Analyzing Input and Output Representations for Speech-Driven Gesture Generation

Taras Kucherenko
KTH Royal Institute of Technology
Stockholm, Sweden
tarask@kth.se

Dai Hasegawa
Hokkai Gakuen University
Sapporo, Japan
dhasegawa@hgu.jp

Gustav Eje Henter
KTH Royal Institute of Technology
Stockholm, Sweden
ghe@kth.se

Naoshi Kaneko
Aoyama Gakuin University
Sagamihara, Japan
kaneko@it.aoyama.ac.jp

Hedvig Kjellström
KTH Royal Institute of Technology
Stockholm, Sweden
hedvig@kth.se

ABSTRACT

This paper presents a novel framework for automatic speech-driven gesture generation, applicable to human-agent interaction including both virtual agents and robots. Specifically, we extend recent deep-learning-based, data-driven methods for speech-driven gesture generation by incorporating representation learning. Our model takes speech as input and produces gestures as output, in the form of a sequence of 3D coordinates.

Our approach consists of two steps. First, we learn a lower-dimensional representation of human motion using a denoising autoencoder neural network, consisting of a motion encoder *MotionE* and a motion decoder *MotionD*. The learned representation preserves the most important aspects of the human pose variation while removing less relevant variation. Second, we train a novel encoder network *SpeechE* to map from speech to a corresponding motion representation with reduced dimensionality. At test time, the speech encoder and the motion decoder networks are combined: *SpeechE* predicts motion representations based on a given speech signal and *MotionD* then decodes these representations to produce motion sequences.

We evaluate different representation sizes in order to find the most effective dimensionality for the representation. We also evaluate the effects of using different speech features as input to the model. We find that mel-frequency cepstral coefficients (MFCCs), alone or combined with prosodic features, perform the best. The results of a subsequent user study confirm the benefits of the representation learning.

KEYWORDS

Gesture generation, social robotics, representation learning, neural network, deep learning, virtual agents

ACM Reference Format:
Taras Kucherenko, Dai Hasegawa, Gustav Eje Henter, Naoshi Kaneko, and Hedvig Kjellström. 2019. Analyzing Input and Output Representations for Speech-Driven Gesture Generation. In *ACM International Conference on Intelligent Virtual Agents (IVA '19), July 2–5, 2019, Paris, France.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3308532.3329472

1 INTRODUCTION

Conversational agents in the form of virtual agents or social robots are rapidly becoming wide-spread and many of us will soon interact regularly with them in our day-to-day lives. Humans use non-verbal behaviors to signal their intent, emotions and attitudes in human-human interactions [19, 25]. Similarly, it has been shown that people read and interpret robots' non-verbal cues similarly to non-verbal cues from other people [3]. Robots that are equipped with such non-verbal behaviors have shown to positively affect people's perception of the robot [32]. Conversational agents therefore need the ability to perceive and produce non-verbal communication.

An important part of non-verbal communication is gesticulation: gestures made with hands, arms, head pose and body pose communicate a large share of non-verbal content [26]. To facilitate natural human-agent interaction, it is hence important to enable robots and embodied virtual agents to accompany their speech with gestures in the way people do.

Most existing work on generating hand gestures relies on rule-based methods [5, 16, 27]. These methods are rather rigid as they can only generate gestures that are incorporated in the rules. Writing down rules for all possible gestures found in human interaction is highly labor-intense and time-consuming. Consequently, it is difficult to fully capture the richness of human gesticulation in rule-based systems. In this paper, we present a solution that eliminates this bottleneck by using a data-driven method that learns to generate human gestures from a dataset of human actions. More specifically, we use speech data, as it is highly correlated with hand gestures [26] and has the same temporal character.

To predict gestures from speech, we apply Deep Neural Networks (DNNs), which have been widely used in human skeleton modeling for motion prediction [24] as well as classification [4]. We further apply representation learning on top of conventional speech-input, gesture-output DNNs. Representation learning is a branch of unsupervised learning aiming to learn a better representation of the data. Typically, representation learning is applied to make a subsequent learning task easier. Inspired by previous successful applications to learning human motion dynamics, for example in prediction [4] and motion synthesis [12], this paper applies representation learning to the motion sequence, in order to extend previous approaches for neural-network-based speech-to-gesture mappings [13, 35].

The contributions of this paper are two-fold:

(1) We propose a novel speech-driven non-verbal behavior generation method that can be applied to any embodiment.
(2) We evaluate the importance of representation both for the motion (by doing representation learning) and for the speech (by comparing different speech feature extractors).

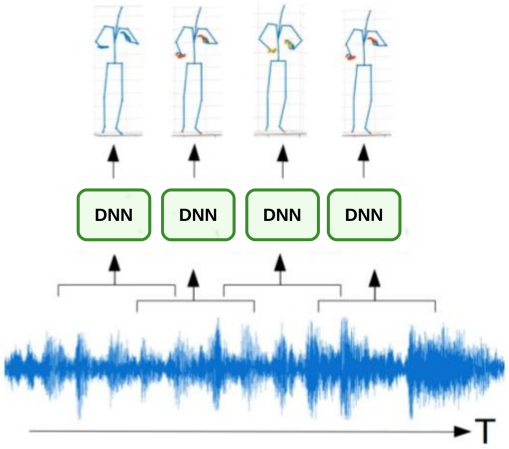


Figure 1: Framework overview. The Deep Neural Network (DNN) green boxes are further described in Figures 2 and 3.

We analyze which motion representation size yields the best results for the speech-driven gesture generation. Moreover, we numerically evaluate which speech features are most useful. Finally, we perform a user study, which finds that representation learning improved the perceived naturalness of the gestures over the baseline model.

