# Towards Multimodal Human-Like Characteristics and Expressive Visual Prosody in Virtual Agents

Mireille Fares

# Towards Multimodal Human-Like Characteristics and Expressive Visual Prosody in Virtual Agents

Mireille Fares
ISIR Lab and STMS-IRCAM Lab, CNRS, Sorbonne Université
Paris, France
fares@isir.upmc.com

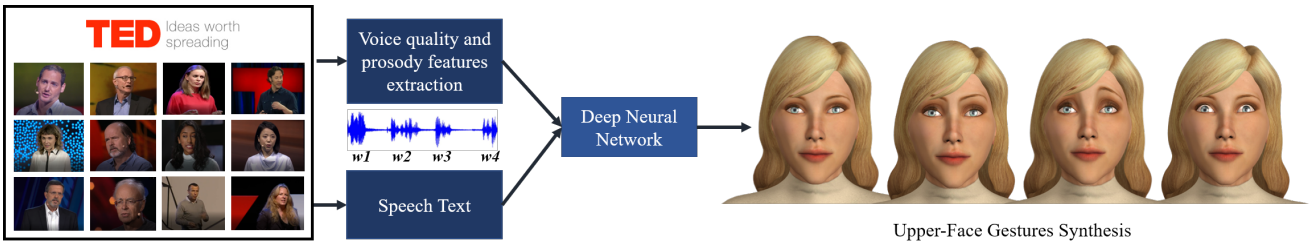


**Figure 1: A deep learning approach is used to generate upper-face gestures and is trained using facial gestures, audio features, and speech text extracted from TED talks.**

## ABSTRACT

One of the key challenges in designing Embodied Conversational Agents (ECA) is to produce human-like gestural and visual prosody expressivity. Another major challenge is to maintain the interlocutor's attention by adapting the agent's behavior to the interlocutor's multimodal behavior. This paper outlines my PhD research plan that aims to develop convincing expressive and natural behavior in ECAs and to explore and model the mechanisms that govern human-agent multimodal interaction. Additionally, I describe in this paper my first PhD milestone which focuses on developing an end-to-end LSTM Neural Network model for upper-face gestures generation. The main task consists of building a model that can produce expressive and coherent upper-face gestures while considering multiple modalities: speech audio, text, and action units.

## KEYWORDS

Multi-modality; Upper-face expressivity; Visual prosody; Speech; Neural Networks; Embodied Conversational Agents

## 1 INTRODUCTION

The first form of communication in the lifespan of humans is non-verbal communication. Before humans evolved their ability to speak and use language, they were able to communicate using non-verbal channels of communication [21]. All non-verbal cues are involved in a Human-Human Interaction (HHI): body, face, voice, appearance, touch, distancing, timing, and other physical cues. Non-verbal behaviors convey tremendous information to the interlocutors. One important channel of communication in HHI is the human face. Humans use their gaze to convey their desire to switch speaking turns, and their hands movements to express their thoughts [5]. A variety of verbal, emotional, and conversational cues are displayed on the face while interacting. As a matter of fact, during speech, humans continually employ various facial gestures, known as "visual prosody"[16], which is a form of facial or head movement produced in conjunction with verbal communication. Facial gestures are consciously or unconsciously used to adjust speech, accentuate words or word segments, or mark speech pauses. These gestures involve different head movements, blinks, eyebrow gestures, gaze, frowning, nose wrinkling or lips moistening [39]. Speech driven facial gestures are associated with prosody and para-linguistic information. Speech prosody refers to various speech characteristics like intonation, rhythm, and stress. Para-linguistic information refers to the cues, which can be used to convey emotion, such as pitch, volume and speech intonation.

My PhD thesis aims to examine and understand the mechanisms that govern a human-agent multimodal interaction, and generate convincing expressive, coherent and human-like behavior in Embodied Conversational Agents (ECA). For this purpose, my work focuses on the ECA's gestural and visual prosody expressivity, as well as on making the ECA capable of adapting its behavior to its interlocutor's multimodal behavior. As a first step in this work, we focus on developing a coherent human-like facial expressivity in ECAs. In particular, we focus on developing a model that predicts expressive upper facial movements such as eyebrows and eyelids movements. The prediction will be based on different modalities which are audio data, text data, and specific facial muscles data. To the best of my knowledge, predicting upper-face movements based on all the previously discussed different modalities has not yet been investigated. As a starting point, we will develop an end-to-end LSTM neural network architecture that predicts upper-face gestures, based on both audio data and text data. Recurrent Neural Networks (RNN) and Long Short-Term Memory (LSTM) architectures are used to exploit the temporal dependencies in audio data and model the prosody variations. The overall architecture is illustrated in Figure 1. We consider multimodal cues when modelling the ECA's behavior, since data coming from multiple sources improve RNNs, produce complementary information, and convey patterns that are not discernible when working with individual modalities.

