2024-03

Data-driven Communicative Behaviour Generation: A Survey

Oralbayeva, N

https://pearl.plymouth.ac.uk/handle/10026.1/22435

10.1145/3609235
ACM Transactions on Human-Robot Interaction
Association for Computing Machinery (ACM)

Data-Driven Communicative Behaviour Generation: A Survey

NURZIYA ORALBAYEVA, Department of Robotics and Mechatronics, School of Engineering and Digital Sciences, Nazarbayev University, Kazakhstan

AMIR ALY, School of Engineering, Computing and Mathematics, University of Plymouth, United Kingdom

ANARA SANDYGULOVA*, Department of Robotics and Mechatronics, School of Engineering and Digital Sciences, Nazarbayev University, Kazakhstan

TONY BELPAEME, Ghent University, IDLab - imec, Belgium

The development of data-driven behaviour generating systems has recently become the focus of considerable attention in the fields of human-agent interaction (HAI) and human-robot interaction (HRI). Although rule-based approaches were dominant for years, these proved inflexible and expensive to develop. The difficulty of developing production rules, as well as the need for manual configuration in order to generate artificial behaviours, places a limit on how complex and diverse rule-based behaviours can be. In contrast, actual human-human interaction data collected using tracking and recording devices makes human-like multimodal co-speech behaviour generation possible using machine learning and specifically, in recent years, deep learning. This survey provides an overview of the state-of-the-art of deep learning-based co-speech behaviour generation models and offers an outlook for future research in this area.

CCS Concepts: • **Computer systems organisation** → Robotics.

Additional Key Words and Phrases: datasets, neural networks, data-driven behaviour generation

ACM Reference Format:

Nurziya Oralbayeva, Amir Aly, Anara Sandygulova, and Tony Belpaeme. 2021. Data-Driven Communicative Behaviour Generation: A Survey. In . ACM, New York, NY, USA, 38 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Recent years have seen an increase in the development of systems for the generation of human-like communicative behaviour. This is driven by the need for socially interactive virtual and robotic agents in various domains. For instance, artificial agents may range from household service robots to museum guide avatars and social robots in education and medicine, whose primary function is not only to assist people but to connect with people through effectively producing social signals [13].

Research has long established a rule-based approach as an advantageous one in human behaviour generation [12, 109, 141]. However, in light of state-of-the-art developments, major issues in the rule-based approach have been identified. While it is efficient in producing human behaviours for a single or a limited number of modalities, its is hampered by the need for explicitly formulating rules, resulting in a practical limit on the number of rules, which in turn curbs the expressiveness of behaviour [62]. Additionally, rule-based systems typically fall short of producing multimodal behaviours, as the number of rules increases rapidly when new modalities are added [170]. Recent evidence

*Corresponding author: anara.sandygulova@nu.edu.kz

Manuscript submitted to ACM

suggests that rule-based models seem to fail when producing natural variations of human behaviour, often because they do not cover the entire range of behaviour or their naturalness is found to be lacking [125].

In contrast, models that are trained by learning from available corpora of speech, text, audio, and multimodal data allow for a more robust human-agent interaction, as they can learn correlated behaviour which is difficult or labour-intensive to capture in rules. For example, it is believed that computational models based on data hold promise in uncovering the complex relationships between verbal and non-verbal human behaviours [124, 218]. Advances in the deep learning and machine learning models, and the availability of large datasets have led to a growing interest in data-driven systems for behaviour generation [85, 111, 228], dialogue systems [173], and speech synthesis systems [197, 211]. The data-driven approach to interaction design is deemed to improve on the labour-intensive rule-based approach. Human behaviours are generally produced through various modes that make communication multimodal [7]. Those are primarily speech and different types of bodily gestures such as facial gestures, movements of the head, and manual (hand, arm, shoulder) gestures [7]. These all play an integral role in conveying social signals and information [147]. Moreover, the affective states of an interlocutor are consciously or unconsciously communicated by means of these verbal and non-verbal communicative channels [7]. Data from several studies suggest that robots and virtual agents able to cause affect in human users are perceived as more vivid and human-like [54, 160].