Our work here extends our previous publication [21] by expanding the method description, investigating additional input features and significantly widening the scope of the objective evaluation. A video summary of this paper with visual examples is available at youtu.be/Iv7UBe92zrw.

2 REPRESENTATION LEARNING FOR SPEECH-MOTION MAPPING

2.1 Problem formulation

We frame the problem of speech-driven gesture generation as follows: given a sequence of speech features $\boldsymbol{s} = [\boldsymbol{s}_t]_{t=1:T}$ extracted from segments (frames) of speech audio at regular intervals t, the task is to generate a corresponding gesture sequence $\hat{\boldsymbol{g}} = [\hat{\boldsymbol{g}}_t]_{t=1:T}$ that a human might perform while uttering this speech.

A speech segment s_t would be typically represented by some features, such as Mel-Frequency Cepstral Coefficients [9], MFCCs, (which are commonly used in speech recognition) or prosodic features including pitch (F0), energy, and their derivatives (which are commonly used in speech emotion analysis). Similarly, the ground truth gestures g_t and predicted gestures $\hat{g}_t$ are typically represented as 3D-coordinate sequences: $\boldsymbol{g}_t = [x_{i,t}, y_{i,t}, z_{i,t}]_{i=1:n}$, n being the number of keypoints of the human body (such as shoulder, elbow, etc.) that are being modelled.

The most recent systems tend to perform mappings from $\boldsymbol{s}$ to $\hat{\boldsymbol{g}}$ using a neural network (NN) learned from data. The dataset typically contains recordings of human motion (for instance from a motion capture system) and the corresponding speech signals.

2.2 Baseline speech-to-motion mapping

Our model builds on the work of Hasegawa et al. [13]. In this section, we describe their model, which is our baseline system.

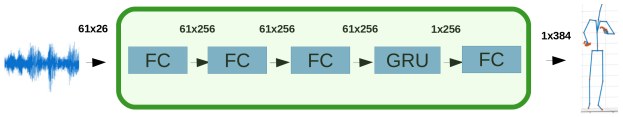


Figure 2: Baseline DNN for speech-to-motion mapping. The green box identifies the part used for the DNN in Figure 1.

The speech-gesture neural network [13] takes a speech sequence as input and generates a sequence of gestures frame by frame. As illustrated in Figure 1, the speech is processed in overlapping chunks of $C = 30$ frames (like in [13]) before and after the current time t. (The offset between frames in the figure is exaggerated for demonstration purposes.) An entire speech-feature window is fed into the network at each time step t: $\mathrm{NN}_{\mathrm{input}} = [s_{t-C}, ...s_{t-1}, s_t, s_{t+1}, ...s_{t+C}]$. The network is regularized by predicting not only the pose but also the velocity as output: $\mathrm{NN}_{\mathrm{output}} = [g_t, \Delta g_t]$. While incorporating the velocity into test-time predictions did not provide a significant improvement, the inclusion of velocity as a multitask objective during training forced the network to learn motion dynamics [35].

The **baseline neural network architecture** is illustrated in Figure 2. First, MFCC features are computed for every speech segment. Then three fully connected layers (FC) are applied to every chunk s_t. Next, a recurrent network layer with Gated Recurrent Units (GRUs) [8] is applied to the resulting sequence. Finally, an additional linear, fully-connected layer is used as the output layer.

We note that the baseline network we described is a minor modification of the network in [13]. Specifically, we use a different type of recurrent network units, namely GRUs instead of B-LSTMs. Our experiments found that this cuts the training time in half while maintaining the same prediction performance. We also used shorter window length for computing MFCC features, namely 0.02 s instead of 0.125 s, since MFCCs were developed to be informative about speech for these window lengths. The only other difference against [13] is that we did not post-process (smooth) the output sequences.

2.3 Proposed approach

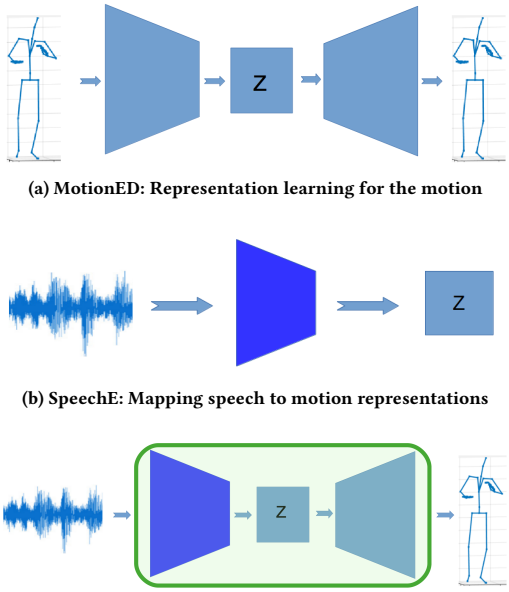
Our intent is to extend the baseline model by leveraging the power of representation learning. Our proposed approach has three steps:

(1) We apply representation learning to learn a motion representation $\boldsymbol{z}$.
(2) We learn a mapping from the chosen speech features $\boldsymbol{s}$ to the learned motion representation $\boldsymbol{z}$ (using the same NN architecture as in the baseline model).
(3) The two learned mappings are chained together to turn speech input $\boldsymbol{s}$ into motion output $\boldsymbol{g}$.

Motion representation learning

Figure 3a illustrates representation learning for human motion sequences. The aim of this step is to reduce the motion dimensionality, which confers two benefits: 1) simplifying the learning problem by reducing the output space dimensionality; and 2) reducing redundancy in the training data by forcing the system to concentrate important information to fewer numbers.

