Automatic gesture generation for virtual humans with deep and temporal learning

Ylva Ferstl & Rachel McDonnell

Abstract

With increasingly sophisticated technical and visual design, virtual humans are now finding numerous applications, as instructors in virtual classrooms, for human-computer interfaces, in video games, and more. As their behaviour becomes more and more automated, a key challenge remains the generation of believable gestures that are tightly linked to the uttered content. Without appropriate non-verbal behaviour, virtual humans quickly appear oddly rigid, eerie, and unappealing. With our proposed framework, we aim to develop more life-like virtual conversational agents by providing a fully automatic system for gesture generation from live speech. Our model does not rely on hand-annotation of data, and bases the selection of gestural behaviour on prosodic, syntactic, as well as semantic analyses. Furthermore it is not restricted to a set of predefined gestural signs. A major strength of the framework is the utilization of deep learning for the association of speech with gestures with the inclusion of temporal relations between gestures.

Introduction

Virtual humans are becoming more and more popular for many applications, such as human-computer interfaces (e.g. virtual museum guides [1]) and personalized training (e.g. virtual patients for medical training [2]), including training of interpersonal skills. Coupling the verbal output with appropriate gesture not only makes interactions with these virtual agents more engaging [3], but is also essential for adequately mimicking real human interactions, in which non-verbal behaviour plays a major role in conveying information [4], [5]. Furthermore, users detect whether virtual human's gestures are consistent with the produced speech [6]. Creating such consistent and believable non-verbal behaviour for virtual agents is, however, a complex problem. Systems often resort to specifically creating animations for pre-defined utterances [7], [8]; procedurally generating appropriate gestures for live speech is much more challenging. While previous models have shown good results for simple beat gestures by relying on prosodic analysis [9], generation of more meaningful gestures has usually relied on incorporating handcrafted rules [10]–[12]. Going beyond beat gestures with a data-driven system, Chiu & Marsella (2015) [13] have published work that generates more complex gestures by using deep learning and temporal modelling without explicit rules. However, the authors rely on a set of pre-defined gestural signs, limiting the gesture output to this set. Furthermore, they require these gestural sign to be hand-annotated in a training set, limiting the amount of feasibly acquirable training data.

With our model, we extend previous approaches of gesture generation to a fully automatic system for meaningful gestures that does not rely on any hand-annotated data or handcrafted rules. Two key parts in our model are a deep learning structure with temporal modelling capacities, and the representation of natural motion in a low-dimensional space. We aim to utilize a recurrent neural network structure, merging the abilities of a classic deep network of learning complex relationships between multiple features, with the abilitiy to take context into account. The representation of motion in a lower-dimensional space reduces the complexity of the motion, allowing us to learn from the whole gestural space (instead of a set of pre-defined gestural signs).

Three types of speech features are fed into the system to choose appropriate gestures. The first are prosodic features, which have been shown to be associated with kinematic features of gestures [14], by reflecting emotional state [15] and intended emphasis [16], and previous gesture generation approaches have shown some success in utilizing them [10], [13], [17]. Secondly, we extract semantic features in order to be able to create meaningful gestures that go beyond beat gestures. We especially want to incorporate important markers such as negations, self-referrals, and direct addresses, but aim to explore beyond these as well.

Thirdly, we automatically mark each word with its associated part-of-speech tag that which may be associated with gesture style [18][19] and can help disambiguate word meanings.

Figure 1 shows an overview of the proposed system pipeline. The input to the system is live speech; this is forwarded to an audio processer that extracts prosodic features and a transcriber that converts the audio signal to text. From this transcript, a part-of-speech (POS) tagger labels the text with the grammatical role of each word, and a semantic processer uses this together with the transcript to deduct some semantic structures, such as negations. The prosodic and semantic features and the POS tags are then fed into a deep network.

The network combines the advantages of deep learning for learning the complex relationship between speech and gestures with the ability to represent the dynamic temporal relationship between sequences of gestures and speech. This kind of network has shown promise in previous works relating to human motion [20], [21], as well as speech [22], [23]. The network learns the association of the previously extracted speech features with motion features that represent gestures in a lower dimensional space.