## 2 BACKGROUND AND RELATED WORK

In this work, we are interested in the generation of sequences of upper-face gestures based on multimodal input sequences such as speech audio and text. Speech related facial gestures are associated with prosody and paralinguistic information. In this section we recall major works related to psychological and paralinguistic research, as well as gesture generation systems.

### 2.1 Psychological and paralinguistic work

Eyebrow motion and speech have been shown to be strongly correlated by Ekman [14]. His findings show that eyebrow movement happen during thinking pauses, or to emphasize a word or sequence of words. Eyebrows are either raised or lowered when the speaker is thinking. According to [11], eyebrows movements are the most relevant and frequent facial gestures that are used during conversations. In [10] they investigate the relation between fundamental frequency (F0) variations and eyebrow movements. On one hand, their findings show that F0 variations and eyebrow movements are highly correlated during speech. On the other hand, they conclude that eyebrow movements also occur during pauses. Therefore, F0 and eyebrow movements are not directly linked, but they are the results of linguistic and conversational choices. They are also used to reassure the speaker that the attention of his/her listener is still captivated. Furthermore, according to [10], eyebrows also mirror the listener's amount of understanding, and can be used as a backchannel.

### 2.2 Gesture generation systems

There have been major efforts in developing gesture synthesizing systems for virtual agents. Several works have been done for hand gesture generation, expressive upper/lower face animation, and head motion generation. In [9], they work on generating expressive facial movement synchronized with the audio of input utterances. In [30], they generate lower facial movements based on a deep learning approach that uses a sliding window predictor that learns nonlinear mappings from phonemes to mouth motion. Kucherenko[22] proposes a speech driven gesture generation by learning a lower dimensional representation of human movements using a denoising autoencoder neural network, and an encoder network that maps speech to movement representation with a low dimensionality. Their findings show that mel-frequency cepstral coefficients (MFCC) used alone or along with other prosodic speech features performs the best for generating gestures. In [38], they propose a facial gesture generation system for ECAs. Lip movements are produced based on input speech signal. In their work, virtual speakers can read given text and transform it into the corresponding speech and facial gesturing. Hofer [20] proposes a speech driven head motion sequence prediction based on Hidden Markov Models (HMM). Another technique for speech driven head motion synthesis is proposed in [18], they use deep neural networks with stacked bottleneck features, along with an LSTM network. Marioooryad and Busso [25] built a facial animation framework based on speech to generate head and eyebrows motion using Dynamic Bayesian Networks. Lu and Shimodaira[24] propose an approach of head motion prediction that is based on speech waveforms. They assert that it is more powerful to use waveforms directly to predict head motion, without extracting any features. They propose a canonical correlation constrained autoencoder to extract the relevant data from waveforms, minimize the errors, and maximize the canonical correlation with head movements. In [13], they propose an animation model based on HMM based on statistical model that links speech prosody with facial gestures. In [34], they produce videos of talking heads based on a person's image, and audio speech. They generate lip motion that is in sync with speech, as well as facial expressions like blinks and eyebrow motion. Their approach is based on GAN that uses three discriminators which goal is to produce reasonable expressions, audio-visual synchronization, and video frames.

## 3 STATEMENT OF THESIS

The present paper is part of a thesis that aims to better understand and model the mechanisms that govern human-machine multimodal interaction. It aims to address the following research questions:

(1) `Expressivity`: How to generate the ECA's visual prosody expressivity, and more specifically its upper-face gesture expressivity, as well as its head motion, based on speech ?
(2) `Adaptation`: How to generate an appropriate behavior for the ECA so that it is adapted to the multimodal behavior of its interlocutor? How to leverage the ECA's expressivity to achieve a good adaptation of its behavior along with its interlocutor's behavior?