To learn motion representations, we used a neural network structure called a Denoising Autoencoder (DAE) [37] with one hidden layer ($\boldsymbol{z}$). This network learns to reconstruct the original data from input examples with additive noise while having a bottleneck layer



(a) MotionED: Representation learning for the motion

(b) SpeechE: Mapping speech to motion representations

(c) Combining the learned components: SpeechE and MotionD

Figure 3: How the proposed encoder-decoder DNN for speech-to-motion mapping is constructed. The green box denotes the part of the system used for the DNN in Figure 1.

in the middle. This bottleneck forces the network to compute lower dimensional representation. The network can be seen as a combination of two networks: *MotionE*, which encodes the motion $\boldsymbol{m}$ to the representation $\boldsymbol{z}$ and *MotionD*, which decodes the representation $\boldsymbol{z}$ back to the motion $\boldsymbol{m}$:

$$\boldsymbol{z} = MotionE(\boldsymbol{m}) \quad (1)$$

$$\hat{\boldsymbol{m}} = MotionD(\boldsymbol{z}) \quad (2)$$

The neural network learns to reconstruct the original motion coordinates as closely as possible by minimizing the mean squared error (MSE) loss function: $\text{MSE}(\boldsymbol{m}, \hat{\boldsymbol{m}}) = \|\hat{\boldsymbol{m}} - \boldsymbol{m}\|_2^2$.

Encoding speech to the motion representation
Figure 3b illustrates the principle of how we map speech to motion representation. Conceptually, the network performing this task fulfills the same role as the baseline network in Section 2.2. The main difference versus the baseline is that the output of the network is not raw motion values, but a compact, learned representation of motion. To be as comparable as possible to the baseline, we use the same network architecture to map speech to motion representations in the proposed system as the baseline used for mapping speech to motion. We call this network *SpeechE*.

Connecting the speech encoder and the motion decoder
Figure 3c illustrates how the system is used at testing time by chaining together the two previously learned mappings. First, speech input is fed to the *SpeechE* encoding net, which produces a sequence of motion representations. Those motion representations are then decoded into joint coordinates by the *MotionD* decoding net.

2.4 Implementation

The baseline neural network
Figure 2 shows the structure and layer sizes of the neural network used in the baseline system. As seen, the network inputs contained 61×26 elements, comprising 26-dimensional speech-derived MFCC vectors from the current frame plus 30 adjacent frames both before and after it, resulting in a total of 61 vectors in the input (While we describe and explore other audio features in 3.2, the baseline model only used MFCCs, to be consistent with [13]). The Fully Connected (FC) layers and the Gated Recurrent Unit (GRU) layers both had a width of 256 and used the ReLU activation function. Batch normalization and dropout with probability 0.1 of dropping activations were applied between every layer. Training minimized the mean squared error between predicted and ground-truth gesture sequences using the Adam optimizer [18] with learning rate 0.001 and batch size 2048. Training was run for 120 epochs, after which no further improvement in validation set loss was observed. Save for batch size and the number of epochs these hyperparameters were taken from the baseline paper [13].

The denoising autoencoder neural network
We trained a DAE with input size 384 (64 joints: 192 3D-coordinates and their first derivatives) and one hidden, feedforward layer in the encoder and decoder. Hyperparameters were optimized on our validation dataset, described in Section 3.1. Different widths were investigated for the bottleneck layer (see Section 4.1), with 325 units giving the best valiadation-data performance. Gaussian noise was added to each input with a standard deviation equal to 0.05 times the standard deviation of that feature dimension. Training minimized the MSE reconstruction loss using Adam with a learning rate of 0.001 and batch size 128. Training was run for 20 epochs.

3 EXPERIMENTAL SETUP

This section describes the data and gives technical detail regarding the experiments we conducted to evaluate the importance of input and output representations in speech-driven gesture generation.

3.1 Gesture-speech dataset

For our experiments, we used a gesture-speech dataset collected by Takeuchi et al. [36]. Motion data were recorded in a motion capture studio from two Japanese individuals having a conversation in the form of an interview. An experimenter asked questions prepared beforehand, and a performer answered them. The dataset contains MP3-encoded speech audio captured using headset microphones on each speaker, coupled with motion-capture motion data stored in the BioVision Hierarchy format (BVH). The BVH data describes motion as a time sequence of Euler rotations for each joint in the defined skeleton hierarchy. These Euler angles were converted to a total of 64 global joint positions in 3D. As some recordings had a different framerate than others, we downsampled all recordings to a common framerate of to 20 frames per second (fps). For the representation learning, each dimension was standardized to mean zero and maximum (absolute) value one.

The dataset contains 1,047 utterances[1], of which our experiments used 957 for training, 45 for validation, and 45 testing. The relationship between various speech-audio features and the 64 joint positions was thus learned from 171 minutes of training data at 20 fps, resulting in 206,000 training frames.

3.2 Feature extraction

The ease of learning and the limits of expressiveness for a speech-to-gesture system depend greatly on the input features used. Simple features that encapsulate the most important information are likely to work well for learning from small datasets, whereas rich and complex features might allow learning additional aspects of speech-driven gesture behavior, but may require more data to achieve good accuracy. We experimented with three different, well-established audio features as inputs to the neural network, namely i) MFCCs, ii) spectrograms, and iii) prosodic features.

