estimated model showed a main effect of condition ($p < .001$), with an effect size measured by Wald Chi Square $\chi^2 = 234.8$. Both GT and GT-S were rated significantly higher than all other conditions (all $p < .001$), interestingly, there was no significant difference between the two. SA was rated significantly lower than all other conditions (all $p < .001$, except $p < .01$ w.r.t. SB) and SC was rated significantly higher than SB ($p < .05$). This is comparable to the results reported in the GENEA challenge[28]. No other differences were significant. Ratings are plotted in Fig. 4.

4.2.2 | Discussion

The results of the second evaluation show that our method produces competitive results to state-of-the-art machine learning models with respect to perceived appropriateness. Notably, while our method can be distinguished fairly easily from other generated motion through its procedural transitions between gestures, it was rated on-par or superior with the compared generative approaches. Indeed, for ground truth gestures too, synthetic transitions were rated to match the speech equally well as the full motion capture, indicating that matching individual gestures to the speech produces valid results even when this changes the motion style of the speaker.

The speaker of dataset B, used in this evaluation, shows a very animated speaking style, engaging his whole body rather than producing isolated arm gesture motion. Due to the our approach of focusing on generating arm gesture motion with only some auxiliary idle animation for the torso and head, we produce motion closer matches to the speaking style in dataset A, in which the speaker retains a fairly firm stance and produces more isolated arm gesture motion. We therefore think our method produced particularly suitable results for speaker A. However, due to the fact that the GENEA models[28] were not trained on this dataset, we cannot compare this performance.

While we did not compare these conditions directly, the MM condition in this evaluation appears to perform better than the random condition in Sec. 4.1. Notably, these conditions differ in that MM represents mismatched full motion-capture, whereas the random condition uses synthetic transitions between gesture strokes. The continuous, fluid motion in MM may appear less obviously mismatched to the speech.

5 | DISCUSSION & FUTURE WORK

We propose a full end-to-end system for gesture generation from speech audio through a novel method guaranteeing realistic gesture form. Previous works on machine learned models for gesture generation from speech often produce motion smoother than natural with poorly defined gesture form, as well as relying on an assumed and implicit speech-gesture relationship. In this work, we generate gestures based on expressive gesture parameters shown to be related to the speech prosody and to impact speech-gesture match. By using a parameterized motion representation in lieu of exact joint configurations, we aim to minimize the overfitting to one speaker's data. By selecting appropriate gestures from a motion-captured database, we always produce natural and well defined gesture form.

We evaluated our gesture generation method with two perceptual studies. First, we compare our method to a baseline method selecting random gestures at the same frequency but agnostic to speech emphasis, as well as to the ground truth gestures. Our method of generating gesture sequences for speech proved to outperform un-matched gesture animation. In our second evaluation, we compare our method to five current machine learning models, as well as to ground truth motion capture, mismatched motion capture, and ground truth gestures with synthetic transitions. The results show that our method is comparative in performance to the best of the tested generative models.

Most machine learning approaches, including the compared, generate continuous motion. Due to our gesture-by-gesture synthesis approach, our method may be easier to integrate into existing state-based frameworks used by game developers. Furthermore, by generating and linking individual gestures, our method allows modifying the expression of individual gestures. For example, a recent proposed framework for personality expression of virtual agents modifies the the Laban Effort and Shape parameters of individual gestures given just their start and end frame (which are known variables within our system)[30]. Our method also allows for tuning of gesture frequency, which can be used to modulate perceptions of extraversion[23, 31].

We release 6 hours of natural conversational data with high-quality motion-capture and speech recordings of an English speaker, to our knowledge the largest open-source dataset of its kind. We also release our full database of gestures, segmented from 10 hours of motion-capture, a total of almost 23,700 samples, as a major resource for gesture research.

Our segmentation of gesture motion into individual gestures may also be useful in developing semantic-aware systems. Wrist trajectories of individual gestures may be analyzed to determine simple gesture shapes, such as wiping. Combined with lexical analysis, such as negation tagging, we are interested in exploring integration of semantically meaningful gesture parameters. The current lack of lexical matching is also one potential reason for disliked gestures within generated sequences. Many of the gestures in the database are iconic gestures: gestures visualizing physical properties and describing the semantic content of the verbalisation. When searching the database for a matching gesture, we only take into account qualitative measures of the gesture (the five gesture parameters), without considering semantic content. Therefore, we sometimes find a gesture match that produces a semantic mismatch with the speech. This is different from many machine learning gesture generation models which largely focus on generating beat gestures, gestures without specific meaning but linked to the rhythm and pace of the speech.

