Generating coherent spontaneous speech and gesture from text

Submission 1054

Abstract

Embodied human communication encompasses both verbal (speech) and non-verbal information (e.g., stance and gesture). Recent advances in machine learning have substantially improved the technologies for generating synthetic versions both of these types of data: On the speech side, text-to-speech systems are now able to generate highly convincing, spontaneous-sounding speech using unscripted speech audio as the source material. On the motion side, probabilistic motion-generation methods can now synthesise vivid and lifelike speech-driven 3D gesticulation. In this paper, we put these two state-of-the-art technologies together in a coherent fashion for the first time. Concretely, we demonstrate a proof-of-concept system trained on a single-speaker audio and motion-capture dataset, that is able to generate both speech and full-body gestures together from text input. In contrast to previous approaches for joint speech-and-gesture generation, we generate full-body gestures from speech synthesis trained on recordings of spontaneous speech from the same person as the motion-capture data. We illustrate our results by visualising gesture spaces and text-speech-gesture alignments, and through a demonstration video.

Keywords: Gestures, Speech, Data-driven animation

1. Introduction

Recent years have witnessed the rise of autonomous agents like digital assistants, embodied conversational agents and social robots. While some of these communicate only through speech, others also exhibit non-verbal behaviours like gestures, head, and facial movements. Currently, both of these modalities are far from natural and are usually criticised for appearing stiff and "robotic". This is not surprising, given that text-to-speech (TTS) is typically trained on corpora of speech read aloud, while gestures often comprise concatenated snippets of "canned" motion. This stands in stark contrast to the smooth and spontaneous behaviours associated with human-human communication. Recently, independent research in both TTS and gesture synthesis has shown that deep learning can remedy some of these issues. In speech synthesis, Tacotron [SPW*18] has ushered in a quantum leap in modelling speech intonation and prosody, greatly increasing the vividness of synthesised speech, and also enabled successful TTS from diverse and messy data like spontaneous speech recordings [SHBG19b]. For motion synthesis, emerging probabilistic modelling techniques such as normalising flows [PNR*19] are able to learn to generate convincing and controllable motion from motion-capture data [HAB19]. Trained on the same source material, consisting of both spontaneous speech and motion, researchers have generated highly natural-sounding speech synthesis and full-body gestures [Aut].

In this paper, we demonstrate for the first time how existing state-of-the-art, data-driven speech synthesis and gesture generation systems can be put together, to synthesise both speech and gesture at the same time simply from text input. Our key innovations are:

1. We generate hand and full-body gestures from synthetic speech.
2. We drive gesture-generation using TTS trained on spontaneous speech recordings. This is important since previous TTS systems have been based on recordings of texts being read aloud, which is not ideal since humans do not gesture when reading.
3. We train speech and gesture-generation on data from one and the same person, so that the gesture behaviour matches the speech.
4. We utilise the synthetic speech acoustics to drive the gesture generation, so that gestures and speech are inherently aligned.

To the best of our knowledge, our system is the first data-driven gesture-generation system capable of producing co-speech gesticulation from synthetic speech.

2. Background on speech synthesis

While the task of artificially creating human speech audio has been treated like a machine-learning problem since at least the 1990s, the deep-learning revolution has led to several paradigm shifts in TTS technology. Neural vocoders have revolutionised the signal quality of synthetic audio, but – more importantly for our purposes – sequence-to-sequence models with neural attention, e.g., Tacotron 2 [SPW*18], have ushered in a step change in the ability to create speech with vivid and lifelike prosody (intonation, dynamics, etc.) from speech alone. These techniques have also shown unprecedented versatility and robustness to challenging data, with Székely et al. [SHBG19b] recently demonstrating that Tacotron 2 is able to create highly convincing TTS from recordings of spontaneous speech with imprecise automatic transcriptions. This contrasts against traditional TTS systems, which to attain acceptable quality had to be built from recordings of isolated sentences accurately read aloud. The new ability to produce convincingly spontaneous-sounding synthetic speech is a game-changer, since the vast majority of human speech is spontaneous and part of a dialogue, and is particularly relevant to our interest in generating coherent spontaneous co-speech gestures.

Another issue with spontaneous speech is that it is not comprised of easily separable, well-defined sentences in the linguistic sense; however, Tacotron and other established TTS engines expect isolated utterances, and will run out of memory when applied to longer chunks of speech. Székely et al. solved this problem by proposing to look at breathing in the spontaneous speech recordings [SHG19], specifically using automatically-detected breath events as segmentation points of spontaneous speech corpora for TTS. This is compelling since breaths relate to the speech planning process and tend to be highly correlated with major prosodic breaks. As a byproduct, this allows control over breathing in the synthetic speech. A similar annotation-based approach can also enable control over unplanned conversational phenomena such as disfluencies [SHBG19a]. We leverage these breakthroughs for the speech synthesis in this paper.