The low-dimensional motion features are computed with Gaussian process dynamical models (GPDMs), as proposed by [24] and [25]. GPDMs are latent variable models that can learn the high-dimensional dynamics of motion capture data. Thus, the high-dimensional motion capture space of natural gestures is encoded in a low-dimensional manifold that then makes association learning of the motion and speech features more feasible. After learning the motion model in the training phase, the GPDM step can receive a batch of motion features from the deep network and map them back onto the full-dimensional motion space during execution. This recreated motion data is then sent to an animation system set up in the Unity game engine.

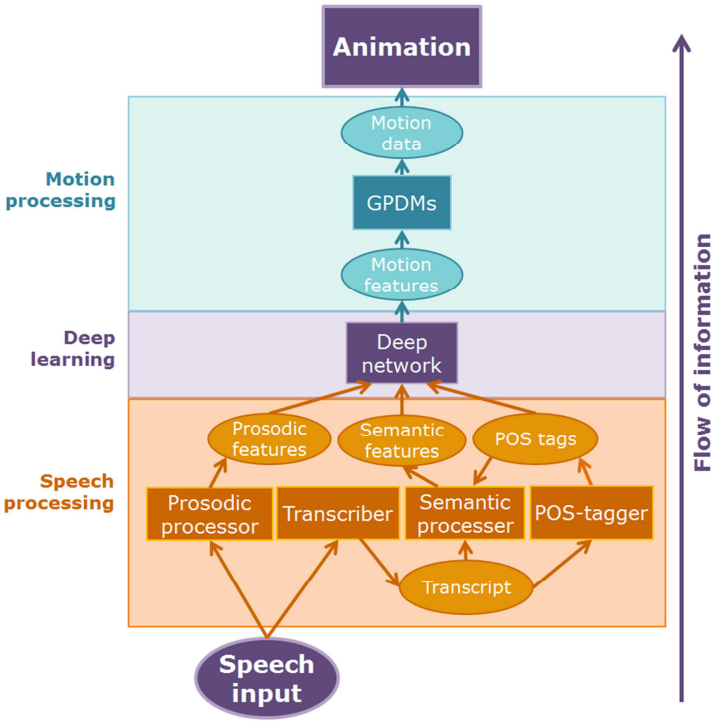

Figure 1: Overview over the gesture generation pipeline. The input into the system is live speech, from which prosodic features, semantic features, and POS tags are extracted. These features are fed into a deep CRF network, where they are associated with low-dimensional motion features. The motion features are mapped back onto the full-dimensional motion space and sent to the animation system.

In summary, our model extends previous systems by using training data that is completely automatically labelled, without needing any hand-annotation, increasing the amount of feasibly acquirable training data, it incorporates prosodic, semantic, as well as syntactic features, and it utilizes a powerful deep learning structure in order to go beyond simple beat gestures.

References

[1] W. Swartout, D. Traum, R. Artstein, D. Noren, P. Debevec, K. Bronnenkant, J. Williams, A. Leuski, S. Narayanan, D. Piepol, C. Lane, J. Morie, P. Aggarwal, M. Liewer, J. Y. Chiang, J. Gerten, S. Chu, and K. White, "Ada and grace: Toward realistic and engaging virtual museum guides," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 6356 LNAI, pp. 286–300,

2010.

[2] R. C. Hubal, P. N. Kizakevich, C. I. Guinn, K. D. Merino, and S. L. West, “The virtual standardized patient,” *Med. Meets Virtual Real.*, pp. 133–138, 2000.

[3] M. Salem, K. Rohlfing, S. Kopp, and F. Joublin, “A friendly gesture: Investigating the effect of multimodal robot behavior in human-robot interaction,” *Proc. - IEEE Int. Work. Robot Hum. Interact. Commun.*, pp. 247–252, 2011.

[4] R. Pally, “A Primary Role for Nonverbal Communication in Psychoanalysis,” *Psychoanal. Inq.*, vol. 21, no. 1, pp. 71–93, 2008.

[5] S. Goldin-Meadow, “The role of gesture in communication and thinking,” *Trends Cogn. Sci.*, vol. 3, no. 11, pp. 419–429, 1999.

[6] C. Ennis, R. McDonnell, and C. O’Sullivan, “Seeing is believing,” *ACM Trans. Graph.*, vol. 29, p. 91, 2010.