In terms of implementation, 26 MFCCs were extracted with a window length of 0.02 s and a hop length of 0.01 s, which amounts to 100 analysis frames per second. Our spectrogram features, meanwhile, were 64-dimensional and extracted with the window length and hop size 0.005 s, yielding a rate of 200 fps. Frequencies that carry little speech information (below the hearing threshold of 20 Hz, or above 8000 Hz) were removed. Both the MFCC and the spectrogram sequences were downsampled to match the motion frequency of 20 fps by replacing every 5 (MFCCs) or 10 (spectrogram) frames by their average. (This averaging prevents aliasing artifacts.)

As an alternative to MFCCs and spectrum-based features, we also considered prosodic features. These differ in that prosody encompasses intonation, rhythm, and anything else about the speech outside of the specific words spoken (e.g., semantics and syntax). Prosodic features were previously used for gesture prediction in early data-driven work by Chiu & Marsella [6]. For this study, we considered pitch and energy (intensity) information. The information in these features has a lower bitrate and is not sufficient for discriminating between and responding differently to arbitrary words, but may still be informative for predicting non-verbal emphases like beat gestures and their timings.

We considered four specific prosodic features, extracted from the speech audio with a window length of 0.005 s, resulting in 200 fps. Our two first prosodic features were the energy of the speech signal and the time derivative (finite difference) of the energy series. The third and fourth features were the logarithm of the F0 (pitch) contour, which contains information about the speech intonation, and its time derivative. We extracted pitch and intensity values from audio using Praat [2] and normalized pitch and intensity as in [6]: the pitch values were adjusted by taking $\log(x+1)-4$ and setting negative values to zero, and the intensity values were adjusted by taking $\log(x)-3$. All these features were again downsampled to the motion frequency of 20 fps using averaging.

3.3 Numerical evaluation measures

We used both objective and subjective measures to evaluate the different approaches under investigation. Among the former, two kinds of error measures were considered:

Average Position Error (APE) The APE is the average Euclidean distance between the predicted coordinates $\hat{g}$ and the original coordinates g:

$$\mathrm{APE}(g_t^n, \hat{g}_t^n) = \frac{1}{DT}\sum_{t=1}^{T}\sum_{d=1}^{D}\|g_t^n - \hat{g}_t^n\|_2 \tag{3}$$

where T is the total duration of the sequence, D is the dimensionality of the motion data and n is a sequence index.

Motion Statistics We considered the average values and distributions of acceleration and jerk for the produced motion.

We believe the motion statistics to be the most informative for our task: in contrast to tracking, the purpose of gesture generation is not to reproduce one specific *true* position, but rather to produce a plausible candidate for natural motion. Plausible motions do not require measures like speed or jerk to closely follow the original motion, but they should follow a similar distribution. That is why we study distribution statistics, namely average speed and jerk.

Since there is some randomness in system training, e.g., due to random initial network weights, we evaluated every condition five times and report the mean and standard deviation of those results.

4 RESULTS AND DISCUSSION

This section presents an analysis of the performance of the gesture-prediction system. We investigate different design aspects of our system that relate to the importance of representations, namely the speech and motion representations used.

4.1 Importance of motion encoding

We first evaluated how different dimensionalities for the learned motion representation affected the prediction accuracy of the full system. Figure 4 graphs the results of this evaluation. In terms of average position error (APE) (see Figure 4a) the optimal embedding-space dimensionality is clearly 325, which is smaller than the original data dimensionality (384). Motion jerkiness (see Figure 4b) is also lowest for dimensionality 325, but only by a slight margin compared to the uncertainty in the estimates. Importantly, the proposed system performs much better than the baseline [13] on both evaluation measures. The difference in the average jerk, in particular, is highly significant. This validates our decision to use representation learning to improve gesture generation models. While motion jerkiness can be reduced through post-processing, as in [13], that does not address the underlying shortcoming of the model.

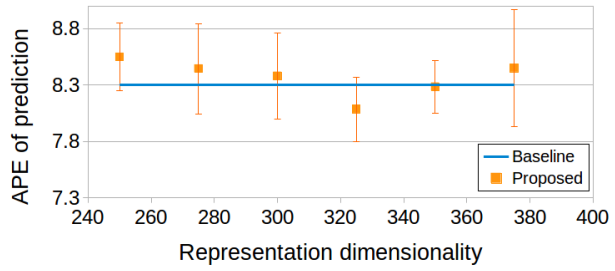
The numerical results are seen to vary noticeably between different runs, suggesting that training might converge to different local optima depending on the random initial weights.

4.2 Input speech representation

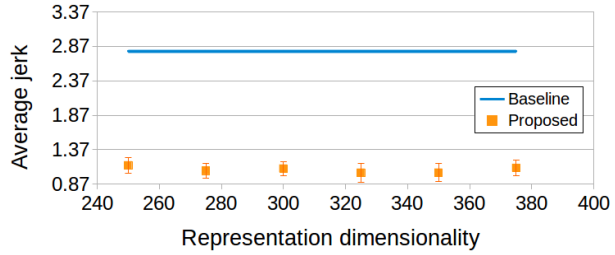
Having established the benefits of representation learning for the output motion, we next analyze which input features perform the best for our speech-driven gesture generation system. In particular, we compare three different features – MFCCs, raw power-spectrogram values, and prosodic features (log F0 contour, energy, and their derivatives) – as described in Section 3.2.

From Table 1, we observe that MFCCs achieve the lowest APE, but produce motion with higher acceleration and jerkiness than the spectrogram features do. Spectrogram features gave suboptimal

[1]The original paper reports 1,049 utterances, which is a typo.



(a) Average position error (APE). The baseline APE (blue line) is 8.3±0.4.



(b) Average jerk. Baseline jerk is 2.8±0.3 while ground-truth jerk is 0.54.

Figure 4: Effect of learned-representation dimensionality in the proposed model.

Table 1: Objective evaluation of different speech features, averaged over five re-trainings of the system.