## 4 RESEARCH METHODOLOGY

As discussed in the previous section, this work will be built upon two main pillars. First, we will model the agent's visual prosody expressiveness using RNNs that are currently commonly used for modelling voice and gestures [29, 35, 36]. The RNN will be used to model gestural variability at the sentence level, or the whole text. The agent's coherent multimodal behavior will be learned using multimodal attention mechanisms applied on the gestures that are generated. Second, we will focus on rendering the agent's behavior coherently adapted to that of its interlocutor using interactive and imitation learning. We will explore different model architectures and therefore train them using different databases such as TEDx [31], Noxi [7], and IEMOCAP [6], to make the agent learn the adaptation during the interactions between the different interlocutors. At the time of this writing, we are developing a model that predicts expressive upper-face gestures. We plan to predict the sequence of Action Units that correspond to upper-face muscles. Different architectures will be explored such as RNNs, and transformers [33]. As a first step, we decided to go for RNN architectures since RNN have a strong performance in sequence modeling [19], and they have the capacity of integrating temporal contextual information. We are using Bi-directional LSTM networks cells in RNN, since they are good in modeling time-dependent and sequential data [19].

## 5 UPPER-FACE GESTURES SYNTHESIS

To address the first research question discussed earlier, we propose a learning-based upper-face gestures generation model that is learned from a series of TEDx talks [31]. For this purpose, our work is based

on Facial Action Coding System (FACS) [15], a system that describes the facial movements based on 44 Action Units (AUs). The model architecture is an end-to-end neural network model that consists of several encoders to encode the input features, and a decoder to generate the sequence of AUs that are related to eyebrow and eyelid movement. The following subsections describe the details of our chosen dataset, our chosen features, the preprocessing phase, as well as the overall network architecture.

## 5.1 TEDx Dataset

TED (Technology, Entertainment, Design) conferences are conferences where people share their major research or ideas from multiple disciplines with their audience. Using TED talks as our dataset has many advantages. First of all, these talks contain myriad of presentation topics, each presented by a unique speaker. Each speaker has his/her communicative style, and all of them have the same goal which is to captivate the audience. TED talks are well recorded and last long. The speakers' speech and gestures are highly expressive, intense, and energized. Therefore, we expect that the speakers use expressive facial expressions, hand/arm gestures, speech, and body postures. A great advantage of using this database is that videos come with their corresponding transcript. TEDx talks [31] were obtained from YouTube. We collected the same talks that were used in [37], who worked on the generation of hand gestures. In total, we collected 1760 videos with their transcripts which include the timestamps of segments of words. In our case, we are only interested by the shots where each speaker's face is visible enough. To minimize the errors, videos were segmented into shots using PySceneDetect [4], and filtered under the following conditions:

- Speaker's face is visible
- Speaker's face is not far from the camera (not small)
- One face is in the shot (speaker's face)
- Speaker is facing the camera that is in front of her/him
- No still pictures in the shot

Video segmentation and filtering was also done in [37], but in our case, since we are working on facial gestures, we are interested in video segments that meet the previously stated conditions. In the case of [37], they were interested in the shots containing visible hand gestures. Their conditions are different, and therefore their dataset is different.

## 5.2 Data representation and preprocessing

In this work, we consider several modalities for our model: speech audio, text, and AUs. The following features were selected to be used for each modality:

- Action Units features: The AUs that represent eyebrows and eyelids movements are AU01, AU02, AU04, AU05, AU06, and AU07. The AUs intensities features were extracted using OpenFace [3] from the video segments that we have previously selected for our TEDx dataset. In OpenFace, each generated AU intensity is given a "Success" score (0 or 1), and a "Confidence" level (between 0 and 1). For each AU intensities of each shot of interest, we applied a median filter with a window size equals to 7, to remove noises and deal with the cases where the confidence is low. As discussed earlier, video segments last less than 5 seconds, which allowed us to apply linear interpolation on the resulting data to deal with the cases where OpenFace results were not successful (Success = 0). Giving that AUs values are continuous, they were quantized to generate a finite range of discrete integers. In fact, in deep learning, quantized representations are used to highly reduce the model size and energy consumption, by storing weights using a compact format such as integers instead of floating numbers [17]. In this work, we used autoencoders for dimensionality reduction. An autoencoder compresses the input data into a lower-dimensional representation, and then converts it back to a reconstruction of the original input. The encoder and decoder of the autoencoder were trained jointly but they are used independently. The encoder part of the network is used to compress the AUs intensities into a lower-dimensional representation, which is fed as an input to our model architecture.
- Audio features: The audio features that we are considering in our model are prosodic features, voice quality features, as well as Mel Frequency features. F0 was extracted using SWIPE estimator [8]. Voice quality features which were proved to improve the expressiveness [26] are: Jitter, Shimmer, Harmonic-to-Noise Ratio, and the Hammarberg index. The latter features were extracted with OpenSmile [1]. IrcamAlign [23] was used to perform alignment for speech signals into phones and diphones, providing a confidence level for each phone. It was also used to extract the phonological structure such as syllables, words and breath sequences from the resulting aligned sequence of phones. As a starting point, and for simplicity, only word-level F0 sequences are considered as input audio features to our model. The other features will be considered later on. We considered the F0 values with a confidence level greater than or equal to 0.3. The ones with confidence level less than 0.3 were replaced by values of 0. In addition to that, for each sequence of F0s corresponding to a word, we applied linear interpolation and extrapolation in order to get a complete sequence of non-zero F0 values. Since F0 values are continuous, they were quantized to generate a finite range of discrete integers. The vocal speech of a typical adult male has a F0 ranging from 85 to 180 Hz. That of a typical adult female ranges from 165 to 255 Hz[2, 32]. In this work, F0 values were restricted to the range of 50 to 550Hz, which is enough to enclose the vocal ranges of both male and female speakers. As done with AUs, we trained an autoencoder that compresses the input F0 data into a lower-dimensional representation, and then converts it back to a reconstruction of the original input. The encoder and decoder of the autoencoder were trained together but are used separately. The low dimensional representation of F0s generated by the encoder is fed into the input of our model.
- Text features: Speech text is represented by a sequence of words, and each word is encoded as a BERT embedding [12]. BERT is the first deeply bidirectional unsupervised language representation, that was trained using a large corpus of sentences, and it produces powerful representation of words: its embeddings are jointly conditioned on both left and right contexts simultaneously, unlike other tools [27] [28].