Compared to other recent reviews [127, 226], this survey intends to take stock of the dynamically expanding field of co-speech gesture and behaviour generation for anthropomorphic agents, and of the methodological approaches used for the evaluation of such models. We review existing research on data-driven approaches in verbal and non-verbal human behaviour generation and cover progress in data-driven communicative behaviour generation from the last five to six years. Furthermore, this work attempts to identify challenges and directions, and in doing so sets a road-map for future research in this field.

Section 2 explains the methodology for the review. Sections 3, 4, 5, 6 and 7 are dedicated to reviewing data-driven models, generating various communicative behaviours that occur in human-human interactions and designed for human-agent and human-robot interaction scenarios. Section 7 finishes the review and focuses on speech synthesis, the communicative behaviour in which most resources have been invested for arguably the longest period of time and which therefore holds essential lessons for data-driven behaviour generation. Section 8 provides an outlook for the field and concludes the paper.

2 MATERIALS AND METHODS

This paper reviews empirical studies published within the past five to six years (2014-2021), with some exceptions for studies published between 2011 and 2012, and which were considered relevant for this survey. Moreover, reference lists of the selected articles and significant review papers were examined to identify other relevant studies for inclusion. A list of research keywords used in this work are summarized in Table 7 (Appendix A).

A total of 825 records were retrieved from various publication databases. The search result statistics across databases (i.e., Google Scholar, Scopus, Web of Science, ACM, IEEE) can be seen in Figure 1. After retrieving meta-data about the papers, the titles and abstracts of all 825 articles were screened to identify the journal articles and conference papers deserving a full-text review. Papers were withheld when containing appropriate keywords and model descriptions. The number of articles was reduced to 534 after the exclusion of overlapping titles and abstracts. Thus, a total of 291 publications were carried over to the full-text review stage.

During the full-text review only publications were included according to the following criteria, where a work:

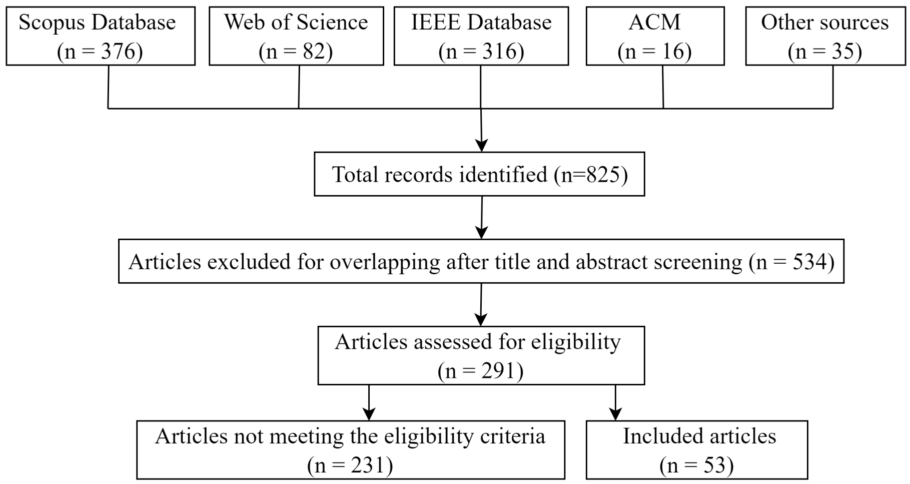

Fig. 1. Literature selection process.

- introduced a model with the capability of training (which in most cases was a neural network);
- relied on a corpus or dataset for training;
- presented clear evaluation metrics;
- presented test-bed platforms for the proposed models.

A paper was excluded if:

- it was focused solely on rule-based systems;
- it did not describe the evaluation metrics;
- it did not provide information on the dataset and corpora for training and validation.

As a result, of 291 works that were considered in the full-text review, 231 works with no evaluation metrics or corpora were excluded. Among them were articles describing rule-based models, which were out of the scope of this survey and hence were removed from the review. The final list of publications thus contained 53 papers meeting the eligibility criteria. The selected papers are organized according to the type of behaviour presented in separate sections in this survey. Note that we are agnostic about the form of the agent on which the behaviour is produced: this survey focuses on the generation of behaviours for both humanoid and non-humanoid robots as well as virtual conversational agents and avatars.