3. Background on gesture synthesis

In general, the synthesis of non-verbal human behavior has gradually moved from rule-based systems [WMK14] towards data-driven approaches in recent decades. Since we continue this line of research, we review only data-driven methods in this paper. Most data-driven gesture-generation systems only consider speech audio as input, e.g., [SB19, HKS*18, KHH*19, GBK*19]. Sadoughi et al. [SB19] learned probabilistic graphical models for generating a discrete set of hand and head gestures. Hasegawa et al. [HKS*18] designed a speech-driven neural network to predict 3D motion sequences, which Kucherenko et al. [KHH*19] extended to include representation learning of the poses. Ginosar et al. [GBK*19] developed a system for predicting 2D gesture poses from YouTube videos in the wild. However, all of these models were validated only on held-out recorded natural speech inputs, and their ability to generate appropriate gesticulation when applied to synthetic speech audio is unknown.

Despite the dearth of co-speech gestures generated from synthetic speech, TTS has been investigated for driving the synthesis of other aspects of non-verbal behavior instead, such as facial expressions [CBL*19] and head motion [SLB17]. Joint generation of speech and facial expression is a particularly well-studied field called *audio-visual speech synthesis*; see [MV15] for a review. Our task is different from audio-visual speech synthesis, as we generate full-body gestures driven by the synthetic speech, but not facial expressions. This is arguably a more challenging task, given that the correlation between speech and full-body gestures is lower than that for facial movements.

Text has also been used as input for gesture generation, but less frequently, and mainly with rule-based models, [RPCM18] being a recent example. Only a handful of text-input models ([IMMI18, YKJ*19]) are data-driven. Specifically, Ishi et al. [IMMI18] generated gestures from text through a series of probabilistic mappings from words to hand-gesture clusters, via word concepts and gesture types. Yoon et al. [YKJ*19], meanwhile, learned to map utterance texts (represented by word vectors) to gestures by a sequence-to-sequence recurrent neural network. If paired with a TTS system, these approaches can produce both synthetic gestures and synthetic speech audio. However, neither of these methods take the the audio signal of the speech into account, making them inappropriate for modeling gesture-speech time alignment. This is problematic for gesture timing in general, and beat gestures in particular.

4. System description

We construct our multimodal speech-and-gesture-generation system by chaining together two essential pieces, namely a spontaneous text-to-speech synthesiser and a speech-audio-driven gesture generator. In this section we describe these two components. Both components were trained using the Trinity Speech and Gesture corpus [FM18], which comprises 31 episodes (244 minutes) of full-body motion capture (69-joint skeleton) and audio of one male actor speaking spontaneously (with disfluencies and fillers) on different topics while shifting stance and moving freely through the capture area. We held out the last episode as test data.

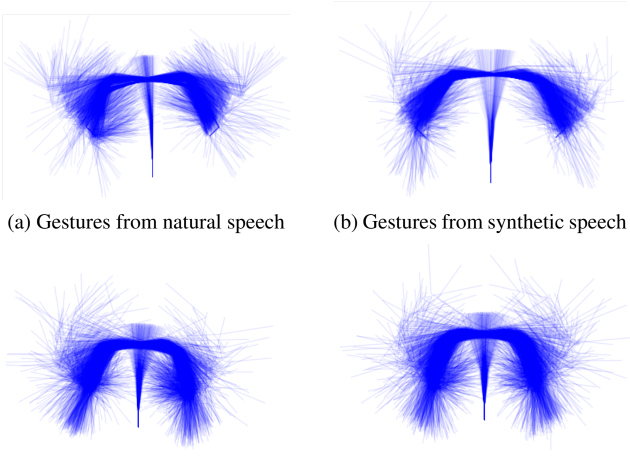
4.1. Spontaneous speech synthesis

Our data annotation and speech synthesis pipeline followed the approach described by Székely et al. [SHBG19b]. To detect breath events and segment the corpus, we used a simplified version of the speaker-dependent breath detector proposed in [SHG19]. This enabled us to segment the audio recordings into 3,487 automatically-detected *breath groups* (speech segments delineated by consecutive breaths). To create a synthesis corpus, these breath groups were paired up into bigrams, which allows synthesising longer utterances and for taking context into account across utterances. The voice was then built based on the Tacotron 2 spectrogram-prediction framework [SPW*18], using public resources for transcription, TTS, pre-training, phonetisation, and waveform synthesis, following [SHBG19b]. Breath events and hesitations (uh and um) were given a unique phone symbol.