Model/feature	APE	Acceleration	Jerk
Static mean pose	8.95	0	0
Prosodic	8.56±0.2	0.90±0.03	1.52±0.07
Spectrogram	8.27±0.4	**0.51**±0.07	**0.85**±0.12
Spectr. + Pros.	8.11±0.3	0.57±0.08	0.95±0.12
MFCC	**7.66**±0.2	0.53±0.03	0.91±0.05
MFCC + Pros.	**7.65**±0.2	0.58±0.06	0.97±0.11
Baseline [13] (MFCC)	8.07±0.1	1.50±0.03	2.62±0.05
Ground truth	0	0.38	0.54

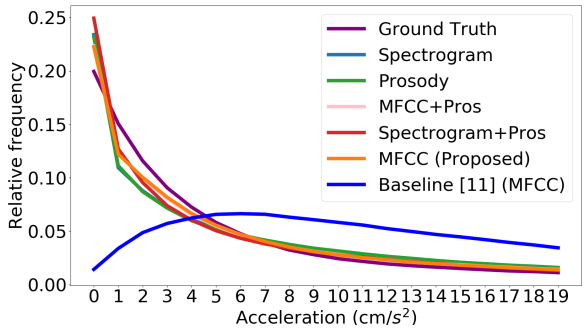
APE, but match ground-truth acceleration and jerk better than the other features we studied.

4.3 Detailed performance analysis

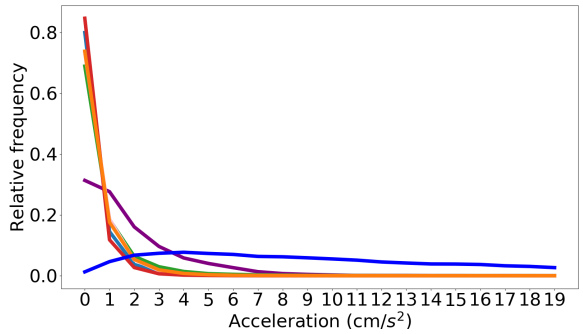
The objective measures in Table 1 do not unambiguously establish which input features would be the best choice for our predictor. We therefore further analyze the statistics of the generated motion, particularly acceleration. Producing the right motion with the right acceleration distribution is crucial for generating convincing motions, as too fast or too slow motion does not look natural.

To investigate the motion statistics associated with the different input features, we computed acceleration histograms of the generated motions and compared those against histograms derived from the ground truth. We calculated the relative frequency of different acceleration values over frames in all 45 test sequences, split into bins of equal width. For easy comparison, our histograms are visualized as line plots rather than bar plots.

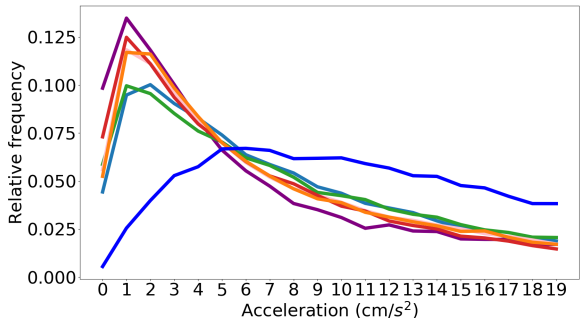
Figure 5a presents acceleration histograms across all joints for different input features. The acceleration distribution of the baseline model deviate more from the ground truth than our model does.



(a) Average acceleration histogram.



(b) Acceleration histogram for shoulders. Legend as in (a).



(c) Acceleration histogram for hands Legend as in (a).

Figure 5: Acceleration distributions given different speech features. Firstly, the motion produced from our model (with any input feature) is more similar to the acceleration distribution of the ground-truth motion, compared to motion from the baseline model. Secondly, we find that MFCCs produce an acceleration distribution most similar to the ground truth, especially for the hands, as shown in (c).

Table 2: Statements evaluated in user study

Scale	Statement (translated from Japanese)
Naturalness	Gesture was natural Gesture was smooth Gesture was comfortable
Time consistency	Gesture timing was matched to speech Gesture speed was matched to speech Gesture pace was matched to speech
Semantic consistency	Gesture was matched to speech content Gesture well described speech content Gesture helped me understand the content

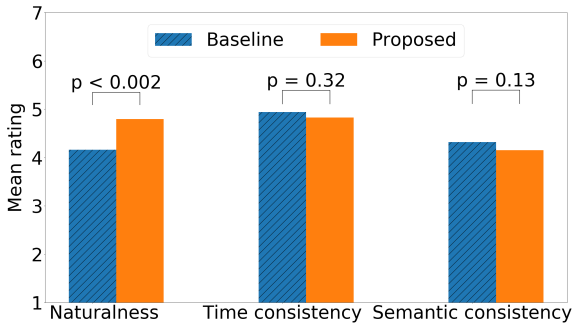


Figure 6: Results from the user study. We note a significant difference in naturalness, but not the other scales.

Since the results in Figure 5a are averaged over all joints, they do not indicate whether all the joints move naturally. To address this we also analyze the acceleration distribution for certain specific joints. Figure 5b shows an acceleration histogram calculated for the shoulders only. We see that our system with all the speech features has acceleration distributions very close to one another, but that all of them far away from the actual data. A possible explanation for this could be that shoulder motion might be difficult to predict from the speech input, in which case the predicted motion is likely to stay close the mean shoulder position.

Figure 5c shows acceleration histograms for the hands. Hands convey the most important gesture information, suggesting that this plot is the most informative. Here, the MFCC-based system is much closer to the ground truth. Combining MFCCs and prosodic features resulted in similar performance as for MFCC inputs alone. This could be due to redundancy in the information exposed by MFCCs and prosodic features, or due to our networks and optimizer not being able to exploit synergies between the two representations.