3 HEAD GESTURES

Head gestures constitute an important part of human body language during communication and co-occur with speech. Speech-driven head gesture synthesis through data-driven approaches has attracted attention over the last decade. Unlike rule-based models for gesture synthesis, data-driven models can learn dependencies between data so as to map a sequence of speech features to meaningful head animations. The related literature shows different frameworks employing Deep Neural Networks (DNNs) [184], Bi-directional Long Short-Term Memory (BLSTM) networks [172], and deep generative models [72, 179], which are capable of learning the temporal and cross-modal dependencies of continuous signals.

Ding et al. [45] discussed a Deep Neural Network (DNN) for synthesizing head motion from speech features. To this end, they pre-trained a Deep Belief Network (DBN) [89], using stacked Restricted Boltzmann Machines (RBMs) [178] with a target layer for fine-tuning the DBN model parameters, creating a DNN model. The objective evaluation criteria depend on three measures: Canonical Correlation Analysis (CCA) [83], Average Correlation Coefficient (ACC) [159], and Mean Square Error (MSE) [6] for the differences between predicted head movements with respect to ground truth movements, where the results show that the generative pre-trained DNN model outperformed the randomly initialized network trained through back propagation. Furthermore, Ding et al. [47] showed that this DNN model outperformed a traditional Hidden Markov Model (HMM) approach for head motion synthesis from speech [91] in the CCA analysis.

Ding et al. [46] compared two types of neural network models, BLSTM and feed-forward networks, to learn the correspondences between speech and head motion. The results show that the BLSTM model significantly reduced the root mean squared error (RMSE) – of predicted movements with respect to ground truth movements – compared to that of the feed-forward model that does not converge when the number of hidden layers is bigger than two. Furthermore, the BLSTM model, with different numbers of hidden layers, achieves a better performance than that of the feed-forward model in the Canonical Correlation Analysis (CCA) [83]. Over and above, a hybrid network composed of two BLSTM layers and one feed-forward layer in between, shows a higher performance in objective evaluations and in subjective evaluation - measuring the naturalness of head motion - than a separate BLSTM model and the other stacked network architectures.

Haag and Shimodaira [82] presented a bottleneck Deep Neural Network (DNN) architecture, where bottleneck features – resulting from a DNN model containing a hidden bottleneck layer and trained on the features of speech and head motion – are used with speech features as input to another DNN model with a BLSTM layer in a forward pass in order to synthesize head motion. These bottleneck features can capture the dependencies between the features of speech and head motion curves, which allows for improving the accuracy of generating head movements. They report that bottleneck features enhanced the performance of the DNN-BLSTM architecture and achieved better scores in the Canonical Correlation Analysis (CCA) [83] than when they were not present in the architecture.

Greenwood et al. [77] introduced a Bi-directional Long Short-Term Memory (BLSTM) model to predict head motion from speech and further extended the model through conditioning by a prior motion input in order to limit the possible head motion predictions for speech. Moreover, they proposed a generative Conditional Variational Autoencoder (CVAE) [179] using BLSTM models as encoder and decoder to map speech to head motion. This last model allows for predicting a variety of output head motion curves for the same speech input by sampling from the Gaussian space and conditioning on speech features.

Sadoughi and Busso [165] presented a conditional Generative Adversarial Network (GAN) [72] with BLSTM cells for generating head movements for speech segments. It learns, during training, the conditional distributions of head motion curves and prosodic features of speech. The performance of the proposed model was compared with a Dynamic Bayesian Network (DBN) [132] and a BLSTM model [46]. The results show that the proposed conditional GAN model outperformed of the baseline DBN and BLSTM models in terms of the log-likelihood measures as well as in subjective evaluation.