4.2. Gesture generation

For full-body gesture and stance generation we used a technique called MoGlow [HAB19], adapted to gestures as described in [Aut]. The model is autoregressive, and learns the distribution for the next motion pose given previous poses and a context window of speech features from the corresponding speech audio. New motion is generated by sampling a sequence of random poses from this next-step distribution. This produces different (but plausible) gestures every time, even if the speech input is the same.

Setup and data processing followed [Aut]. The aligned gestures and natural speech acoustics were downsampled to 20 frames (poses and mel-spectra) per second. Poses were parameterised using joint rotations, complemented by the forwards, sideways, and angular displacement of the root (hip), to enable the system to generate small steps and changes in stance based on the input. The data was also augmented by lateral mirroring. We used the same


(a) Gestures from natural speech

(b) Gestures from synthetic speech

(c) Training data before mirroring

(d) Mirrored training data

Figure 1: Gesture space visualisation for model predictions from natural and synthetic speech, as well as for training data.

network size as in [Aut] and trained the model for 80,000 steps using a data dropout rate of 0.4 and a learning rate decaying from $lr_{\max} = 2 \cdot 10^{-3}$ to $lr_{\min} = 5 \cdot 10^{-4}$. At synthesis time, 20 fps mel-spectrum features from either natural or synthetic speech waveforms were fed into the system to sample synthetic pose sequences.

5. Evaluation and discussion

We demonstrate the efficacy of our proof-of-concept system in several ways. The enclosed video (supplementary material) of the system presenting itself shows the approach in action; it will be made publicly available conditional on paper acceptance. We also evaluate the gesture spaces induced by natural and synthetic speech input to the gesture generator, and study the alignment between speech input and sampled output gestures. The stimuli for these evaluations were three 12–16 s long, semantically-coherent utterances selected from the held-out data, each comprising three sequential breath groups. To generate synthetic speech, we transcribed the words and original breath locations from the selected utterances.

5.1. Gesture space comparison

Many data-driven gesture synthesis methods suffer from regression to the mean pose, making synthetic gestures more constrained and less vivid than the natural training gestures, as illustrated in [FNM19]. To check this, we visualise the gesture spaces of model-synthesised and natural gestures by overlaying a hip-centered upper-body skeleton of the gesticulating avatar extracted every 8th frame for 15 samples (12–16s long) from each of the three selected utterances. The result is shown in Fig. 1. We see that the gesture spaces for gestures synthesised from the model are virtually the same as the natural gesture space of the training data. This contrasts against other gesture-synthesis methods such as those visualised in [FNM19, Fig. 1]. Crucially, gestures generated from synthetic speech acoustics (Fig. 1a) essentially fill the same space as gestures synthesised from natural speech input (Fig. 1b). This suggests that our approach generalises well to synthetic speech and still occupies the entire gesticulation space that a human naturally would use (as also seen in our demonstration video). Making motion statistics, e.g., the spacial extent of gestures, match natural gesticulation is also important since these statistics correlate with the perception of personality traits and emotional states [CN19, KG10].

5.2. Speech-gesture alignment

A key aspect of our approach, where the gesture synthesis is driven by synthetic speech, is that it can generate co-speech gestures that align with speech timings. Since the approach is probabilistic, we can get a good impression of this alignment using only a single utterance but many random draws, as seen in Fig. 2, which focusses on the first 8 s of one of the selected stimulus utterances. The figure uses kernel density estimation to visualise the distribution of the top peaks in hand velocity – taken as indicators of peak gesticulation intensity – across 300 randomly sampled pose sequences for the same input audio. There is clear variation between realisations, just as human gestures never are the same every time. However, the gesticulation intensity is not uniformly random, but exhibits a consistent structure, with pronounced peaks near emphasised speech segments and low gesture activity near long pauses. A deterministic gesture-generation method would require us to analyse hundreds of different utterances to reliably see this kind of structure. The differences in velocity distribution between gestures driven by natural speech audio (blue, top pane) and gestures driven by synthesised speech (red, lower pane) in Fig. 2 are not unexpected, since the intonation and emphasis of the text-derived synthetic speech is not the same as in the natural speech. In general, synthetic-speech audio spectrograms are blurrier and less distinct than those derived from natural speech, which is an artefact of excessive averaging (oversmoothing), analogous to regression towards the average pose seen in [FNM19]. However, sequence-to-sequence TTS produces much more lively and less averaged intonation than earlier systems, which we hypothesise helps our approach generate consistent and distinct gesticulation peaks also when driven by synthetic speech.