### 5.3 Network Architecture

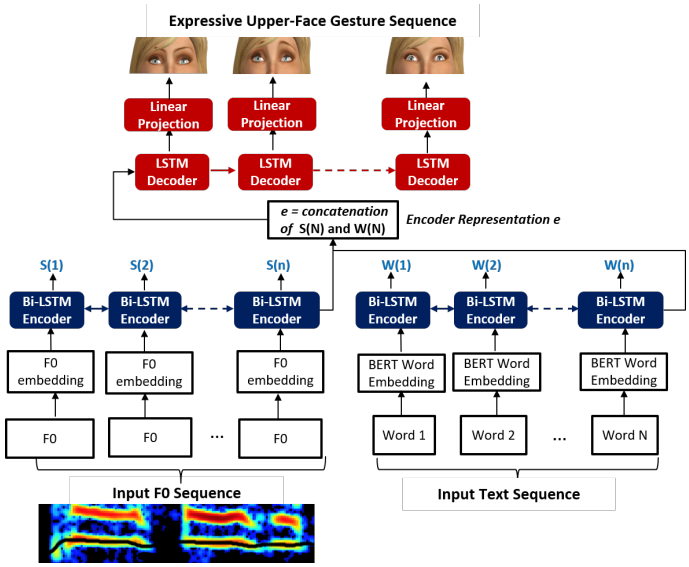


**Figure 2: Sequence to Sequence Network Architecture.**

Our problem consists of mapping a sequence of words, and a sequence of F0s to a sequence of action units AU01, AU02, AU04, AU05, AU05, AU06, and AU07. This problem is similar to the neural machine translation problem which consists of mapping a sequence of words to another sequence of words in another language. The model used in such problems is called Sequence to Sequence (Seq2Seq). As a starting point, we use the Seq2Seq model to address our research question of visual prosody expressivity: we propose an end-to-end LSTM Neural Network model to predict eyebrows and eyelids movements based on sequences of words and F0. The overall architecture of the proposed method is illustrated in Figure 2. The input features of the model are the following: word embeddings produced by BERT, as well as F0 embeddings generated by the Encoder network of the AutoEncoder discussed earlier. The inputs size of the model is equal to the Inter-Pausal Unit (IPU) which includes the words said before and after pauses that last longer than 200 milliseconds. The encoders process the input words and the input F0s one by one. The results are transmitted to the decoder to generate sequences of AUs that are related to the corresponding eyebrow and eyelid movements. The decoder network of the AUs AutoEncoder we discussed earlier is used in this model to predict the AUs given their embedding vector, and therefore reconstruct the corresponding AUs. The decoder is followed by post-linear layers. The model is implemented with an attention mechanism, which is not depicted in Figure 2 for simplicity. All the elements that were discussed in this section were implemented. The next steps are to train the architecture on our dataset.

## 6 FUTURE WORK AND TIMELINE

The research plan and timeline for this dissertation are as follows:

- In year 2020, we plan to complete and train the Seq2Seq model for upper face movement prediction. We plan to test our model with all speech input features that we previously discussed. Furthermore, we will test other model architectures such as the transformers.
- In year 2021, we plan on developing models to generate expressive and coherent head motion. The audience in TEDx is seated in a semi-circular fashion, and therefore TEDx speakers tend to continuously move their heads. This is why we plan to train the models using other datasets such as [7] [6]. In addition to that, we plan on developing a model to produce appropriate behavior for the ECA, adapted to the multimodal behavior of its interlocutor. We will be conducting several experiments to evaluate each of our developed models.
- Starting the end of year 2021 onward, we plan to wrap up the dissertation work by adding any missing components and start writing the overall PhD thesis.

## 7 EXPECTED CONTRIBUTIONS

We envision the following major contributions for this thesis:

(1) Development of models for multimodal expressive visual prosody synthesis intended to be used with ECAs.
(2) Development of specific models to control ECAs' behavior to make it adapted to the multimodal behavior of their interlocutors.

## ACKNOWLEDGMENTS

I would like to thank my PhD advisors, Prof. Catherine Pelachaud, and Prof. Nicolas Obin for their continuous help and support throughout this work. This work was performed within the Labex SMART (ANR-11-LABX-65) supported by French state funds managed by the ANR within the Investissements d'Avenir programme under reference ANR-11-IDEX-0004-02.

balance

## REFERENCES

[1] [n.d.]. OpenSMILE. https://www.audeering.com/opensmile/
[2] Ronald J Baken and Robert F Orlikoff. 2000. *Clinical measurement of speech and voice.* Cengage Learning.
[3] Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. 2016. Openface: an open source facial behavior analysis toolkit. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV).* IEEE, 1–10.
[4] Breakthrough. 2020. Breakthrough/PySceneDetect. https://github.com/Breakthrough/PySceneDetect
[5] Judee K Burgoon, Laura K Guerrero, and Valerie Manusov. 2016. *Nonverbal communication.* Routledge.
[6] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. IEMOCAP: Interactive emotional dyadic motion capture database. *Language resources and evaluation* 42, 4 (2008), 335.
[7] Angelo Cafaro, Johannes Wagner, Tobias Baur, Soumia Dermouche, Mercedes Torres Torres, Catherine Pelachaud, Elisabeth André, and Michel Valstar. 2017. The NoXi database: multimodal recordings of mediated novice-expert interactions. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction.* 350–359.
[8] Arturo Camacho and John G Harris. 2008. A sawtooth waveform inspired pitch estimator for speech and music. *The Journal of the Acoustical Society of America* 124, 3 (2008), 1638–1652.
[9] Yong Cao, Wen C Tien, Petros Faloutsos, and Frédéric Pighin. 2005. Expressive speech-driven facial animation. *ACM Transactions on Graphics (TOG)* 24, 4 (2005), 1283–1302.
[10] Christian Cavé, Isabelle Guaïtella, Roxane Bertrand, Serge Santi, Françoise Harlay, and Robert Espesser. 1996. About the relationship between eyebrow movements and Fo variations. In *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96,* Vol. 4. IEEE, 2175–2178.

[11] Nicole Chovil. 1991. Discourse-oriented facial displays in conversation. *Research on Language & Social Interaction* 25, 1-4 (1991), 163–194.