Table 1. Corpora[1] and evaluation used in the head gesture generation literature

	Corpus[2]			Evaluation	
	Source	**Training**	**Test**	**Objective**	**Subjective**
Ding et al. [45]	Audio-visual dataset from NBC newscast	93 mins from a target news presenter, 120 mins from other 10 presenters	10 mins from the target presenter	Canonical Correlation Analysis (CCA) [83], Average Correlation Coefficient (ACC) [159], and Mean Square Error (MSE) [6]	N/A [3]
Ding et al. [46]	The MNGU0 articulatory corpus [158]	1137 utterances from a single speaker	63 utterances from the single speaker	CCA [83]	A/B preference test (naturalness) [108]
Haag and Shimodaira [82][4]	The University of Edinburgh dataset [81]	N/A[5]	N/A	CCA [83]	MOS (naturalness) [156]
Greenwood et al. [77][6]	Audio-visual dataset collected by the authors	1440 utterances from one actor (~144 mins)	180 utterances from the actor (~18 mins)	N/A	N/A
Sadoughi and Busso [165]	The IEMOCAP database [20]	38 mins from one actor	14 mins from the actor	Log-likelihood measures [64]	Questionnaire, A pairwise comparison

Table 1 summarizes the related information to the corpora and evaluation approaches used in the studies covered in this survey. While most of these studies considered objective measures to evaluate the proposed models, some of them had subjective evaluations. It is noteworthy that the sizes of the corpora and the scale of evaluations are often small; therefore, measuring how appropriate the generated head gestures is not always possible, and new metrics supplementing the existing objective metrics might be needed.

Summary: Head Gestures

- Different data-driven models can be used for successfully generating expressive head motion from speech, all are likely to achieve a satisfactory level of subjective and objective performance.
- Speech and audio representations for head gesture generation are provided in a number of different features, such as acoustic (e.g., mel frequency cepstral coefficients (MFCC) [45, 46, 82], linear prediction coefficients (LPC), the lower representation of speech - FBank), articulatory [82], and prosodic (e.g., frequency and intensity of speech) [165].
- Defining a credible metric for the quality and appropriateness of the generated head motion is still an open challenge.
- The size of the training and test corpora are generally limited, which could affect the quality of the generated gestures. Creating larger corpora for head gesture generation is likely to be a good investment.

4 FACIAL EXPRESSIONS

The human face is an important channel for non-verbal communication [61]. Most research has focused on facial animation to express facial affect (or emotions) Pantic [146], and typically use the facial Action Units (AU) schema by

[0] The reporting of dataset durations for training and test splits from different works in this table and hereinafter was constrained by their availability.
[1] The reporting of dataset durations for training and test splits from different works in this table and hereinafter was constrained by their availability.
[2] Each of the following datasets has been processed by the authors to extract the characteristics of speech and head motion in order to train the proposed models, except in Ding et al. [46] and Sadoughi and Busso [165] where audio-visual data and features are provided [20, 158].
[3] **Not applicable**, w.r.t the evaluation metric, a particular metric is not applied in the work.
[4] The authors did not provide clear information on the size of the training and testing data.
[5] Dataset sizes are **not available**.
[6] Greenwood et al. [77] did not use any objective or subjective measures. Instead, they discussed the characteristics of the generated head motion with respect to the ground truth.

Ekman *et al.* to present facial animations in a numerical manner [50]. Along with the basic emotional model suggested by Ekman, Facial Action Coding system (FACS) [51] – a systematic method for describing and measuring facial movements in response to emotions – is leveraged as a common representation of facial affect in most of the works on facial expression generation. Researchers consider such facial modalities as the gaze, eyebrow actions, head motion [132] or eye behaviour, mouth, eyebrows, nose, the shape of the face, cheeks, wrinkles, neck and even hair [190] and lip motion Mancini et al. [130] to contribute to the facial behaviour and expression generation. While the majority of studies consider facial expressions in close relation to emotions [25, 164], elsewhere research focuses on facial units regardless of emotions, using the term facial gestures [53, 61]. Generally, facial expression generating models are based on Dynamic Bayesian Networks (DBN) [132], Generative Adversarial Networks [72] and Long Short-Term Memory (LSTM) [90]. In this survey, facial expression generation is discussed in two subsections, distinguishing natural facial behaviours (such as blinking, lip-syncing, etc.) and affective facial expressions.