Different techniques may be used to combine individual motion segments for improved between-gesture motion. Motion graphs have been employed for dyadic conversation gestures[9], however, building a motion graph for very large datasets can potentially become problematic. Yang et al.[9] note that the motion variety of conversational gesture requires a much larger graph than was previously used for e.g. locomotion. While they build a graph from 30 minutes of motion data, our work uses 20 times that amount. Constructing and searching a motion graph in our case would require questionable computing power. We instead opted for a more flexible method of motion blending and transition poses, only requiring storage of each gesture's stroke motion data, but theoretically our method can be combined with any other technique of combining motion segments and in future work we would like to improve the realism of the between-gesture motion.

Our gesture database can easily be extended with new motion without associated speech, such as the released 20 hours of the *Talking With Hands 16.2M* conversational dataset. This only requires automatic stroke segmentation as utilized here, and automatic tagging of the gesture parameters. In this regard, our method differs from Yang et al.[9], who base motion selection on the associated audio segment and therefore require every addition to the database to contain synchronized speech audio.

Additional gesture parameters can be used for improved gesture selection. Any automatically extractable motion measure would be readily integratable in our system. To improve the speech-gesture match, a relationship has to be established between the speech prosody and the new gesture parameter. If the parameter is not computable from speech, it may still be used for biased gesture selection after determining a number of suitable gestures from speech-based parameters, for example in order to achieve a specific gesture style.

Gesture selection can be tuned by scaling gesture parameters. This could be used for creating gesture behavior specific to a personality, for example, for an extroverted speaker, predicted gesture size can be scaled up in order to retain predicted size variation but creating gesture sequences with overall larger gestures. In this study we do not adjust for speaker style or personality other than what may be implicitly expressed through the five gesture parameters. For generating animation for one speaker, we allowed gesture selection from either speaker, resulting in a mix of gestures from both speakers within a sequence. This could potentially create style-mismatches both between gestures, and between gesture and speech. Speaker-specific gesture retrieval is possible with the downside of reducing the amount of available gestures.

Our method works for offline gesture synthesis; real-time gesture selection is difficult due to the gesture-before-speech rule in natural speech. Our results are limited to two speakers due to data availability; in future work we hope to be able to include more speakers and validate gesture parameter suitability across additional speakers.[3]

References

1. Ferstl Y, Neff M, and McDonnell R. Understanding the predictability of gesture parameters from speech and their perceptual importance. In: Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents; 2020. p. 1–8.
2. Ferstl Y, Neff M, and McDonnell R. It's A Match! Gesture Generation Using Expressive Parameter Matching. In: Proceedings of the 20th International Conference on Autonomous Agents and Multiagent Systems; 2021. p. 151–158.
3. Cassell J, Vilhjálmsson HH, and Bickmore T. BEAT: the Behavior Expression Animation Toolkit. ACM Transactions on Graphics. 2001;p. 477–486.
4. Thiebaux M, Marsella S, Marshall AN, and Kallmann M. Smartbody: Behavior realization for embodied conversational agents. In: Proceedings of the 7th international joint conference on Autonomous agents and multiagent systems-Volume 1; 2008. p. 151–158.
5. Marsella S, Xu Y, Lhommet M, Feng A, Scherer S, and Shapiro A. Virtual character performance from speech. In: Proceedings of the 12th ACM SIGGRAPH/Eurographics Symposium on Computer Animation; 2013. p. 25–35.
6. Neff M, Kipp M, Albrecht I, and Seidel HP. Gesture modeling and animation based on a probabilistic re-creation of speaker style. ACM Transactions on Graphics. 2008;**27**(1):1–24.
7. Bergmann K, and Kopp S. GNetIc - Using bayesian decision networks for iconic gesture generation. In: International Workshop on Intelligent Virtual Agents. Springer; 2009. p. 76–89.
8. Fernández-Baena A, Montaño R, Antonijoan M, Roversi A, Miralles D, and Alías F. Gesture synthesis adapted to speech emphasis. Speech Communication. 2014;**57**:331–350.
9. Yang Y, Yang J, and Hodgins J. Statistics-based Motion Synthesis for Social Conversations. Computer Graphics Forum. 2020;.
10. Ferstl Y, and McDonnell R. Investigating the use of recurrent motion modelling for speech gesture generation. In: IVA '18: International Conference on Intelligent Virtual Agents (IVA '18); 2018. p. 93–98.
11. Kucherenko T, Jonell P, van Waveren S, Henter GE, Alexandersson S, Leite I, et al. Gesticulator: A framework for semantically-aware speech-driven gesture generation. In: Proceedings of the 2020 International Conference on Multimodal Interaction; 2020. p. 242–250.
12. Ondras J, Celiktutan O, Bremner P, and Gunes H. Audio-driven robot upper-body motion synthesis. IEEE Transactions on Cybernetics. 2020;.