[7] “Gunslinger.” [Online]. Available: http://ict.usc.edu/prototypes/gunslinger/.

[8] D. DeVault, R. Artstein, G. Benn, T. Dey, E. Fast, A. Gainer, K. Georgila, J. Gratch, A. Hartholt, M. Lhommet, and others, “SimSensei Kiosk: A virtual human interviewer for healthcare decision support,” *Proc. 2014 Int. Conf. Auton. agents multi-agent Syst.*, no. 1, pp. 1061–1068, 2014.

[9] C. Chiu and S. Marsella, “Gesture Generation with Low-Dimensional Embeddings,” in *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*, 2014, pp. 781–788.

[10] S. Marsella, Y. Xu, M. Lhommet, A. Feng, S. Scherer, and A. Shapiro, “Virtual character performance from speech,” in *Proceedings of the 12th ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, 2013, pp. 25–35.

[11] M. Thiebaux, S. Marsella, A. N. Marshall, and M. Kallmann, “SmartBody: behavior realization for embodied conversational agents,” in *Proceedings of International Joint Conference on Autonomous Agents and Multiagent Systems*, 2008, no. Aamas, pp. 151–158.

[12] J. Cassell, H. H. Vilhjálmsson, and T. Bickmore, “BEAT: the Behavior Expression Animation Toolkit,” *ACM Trans. Graph.*, pp. 477–486, 2001.

[13] C.-C. Chiu and S. Marsella, “Predicting Co-verbal Gestures: A Deep and Temporal Modeling Approach,” *Int. Conf. Intell. Virtual Agents*, vol. 9238 of th, no. August 2015, pp. 152–166, 2015.

[14] L. Valbonesi, R. Ansari, D. McNeill, F. Quek, S. Duncan, K. E. McCullough, and R. Bryll, “Multimodal signal analysis of prosody and hand motion: Temporal correlation of speech and gestures,” *Proc. Eur. Signal Process. Conf. EUSIPCO 2002*, pp. 75–78, 2002.

[15] K. R. Scherer, R. Banse, H. G. Wallbott, and T. Goldbeck, “Vocal cues in emotion encoding and decoding,” *Motiv. Emot.*, vol. 15, no. 2, pp. 123–148, 1991.

[16] J. Terken, “Fundamental frequency and perceived prominence of accented syllables,” *J. Acoust. Soc. Am.*, vol. 89, no. 4, pp. 1768–1776, 1991.

[17] S. Levine, C. Theobalt, and V. Koltun, “Real-time prosody-driven synthesis of body language,” *ACM Trans. Graph.*, vol. 28, no. 5, p. 1, 2009.

[18] S. Goldin-Meadow, C. Butcher, C. Mylander, and M. Dodge, “Nouns and verbs in a self-styled gesture system: What’s in a name?,” *Cogn. Psychol.*, vol. 27, no. 3, pp. 259–319, 1994.

[19] L. J. Gogate, L. E. Bahrick, and J. D. Watson, “A Study of Multimodal Motherese: The Role of Temporal Synchrony between Verbal Labels and Gestures,” *Child Dev.*, vol. 71, no. 4, pp. 878–894, 2000.

[20] N. Nishida and H. Nakayama, “Multimodal gesture recognition using multi-stream recurrent neural network,” *Pacific-Rim Symp. Image Video Technol.*, pp. 682–694, 2015.

[21] Y. Du, W. Wang, and L. Wang, “Hierarchical recurrent neural network for skeleton based action recognition,” *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pp. 1110–1118, 2015.

[22] A. Graves, A. Mohamed, and G. Hinton, “Speech recognition with deep recurrent neural networks,” *Acoust. Speech Signal Process.*, no. 3, 2013.

[23] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, “Listen, attend and spell: A neural network for large vocabulary conversational speech recognition,” *Acoust. Speech Signal Process.*, pp. 4960–4964, 2016.

[24] J. M. Wang, D. J. Fleet, and A. Hertzmann, “Gaussian Process Dynamical Models for Human Motion,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 2, pp. 283–298, 2008.

[25] J. Wang, D. Fleet, and A. Hertzmann, “Gaussian process dynamical models,” *NIPS*, vol. 18, p. 3, 2005.