4.1 Natural Facial Expressions

The following works center around the facial expressions deemed "independent of facial expressions of emotions" such as raising an eyebrow, winking, shaking the head [53] or blinking and frowning [206].

Taylor et al. [188] proposed to use a Sliding Window Deep Neural Network (SW-DNN) [103] to generate lip movements using the Mel-frequency Cepstral Coefficients (MFCCs) of the speech input from the audio-visual KB-2k [189] speech dataset. The model was benchmarked against the HMM inversion (HMMI) [66] and was also evaluated subjectively for perceived realism alongside ground truth (GT) and HMMI, determining the average response rate. As a result, the SW-DNN model achieved optimal results in generating the output of lip movements and mouth shapes.

van der Struijk et al. [202] developed a generative FACSvatar [7] framework for modelling virtual avatars' facial animation based on Facial Action Coding System (FACS) [161] data. The framework enables a data-driven generation of facial animation through a simple Gated Recurrent Unit (GRU) neural network implemented with Keras[8]. Input data was obtained through OpenFace2, which, from FACS-based [51] input, sent AU eye gaze and head rotation to ZeroMQ in real-time. The subjective evaluation results regarding the generation of facial configurations demonstrated that the DNN model in the machine learning module requires further improvements. Moreover, the performance of the FACSvatar framework was tested on several modules, such as CSV offline, Bridge, AU to Blend Shapes, Visualisation in Unity 3D and Machine Learning. The main limitation of this framework is the shortage of datasets with different AU intensities, which seems to impede the machine learning process.

Jonell et al. [99] proposed a probabilistic method to generate interlocutor-aware facial expressions using four modalities: an interlocutor's acoustic features and facial features as well as the avatar's acoustic features and existing facial features. Although the model resembles the MoGlow [87, 105], it differs by using multiple modalities and encoding each modality by separate networks, such as Multi-layer Perceptrons (MLPs), Recurrent Neural Networks (RNNs) and 1D-convolution networks (CNNs). As an objective measurement, the authors used log-likelihood and its ablations as well as mismatched sequences. As for the subjective evaluation metrics, a user study used a single question across five experiments with the participants on their perceptions of the system. The experimental results demonstrated the significance of multimodal input in generating appealing facial expressions in response to the interlocutor.

[7] A framework which adds and processes data based on Facial Action Coding System (FACS) [161] in real time.
[8] See keras.io

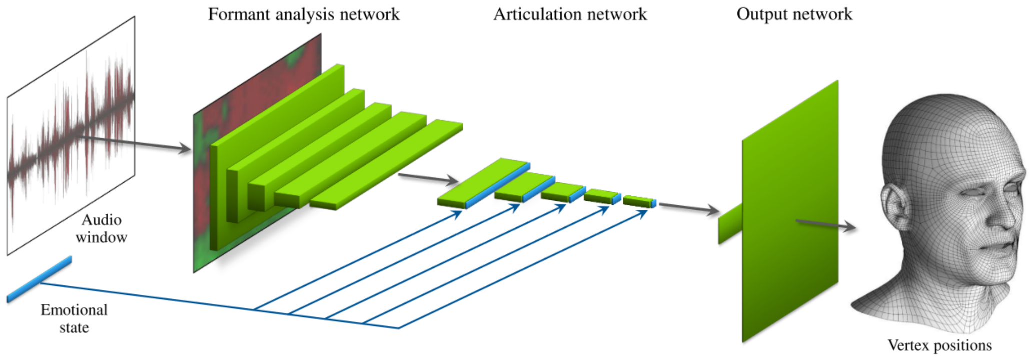

Fig. 2. An illustration of a deep neural model used for generating facial expressions using speech as input, from Karras et al. [101]. The network takes two types of input: half a second of audio and a description of an emotional state. The former (audio) is used to output the 3D vertex positions of a fixed-topology mesh that correspond to the center of the audio window, while the latter (emotional state) disambiguates facial expressions and speaking styles.

4.2 Affective Facial Expressions

This subsection focuses on expressive facial animation generation. Research into the affective facial expression generation in the domain of Embodied Conversation Agents (ECA) has produced some seminal works, such as those by [101, 164], to name but a few. In the following paragraphs, we elaborate on works that consider emotion information, such as the six universally recognized emotions suggested by [52] – happiness, sadness, disgust, anger, fear, and surprise – in the design of facial expression generation models.