[3] This work was funded by Science Foundation Ireland under the ADAPT Centre for Digital Content Technology (Grant 13/RC/2106).

13. Ginosar S, Bar A, Kohavi G, Chan C, Owens A, and Malik J. Learning Individual Styles of Conversational Gesture. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2019. p. 3497–3506.

14. Ferstl Y, Neff M, and McDonnell R. Adversarial gesture generation with realistic gesture phasing. Computers & Graphics. 2020;.

15. Alexanderson S, Henter GE, Kucherenko T, and Beskow J. Style-Controllable Speech-Driven Gesture Synthesis Using Normalising Flows. Computer Graphics Forum. 2020;.

16. Levine S, Krähenbühl P, Thrun S, and Koltun V. Gesture controllers. In: ACM SIGGRAPH 2010 papers; 2010. p. 1–11.

17. Stone M, DeCarlo D, Oh I, Rodriguez C, Stere A, Lees A, et al. Speaking with Hands: Creating Animated Conversational Characters from Recordings of Human Performance. ACM Transactions on Graphics - Proceedings of ACM SIGGRAPH 2004. 2004;**23**(3):506–513. Available from: http://dl.acm.org/citation.cfm?id=1015753.

18. Clavet S, and Büttner S. Motion Matching - The Road to Next Gen Animation. In: Proc. of Nucl.ai; 2015. .

19. Holden D, Kanoun O, Perepichka M, and Popa T. Learned motion matching. ACM Transactions on Graphics. 2020;**39**(4):53–1.

20. Kapadia M, Chiang Ik, Thomas T, Badler NI, and Kider Jr JT. Efficient motion retrieval in large motion databases. In: Proceedings of the ACM SIGGRAPH Symposium on Interactive 3D Graphics and Games; 2013. p. 19–28.

21. Valcik J, Sedmidubsky J, and Zezula P. Assessing similarity models for human-motion retrieval applications. Computer Animation and Virtual Worlds. 2016;**27**(5):484–500.

22. Kendon A. Some relationships between body motion and speech. Studies in dyadic communication. 1972;**7**(177):90.

23. Kim H, Kwak SS, and Kim M. Personality design of sociable robots by control of gesture design factors. In: RO-MAN 2008-The 17th IEEE International Symposium on Robot and Human Interactive Communication. IEEE; 2008. p. 494–499.

24. Kendon A. Gesticulation and speech: Two aspects of the process of utterance. The relationship of verbal and nonverbal communication. 1980;**25**(1980):207–227.

25. Eyben F, Scherer KR, Schuller BW, Sundberg J, Andre E, Busso C, et al. The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing. IEEE Transactions on Affective Computing. 2016;**7**(2):190–202.

26. Eyben F, Wöllmer M, and Schuller B. Opensmile: the munich versatile and fast open-source audio feature extractor. In: Proceedings of the 18th ACM international conference on Multimedia; 2010. p. 1459–1462.

27. Shapiro A, Faloutsos P, and Ng-Thow-Hing V. Dynamic animation and control environment. In: Proceedings of graphics interface 2005. Canadian Human-Computer Communications Society; 2005. p. 61–70.

28. Kucherenko T, Jonell P, Yoon Y, Wolfert P, and Henter GE. The GENEA Challenge 2020: Benchmarking gesture-generation systems on common data. 2020;.

29. Hartholt A, Traum D, Marsella SC, Shapiro A, Stratou G, Leuski A, et al. All together now. In: International Workshop on Intelligent Virtual Agents. Springer; 2013. p. 368–381.

30. Sonlu S, Güdükbay U, and Durupinar F. A Conversational Agent Framework with Multi-modal Personality Expression. ACM Transactions on Graphics. 2021;**40**(1):1–16.

31. Neff M, Wang Y, Abbott R, and Walker M. Evaluating the effect of gesture and language on personality perception in conversational agents. In: International Conference on Intelligent Virtual Agents. Springer; 2010. p. 222–235.

How to cite this article: Ferstl Y., M. Neff, and R. McDonnell (2021), ExpressGesture: Expressive Gesture Generation from Speech through Database Matching, *Computer Animation and Virtual Worlds, Special Issue: CASA 2021*, *2021;00*.