6. Conclusion and future work

We have demonstrated multimodal synthesis of both speech and full-body co-speech gestures from the same input text, by chaining together state-of-the-art text-to-speech and speech-to-gesture systems. Using the latest developments in spontaneous speech synthesis enables gestures to be generated from convincing spontaneous-style TTS audio. This is crucial for appropriateness, since traditional TTS is built from speech read aloud, and humans typically do not gesticulate when reading. By using speech audio and motion capture sourced from a single person, we generate speech and gesture behaviours that match a real human individual. Future work will involve tighter integration between gesture and speech synthesis, by learning a single, unified model that generates both modalities simultaneously. The fact that convincing speech and gesticulation can be generated simply from text input promises to simplify many applications that rely on cumbersome, low-level animation control to produce acceptable gestures. Beside work in autonomous and embodied agents and social robots, the ability to control gesture style [Aut] and unplanned phenomena like speech disfluencies [SHBG19a] opens the door to gesture-perception research similar to [KG10], but with highly realistic stimuli.

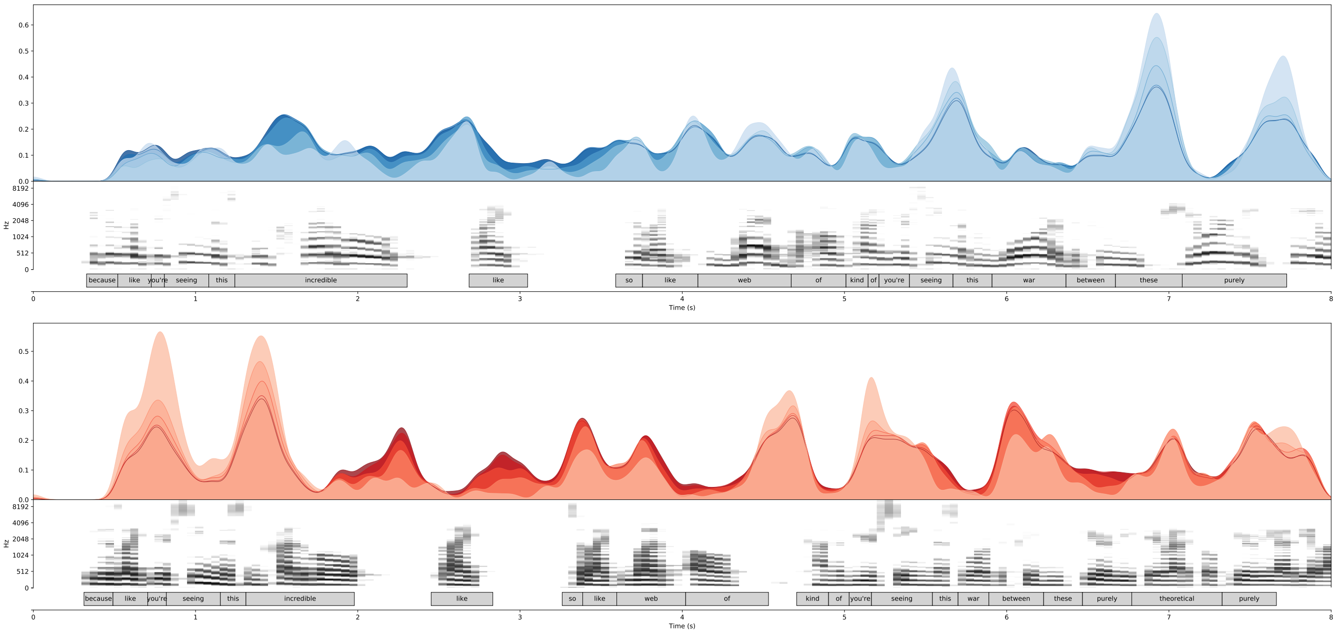


Figure 2: Velocity peak distribution. The colour plots illustrate how the top N ($2 \leq N \leq 12$) hand-velocity peaks of the synthesised motion are distributed for 300 pose sequences sampled for one natural (top/blue) and one synthetic (bottom/red) speech input utterance. Darker shades correspond to higher N. Below each distribution plot is the mel spectrogram (the system input features) and the timings of the spoken words.