Karras et al [101] presented a model based on a deep neural network to generate expressive 3D facial animations from speech audio (Fig. 2). The emotional states were presented as E-dimensional vectors [9] fed to the network as a secondary input. The performance of the proposed model was compared in a subjective user study against video-based performance capture from the DI4D [10] system and dominance model-based animation produced by FaceFX [11] [39] as baselines. While the proposed model was outperformed in the naturalness of the output facial animations by the video-based performance capture model, it showed an outstanding performance over the dominance model. The major shortcoming of the proposed model was caused by its inability to represent eye motion due to mismatches with the audio. Therefore, combining the proposed approach with generative neural networks would provide a better synthesis of such details. While the model succeeded to produce plausible results for several emotional states (e.g., amused, surprised), a larger dataset might be useful to advance the model further.

Huang and Khan [94] introduced a Dyadic Generative Adversarial Network (DyadGAN) model to generate a partner-aware facial expression response in dyadic conversations with a virtual agent. The DyadGAN model follows two stages of GAN; one generates sketch images conditioned on the facial expressions of an interviewee, while the other generates real facial expressions of an interviewer. Experiments with two quantitative metrics - calculating facial expression features and canonical expression descriptors - revealed the model's ability to generate consistent facial expressions with movements from right to left. The overall results demonstrated that the generated interviewer response was consistent with the interviewees' emotions (i.e., joy, anger, surprise, fear, contempt, disgust, sadness, and neutral).

[9] E is a tunable parameter representing an emotional state to the output of each convolution layer.
[10] www.di4d.com
[11] An audio-based facial animation generating system, See www.facefx.com.

However, the authors emphasize directions for further improvements of the model in terms of using a larger dataset with multiple interviewers to enable the generalisation to different identities. Another way of enhancement would be combining the proposed model with a temporal recurrent network, namely, LSTM [90] to obtain video frames of facial expressions.

Sadoughi and Busso [164] presented a BLSTM [232] trained with speech features (i.e., Mel-frequency Cepstral Coefficients (MFCCs)) and the *extended Geneva minimalistic acoustic parameter set* eGeMAPS [57] for emotional speech-driven lip motion generation designed specifically for conversational agents. The proposed approach relied on multitask learning (MTL) [12], which created shared representations for the tasks. The study results were measured objectively through single task learning (STL) [13] and MTL comparison and benchmarked against state-of-the-art baselines [163, 188]. Moreover, the subjective evaluation used Tukey's multiple comparisons test to assess the naturalness of the lip movements. The results demonstrated the advantage of MTL in the generation of lip movements corresponding to the original sequences, achieving the naturalness of animation. It is noteworthy that the MTL-based framework can be trained on partial information (i.e., without necessitating the full labelling of data).

Sadoughi and Busso [167] proposed a Conditional Sequential Generative Adversarial Network (CSG) model that learns the relationships between emotion, lexical content and lip movements using the sceptral and emotional speech features as conditioning inputs to generate expressive and naturalistic lip movements. Compared against three DNN-based baselines [59, 163, 188] with the Parzen estimator [72], the model displayed higher log-likelihood and outperformed other baselines in the objective evaluation. The subjective evaluation results showed a better performance for the CSG model in terms of the naturalness of the generated lip motions. The generated lip movements were also evaluated for their ability to convey emotional cues, manifesting that the CSG model allows conveying expressive cues close to the original recordings.

Otberdout et al. [144] proposed a conditional version of the manifold-valued Wasserstein Generative Adversarial Network [9] to generate facial expressions of six basic emotions [52] from an image of neutral facial expression. To evaluate the model both qualitatively and quantitatively, [144] utilized the Oulu-CASIA [14] [234], MUG Facial Expression [4], and the Extended Cohn Kanade (CK+) [129] datasets. Objective metrics as Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity (SSIM) [213], Inception Score (IS) [16, 80], Average Content Distance (ACD) [15] [193] and its variant ACD-I [16] [235] were used to evaluate the model's performance. The results of both the objective evaluation and comparison with the baselines (MoCoGAN[193], VGAN[205], TGAN[169]) showed that the proposed model outperforms the state-of-the-art in video facial expression generation.