Taken together, Figures 5a-c suggest that motion generated from MFCC features give acceleration statistics as similar or more similar to the ground truth as those of motion generated from other features. Moreover, using MFCCs as input features makes our proposed system consistent with the baseline paper [13].

4.4 User study

The most important goal in gesture generation is to produce motion patterns that are convincing to human observers. Since improvements in objective measures do not always translate into superior subjective quality for human observers, we validated our conclusions by means of a user study comparing key systems.

We conducted a 1×2 factorial design with the within-subjects factor being representation learning (baseline vs. encoded). The encoded gestures were generated by the proposed method from MFCC input. We randomly selected 10 utterances from a test dataset of 45 utterances, for each of which we created two videos using the two gesture generation systems. Visual examples are provided at https://vimeo.com/album/5667276. After watching each video, we asked participants to rate nine statements about the naturalness, time consistency, and semantic consistency of the motion. The statements were the same as in the baseline paper [13] and are listed in Table 2. Ratings used a seven-point Likert scale anchored from strongly disagree (1) to strongly agree (7). The utterance order was fixed for every participant, but the gesture conditions (baseline vs. encoded) were counter-balanced. With 10 speech segments and two gesture-generation systems, we obtained 20 videos, producing 180 ratings in total per subject, 60 for each scale in Table 2.

19 native speakers of Japanese (17 male, 2 female), on average 26 years old, participated in the user study. A paired-sample t-test was conducted to evaluate the impact of the motion encoding on the perception of the produced gestures. Figure 6 illustrates the results we obtained for the three scales being evaluated. We found a significant difference in naturalness between the baseline (M=4.16, SD=0.93) and proposed model (M=4.79, SD=0.89), t=-3.6372, p<0.002. A 95%-confidence interval for the mean rating improvement with the proposed system is (0.27,1.00). There were no significant difference on the other scales: for time-consistency t=1.0192, p=0.32, for semantic consistency t=1.5667, p=0.13. These results indicate that gestures generated by the proposed method (i.e., with representation learning) were perceived as more natural than the baseline.

5 RELATED WORK

While most of the work on non-verbal behavior generation in the literature considers rule-based systems [5, 16, 33], we review only data-driven approaches and pay special attention to methods incorporating elements of representation learning, since that is the direction of our research. For a review of rule-based systems, we refer the reader to Wagner et al. [38].

5.1 Data-driven head and face movements

Facial-expression generation has been an active field of research for several decades. Many of the state-of-the-art methods are data-driven. Several recent works have applied neural networks in this domain [10, 11, 29, 30, 34]. Among the cited works, Haag & Shimodaira [11] use a bottleneck network to learn compact representations, although their bottleneck features subsequently are used to define prediction inputs rather than prediction outputs as in the work we presented. Our proposed method works on a different aspect of non-verbal behavior that co-occurs with speech, namely generating body motion driven by speech.

5.2 Data-driven body motion generation

Generating body motion is an active area of research with applications to animation, computer games, and other simulations. Current state-of-the-art approaches in such body-motion generation are generally data-driven and based on deep learning [28, 40, 41]. Zhou et al. [41] proposed a modified training regime to make recurrent neural networks generate human motion with greater long-term stability, while Pavllo et al. [28] formulated separate short-term and long-term recurrent motion predictors, using quaternions to more adequately express body rotations.

Some particularly relevant works for our purposes are [4, 14, 15, 23]. All of these leverage representation learning (various forms of autoencoders) that predict human motion, yielding accurate yet parsimonious predictors. Habibie et al. [12] extended this general approach to include an external control signal in an application to human locomotion generation with body speed and direction as the control input. Our approach is broadly similar, but generates body motion from speech rather than position information.

5.3 Speech-driven gesture generation

Like body motion in general, gesture generation has also begun to shift towards data-driven methods, for example [6, 7, 13]. Several researchers have tried to combine data-driven approaches with rule-based systems. For example, Bergmann & Kopp [1] learned a Bayesian decision network for generating iconic gestures. Their system is a hybrid between data-driven and rule-based models because they learn rules from data. Sadoughi et al. [31] used probabilistic graphical models with an additional hidden node to provide contextual information, such as a discourse function. They experimented on only three hand gestures and two head motions. We believe that regression methods that learn and predict arbitrary movements, like the one we have proposed, represent a more flexible and scalable approach than the use of discrete and pre-defined gestures.

The work of Chiu & Marsella [6] is of great relevance to the work have presented, in that they took a regression approach and also utilized representation learning. Specifically, they used wrist height in upper-body motion to identify gesticulation in motion capture data of persons engaged in conversation. A network based on Restricted Boltzmann Machines (RBMs) was used to learn representations of arm gesture motion, and these representations were subsequently predicted based on prosodic speech-feature inputs using another network also based on RBMs. Levine et al. [22] also used an intermediate state between speech and gestures. The main differences are that they used hidden Markov models, whose discrete states are less powerful than recurrent neural networks, and that they selected motions from a fixed library, while our model can generate unseen gestures. Later Chiu et al. [7] proposed a method to predict co-verbal gestures using a machine learning setup with feedforward neural networks followed by Conditional Random Fields (CRFs) for temporal smoothing. They limited themselves to a set of 12 discrete, pre-defined gestures and used a classification-based approach.

Recently, Hasegawa et al. [13] designed a speech-driven neural network capable of producing 3D motion sequences. We built our model on this work while extending it with motion-representation learning, since learned representations have improved motion prediction in other applications as surveyed in Section 5.2.