References

[Aut] AUTHORS: Style-controllable speech-driven gesture synthesis using normalising flows. In *EUROGRAPHICS 2020 Submission ID 1201, Conditionally accepted for publication and supplied as additional material* `eg2020_submission_1201.pdf`. 1, 2, 3

[CBL*19] CUDEIRO D., BOLKART T., LAIDLAW C., RANJAN A., BLACK M. J.: Capture, learning, and synthesis of 3D speaking styles. In *Proc. CVPR* (2019), pp. 10101–10111. 2

[CN19] CASTILLO G., NEFF M.: What do we express without knowing?: Emotion in gesture. In *Proc. AAMAS* (2019), pp. 702–710. 3

[FM18] FERSTL Y., MCDONNELL R.: Investigating the use of recurrent motion modelling for speech gesture generation. In *Proc. IVA* (2018). 2

[FNM19] FERSTL Y., NEFF M., MCDONNELL R.: Multi-objective adversarial gesture generation. In *Proc. MIG* (2019). 3

[GBK*19] GINOSAR S., BAR A., KOHAVI G., CHAN C., OWENS A., MALIK J.: Learning individual styles of conversational gesture. In *Proc. CVPR* (2019), pp. 3497–3506. 2

[HAB19] HENTER G. E., ALEXANDERSON S., BESKOW J.: MoGlow: Probabilistic and controllable motion synthesis using normalising flows. *arXiv preprint arXiv:1905.06598* (2019). 1, 2

[HKS*18] HASEGAWA D., KANEKO N., SHIRAKAWA S., SAKUTA H., SUMI K.: Evaluation of speech-to-gesture generation using bi-directional LSTM network. In *Proc. IVA* (2018), pp. 79–86. 2

[IMMI18] ISHI C. T., MACHIYASHIKI D., MIKATA R., ISHIGURO H.: A speech-driven hand gesture generation method and evaluation in android robots. *IEEE Robot. Autom. Lett. 3*, 4 (2018), 3757–3764. 2

[KG10] KOPPENSTEINER M., GRAMMER K.: Motion patterns in political speech and their influence on personality ratings. *J. Res. Pers. 44*, 3 (2010), 374–379. 3

[KHH*19] KUCHERENKO T., HASEGAWA D., HENTER G. E., KANEKO N., KJELLSTRÖM H.: Analyzing input and output representations for speech-driven gesture generation. In *Proc. IVA* (2019). 2

[MV15] MATTHEYSES W., VERHELST W.: Audiovisual speech synthesis: An overview of the state-of-the-art. *Speech Commun. 66* (2015), 182–217. 2

[PNR*19] PAPAMAKARIOS G., NALISNICK E., REZENDE D. J., MOHAMED S., LAKSHMINARAYANAN B.: Normalizing flows for probabilistic modeling and inference. *arXiv preprint arXiv:1912.02762* (2019). 1

[RPCM18] RAVENET B., PELACHAUD C., CLAVEL C., MARSELLA S.: Automating the production of communicative gestures in embodied characters. *Front. Psychol. 9* (2018). 2

[SB19] SADOUGHI N., BUSSO C.: Speech-driven animation with meaningful behaviors. *Speech Commun. 110* (2019), 90–100. 2

[SHBG19a] SZÉKELY É., HENTER G. E., BESKOW J., GUSTAFSON J.: How to train your fillers: uh and um in spontaneous speech synthesis. In *Proc. SSW* (2019), pp. 245–250. 2, 3

[SHBG19b] SZÉKELY É., HENTER G. E., BESKOW J., GUSTAFSON J.: Spontaneous conversational speech synthesis from found data. In *Proc. Interspeech* (2019), pp. 4435–4439. 1, 2

[SHG19] SZÉKELY É., HENTER G. E., GUSTAFSON J.: Casting to corpus: Segmenting and selecting spontaneous dialogue for TTS with a CNN-LSTM speaker-dependent breath detector. In *Proc. ICASSP* (2019), pp. 6925–6929. 2

[SLB17] SADOUGHI N., LIU Y., BUSSO C.: Meaningful head movements driven by emotional synthetic speech. *Speech Commun. 95* (2017), 87–99. 2

[SPW*18] SHEN J., PANG R., WEISS R. J., SCHUSTER M., JAITLY N., YANG Z., ET AL.: Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions. In *Proc. ICASSP* (2018). 1, 2

[WMK14] WAGNER P., MALISZ Z., KOPP S.: Gesture and speech in interaction: An overview. *Speech Commun. 57* (2014), 209–232. 2

[YKJ*19] YOON Y., KO W.-R., JANG M., LEE J., KIM J., LEE G.: Robots learn social skills: End-to-end learning of co-speech gesture generation for humanoid robots. In *Proc. ICRA* (2019). 2