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[13] Yu Ding, Catherine Pelachaud, and Thierry Artieres. 2013. Modeling multimodal behaviors from speech prosody. In *International Workshop on Intelligent Virtual Agents*. Springer, 217–228.

[14] P Ekman. 1979. About brows: emotional and conversational signals in Human Ethology (eds M. von Cranach, K. Foppa, W. Lepenies & D. Ploog) 169–248.

[15] Rosenberg Ekman. 1997. *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS).* Oxford University Press, USA.

[16] H. P. Graf, E. Cosatto, V. Strom, and Fu Jie Huang. 2002. Visual prosody: facial movements accompanying speech. In *Proceedings of Fifth IEEE International Conference on Automatic Face Gesture Recognition*. 396–401.

[17] Yunhui Guo. 2018. A survey on methods and theories of quantized neural networks. *arXiv preprint arXiv:1808.04752* (2018).

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

[19] Sepp Hochreiter and Jürgen Schmidhuber. 1997. LSTM can solve hard long time lag problems. In *Advances in neural information processing systems*. 473–479.

[20] Gregor Hofer and Hiroshi Shimodaira. 2007. Automatic head motion prediction from speech data. (2007).

[21] Mark L Knapp, Judith A Hall, and Terrence G Horgan. 2013. *Nonverbal communication in human interaction.* Cengage Learning.

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

[23] Pierre Lanchantin, Andrew C Morris, Xavier Rodet, and Christophe Veaux. 2008. Automatic Phoneme Segmentation with Relaxed Textual Constraints.. In *LREC*.

[24] JinHong Lu and Hiroshi Shimodaira. 2020. Prediction of head motion from speech waveforms with a canonical-correlation-constrained autoencoder. *arXiv preprint arXiv:2002.01869* (2020).

[25] Soroosh Mariooryad and Carlos Busso. 2012. Generating human-like behaviors using joint, speech-driven models for conversational agents. *IEEE Transactions on Audio, Speech, and Language Processing* 20, 8 (2012), 2329–2340.

[26] Carlos Monzo, Ignasi Iriondo, and Joan Claudi Socoró. 2014. Voice quality modelling for expressive speech synthesis. *The Scientific World Journal* 2014 (2014).

[27] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1532–1543.

[28] Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365* (2018).

[29] Carl Robinson, Nicolas Obin, and Axel Roebel. 2019. Sequence-to-sequence Modelling of F0 for Speech Emotion Conversion. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6830–6834.

[30] Sarah Taylor, Taehwan Kim, Yisong Yue, Moshe Mahler, James Krahe, Anastasio Garcia Rodriguez, Jessica Hodgins, and Iain Matthews. 2017. A deep learning approach for generalized speech animation. *ACM Transactions on Graphics (TOG)* 36, 4 (2017), 1–11.

[31] TEDxTalks. [n.d.]. TEDx Talks. https://www.youtube.com/user/TEDxTalks

[32] IR Titze. 1994. *Principles of Voice Production*. Prentice-Hall Inc.

[33] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*. 5998–6008.

[34] Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. 2019. Realistic speech-driven facial animation with gans. *International Journal of Computer Vision* (2019), 1–16.

[35] Xin Wang, Shinji Takaki, and Junichi Yamagishi. 2017. An RNN-Based Quantized F0 Model with Multi-Tier Feedback Links for Text-to-Speech Synthesis.. In *INTERSPEECH*. 1059–1063.

[36] Yuxuan Wang, Daisy Stanton, Yu Zhang, RJ Skerry-Ryan, Eric Battenberg, Joel Shor, Ying Xiao, Fei Ren, Ye Jia, and Rif A Saurous. 2018. Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis. *arXiv preprint arXiv:1803.09017* (2018).

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

[38] Goranka Zoric, Karlo Smid, and Igor S Pandzic. [n.d.]. Automated Gesturing for Embodied Animated Agent: Speech-driven and Text-driven Approaches. *Journal of Multimedia* 1, 1 ([n. d.]).

[39] Goranka Zoric, Karlo Smid, and Igor S Pandzic. 2007. Facial gestures: taxonomy and application of non-verbal, non-emotional facial displays for embodied conversational agents. *Conversational Informatics: An Engineering Approach* (2007), 161–182.