6 CONCLUSIONS AND FUTURE WORK

This paper presented a new model for speech-driven gesture generation. Our method extends prior work on deep learning for gesture generation by applying representation learning. The motion representation is learned first, after which a network is trained to predict such representations from speech, instead of directly mapping speech to raw joint coordinates as in prior work. We also evaluated the effect of different representations for the input speech. Our code is publicly available to encourage replication of our results.[2]

Our experiments show that representation learning improves the objective and subjective performance of the speech-to-gesture neural network. Although models with and without representation learning were rated similarly in terms of time consistency and semantic consistency, subjects rated the gestures generated by the proposed method as significantly more natural than the baseline.

The main limitation of our method, as with any data-driven method and particularly those based on deep learning, is that it requires substantial amounts of parallel speech-and-motion training data of sufficient quality in order to obtain good prediction performance. In the future, we might overcome this limitation by obtaining datasets directly from publicly-available video recordings using motion-estimation techniques.

6.1 Future work

We see several interesting directions for future research:

Firstly, it is beneficial to make the model probabilistic, e.g., by using a Variational Autoencoder (VAE) as in [20]. A person is likely to gesticulate differently at different times for the same utterance. It is thus an appealing idea to make a conversational agent also generate different gestures every time they speak the same sentence. For this we need to make the mapping probabilistic, to represent a probability distribution over plausible motions and then draw samples from that distribution. VAEs can provide us with this functionality.

Secondly, text should be taken into account, e.g., as in [17]. Gestures that co-occur with speech depend greatly on the semantic content of the utterance. Our model generates mostly beat gestures, as we rely only on speech acoustics as input. Hence the model can benefit from incorporating the text transcription of the utterance along with the speech audio. This may enable producing a wider range of gestures (also metaphoric and deictic gestures).

Lastly, the learned model can be applied to a humanoid robot so that the robot's speech is accompanied by appropriate co-speech gestures, for instance on the NAO robot as in [39].

ACKNOWLEDGEMENT

The authors would like to thank Sanne van Waveren, Iolanda Leite and Simon Alexanderson for helpful discussions. This project is supported by the Swedish Foundation for Strategic Research Grant No.: RIT15-0107 (EACare). This work was also partially supported by JSPS Grant-in-Aid for Young Scientists (B) Grant Number 17K18075.

[2] github.com/GestureGeneration/Speech_driven_gesture_generation_with_autoencoder

REFERENCES

[1] Kirsten Bergmann and Stefan Kopp. 2009. GNetIc–Using Bayesian decision networks for iconic gesture generation. In *International Workshop on Intelligent Virtual Agents (IVA '09)*. Springer, 76–89.

[2] Paul Boersma. 2002. Praat, a system for doing phonetics by computer. *Glot International* 5, 9/10 (2002), 341–345.

[3] Cynthia Breazeal, Cory D Kidd, Andrea Lockerd Thomaz, Guy Hoffman, and Matt Berlin. 2005. Effects of nonverbal communication on efficiency and robustness in human-robot teamwork. In *International Conference on Intelligent Robots and Systems, (IROS '05)*. IEEE, 708–713.

[4] Judith Bütepage, Michael J. Black, Danica Kragic, and Hedvig Kjellström. 2017. Deep representation learning for human notion prediction and classification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR '17)*. IEEE.

[5] Justine Cassell, Hannes Högni Vilhjálmsson, and Timothy Bickmore. 2001. Beat: The behavior expression animation toolkit. In *Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH '01)*. ACM.

[6] Chung-Cheng Chiu and Stacy Marsella. 2011. How to train your avatar: A data driven approach to gesture generation. In *International Workshop on Intelligent Virtual Agents (IVA'11)*. Springer, 127–140.

[7] Chung-Cheng Chiu, Louis-Philippe Morency, and Stacy Marsella. 2015. Predicting co-verbal gestures: A deep and temporal modeling approach. In *International Conference on Intelligent Virtual Agents (IVA '15)*. Springer.

[8] Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder–decoder approaches. *Syntax, Semantics and Structure in Statistical Translation* (2014), 103.

[9] Steven Davis and Paul Mermelstein. 1980. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE International Conference on Acoustics, Speech and Signal Processing (ACASSP '80)* 28, 4, 357–366.

[10] David Greenwood, Stephen Laycock, and Iain Matthews. 2017. Predicting head pose from speech with a conditional variational autoencoder. In *Conference of the International Speech Communication Association (Interspeech '17)*. ISCA, 3991–3995.

[11] Kathrin Haag and Hiroshi Shimodaira. 2016. Bidirectional LSTM networks employing stacked bottleneck features for expressive speech-driven head motion synthesis. In *International Conference on Intelligent Virtual Agents (IVA '16)*. Springer, 198–207.

[12] Ikhsanul Habibie, Daniel Holden, Jonathan Schwarz, Joe Yearsley, and Taku Komura. 2017. A recurrent variational autoencoder for human motion synthesis. *IEEE Computer Graphics and Applications* 37 (2017), 4.

[13] Dai Hasegawa, Naoshi Kaneko, Shinichi Shirakawa, Hiroshi Sakuta, and Kazuhiko Sumi. 2018. Evaluation of speech-to-gesture generation using bi-directional LSTM network. In *International Conference on Intelligent Virtual Agents (IVA '18)*. ACM, 79–86.

[14] Daniel Holden, Jun Saito, and Taku Komura. 2016. A deep learning framework for character motion synthesis and editing. *ACM Transactions on Graphics (TOG)* 35, 4 (2016), 138:1–138:11.

[15] Daniel Holden, Jun Saito, Taku Komura, and Thomas Joyce. 2015. Learning motion manifolds with convolutional autoencoders. In *SIGGRAPH Asia Technical Briefs*. 18:1–18:4.

[16] Chien-Ming Huang and Bilge Mutlu. 2012. Robot behavior toolkit: Generating effective social behaviors for robots. In *International Conference on Human Robot Interaction (HRI '12)*. ACM/IEEE.

[17] Ryo Ishii, Taichi Katayama, Ryuichiro Higashinaka, and Junji Tomita. 2018. Generating body motions using spoken language in dialogue. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents (IVA '18)*. ACM, 87–92.

[18] Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR '15)*.

[19] Mark L Knapp, Judith A Hall, and Terrence G Horgan. 2013. *Nonverbal Communication in Human Interaction*. Wadsworth, Cengage Learning.

[20] Taras Kucherenko. 2018. Data driven non-verbal behavior generation for humanoid robots. In *ACM International Conference on Multimodal Interaction, Doctoral Consortium (ICMI '18)*. ACM, 520–523.

[21] Taras Kucherenko, Dai Hasegawa, Naoshi Kaneko, Gustav Eje Henter, and Hedvig Kjellström. 2019. On the importance of representations for speech-driven gesture generation. In *International Conference on Autonomous Agents and Multiagent Systems (AAMAS '19)*. ACM.

[22] Sergey Levine, Philipp Krähenbühl, Sebastian Thrun, and Vladlen Koltun. 2010. Gesture controllers. *ACM Transactions on Graphics (TOG)* 29, 4 (2010), 124.

[23] Hailong Liu and Tadahiro Taniguchi. 2014. Feature extraction and pattern recognition for human motion by a deep sparse autoencoder. In *International Conference on Computer and Information Technology (CIT '14)*. IEEE, 173–181.

[24] Julieta Martinez, Michael J Black, and Javier Romero. 2017. On human motion prediction using recurrent neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR '17)*. IEEE, 4674–4683.

[25] David Matsumoto, Mark G Frank, and Hyi Sung Hwang. 2013. *Nonverbal Communication: Science and Applications*. Sage.

[26] David McNeill. 1992. *Hand and Mind: What Gestures Reveal about Thought*. University of Chicago Press.

[27] Victor Ng-Thow-Hing, Pengcheng Luo, and Sandra Okita. 2010. Synchronized gesture and speech production for humanoid robots. In *International Conference on Intelligent Robots and Systems (IROS '10)*. IEEE/RSJ.

[28] Dario Pavllo, David Grangier, and Michael Auli. 2018. QuaterNet: A quaternion-based recurrent model for human motion. In *British Machine Vision Conference (BMVC '18)*.

[29] Najmeh Sadoughi and Carlos Busso. 2017. Joint learning of speech-driven facial motion with bidirectional long-short term memory. In *International Conference on Intelligent Virtual Agents (IVA '17)*. Springer, 389–402.

[30] Najmeh Sadoughi and Carlos Busso. 2018. Novel realizations of speech-driven head movements with generative adversarial networks. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '18)*. IEEE, 6169–6173.

[31] Najmeh Sadoughi and Carlos Busso. 2019. Speech-driven animation with meaningful behaviors. *Speech Communication* 110 (2019), 90–100.

[32] Maha Salem, Friederike Eyssel, Katharina Rohlfing, Stefan Kopp, and Frank Joublin. 2013. To err is human (-like): Effects of robot gesture on perceived anthropomorphism and likability. *International Journal of Social Robotics* 5, 3 (2013), 313–323.

[33] Giampiero Salvi, Jonas Beskow, Samer Al Moubayed, and Björn Granström. 2009. SynFace: Speech-driven facial animation for virtual speech-reading support. *EURASIP Journal on Audio, Speech, and Music Processing* (2009), 3.

[34] Supasorn Suwajanakorn, Steven M Seitz, and Ira Kemelmacher-Shlizerman. 2017. Synthesizing Obama: Learning lip sync from audio. *ACM Transactions on Graphics (TOG)* 36, 4 (2017), 95.

[35] Kenta Takeuchi, Dai Hasegawa, Shinichi Shirakawa, Naoshi Kaneko, Hiroshi Sakuta, and Kazuhiko Sumi. 2017. Speech-to-gesture generation: A challenge in deep learning approach with bi-directional LSTM. In *International Conference on Human Agent Interaction (HAI '17)*.

[36] Kenta Takeuchi, Souichirou Kubota, Keisuke Suzuki, Dai Hasegawa, and Hiroshi Sakuta. 2017. Creating a gesture-speech dataset for speech-based automatic gesture generation. In *International Conference on Human-Computer Interaction (HCI '17)*. Springer, 198–202.

[37] Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, and Pierre-Antoine Manzagol. 2010. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of Machine Learning Research* 11, Dec (2010), 3371–3408.

[38] Petra Wagner, Zofia Malisz, and Stefan Kopp. 2014. Gesture and speech in interaction: An overview. *Speech Communication* 57 (2014), 209–232.

[39] 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 *International Conference on Robotics and Automation (ICRA '19)*. IEEE.

[40] He Zhang, Sebastian Starke, Taku Komura, and Jun Saito. 2018. Mode-adaptive neural networks for quadruped motion control. *ACM Transactions on Graphics (TOG)* 37, 4 (2018), 145.

[41] Yi Zhou, Zimo Li, Shuangjiu Xiao, Chong He, Zeng Huang, and Hao Li. 2017. Auto-conditioned recurrent networks for extended complex human motion synthesis. In *International Conference on Learning Representations (ICLR '17)*.