Table 2 presents the summary of the corpora and evaluation metrics used in natural and affective facial expression generation. Corpora-wise, there seems to be large diversity in datasets to train models. In terms of representations, while some opted for Action Units [25], others relied on readily available large databases of facial expressions [61, 94, 202]. Nevertheless, dataset sizes are not always consistent and sufficient for the completely smooth performance of a model.

[12] A strategy that jointly solves related secondary tasks.
[13] A strategy that focuses on solving a primary task only.
[14] A dataset containing 480 videos of basic emotion labels performed by 80 subjects.
[15] ACD measures the content consistency of the generated video based on how well the video preserves identity of the input face [144].
[16] The average distance between each generated frame and the original input frame.

Table 2. Corpora and evaluation used in the facial expression generation literature

	Corpus			Evaluation	
	Source	Training	Test	Objective	Subjective
Taylor et al. [188]	KB-2k audio-visual speech dataset [189]	2300 sentences	100 sentences	MSE [6]	Forced binary choice test [171]
Karras et al. [101]	The emotion database [101]	5min 1s (9034 frames) for Character 1, 3min 45s (6762 frames) for Character 2	57 seconds (1734 frames), 29 seconds (887 frames)	N/A	A/B preference test [108]
Huang and Khan [94]	Dyadic video interviews of 31 students [94]	24 hours of video (1000 short video clips)	N/A	Euclidean distance [56, 148]	N/A
Sadoughi and Busso [164]	The IEMOCAP database [20]	106 sentences	20% of the whole dataset	Concordance Correlation Coefficient (CCC) [163, 192] & Mean Squared Error (MSE) [6]	Questionnaire (10-point Likert scale) using Amazon Mechanical Turk (AMT)
Sadoughi and Busso [167]	The IEMOCAP database [20]	1,898 samples	recordings with 617 speaking turns	Parzen window-based density estimation [72]	Questionnaire (naturalness)
van der Struijk et al. [202]	The MAHNOB Mimicry Database [14]	12 hours (32 recordings)	2.4 hours (6 recordings)	N/A	Questionnaire (5-point Likert scale & open questions)
Jonell et al. [99]	MAHNOB Mimicry database [14] with spontaneous dyadic conversations	9.5 hours[17]	0.74 hour[18] (6.5% of the total dataset)	Log-likelihood values [64] of the model using unmodified and mismatched test sequences	Questionnaire (perception)
Otberdout et al. [144]	Oulu-CASIA dataset [234] MUG-Facial Expression database [4] Extended Cohn Kanade (CK+) dataset [129]	80% of the dataset (384 videos) 1400 videos 327 sequences	20% of the dataset (96 videos)	Geodesic distance between the generated expression dynamics, Inception Score (IS) [80], Peak Signal-to-Noise Ratio (PSNR) [213], Structural Similarity (SSIM) [213], Average Content Distance (ACD) [144], ACD-I [235].	N/A

Summary: Facial Expressions

- Data-driven production of facial expressions, also known as facial gestures, has focused on creating natural (neutral) and affective facial expressions.
- Application domains vary significantly and range from the games industry to HRI.
- In terms of representation, some approaches opt for high-level Facial Action Units and audio-visual features [25], while others rely on readily available large databases of facial expressions [61, 94, 202]. Yet, there is an overall lack of more sophisticated datasets, i.e. with a high spatial and temporal resolution, emotional audio-visual data.
- There is a lack of sophisticated expressive animation rendering toolkits for off-the-shelf production of facial expressions [167].

5 HAND GESTURES

As a natural mode of interaction, hand gestures carry important functions in human-human communication, such as maintaining an image of a concrete or abstract object and idea (iconic and metaphoric gestures), pointing and giving directions (deictic gestures), or emphasizing some parts of the speech (beat gestures) [134]. Hand gestures, including fingers and arms, also act as an independent modality or part of modalities designed for various virtual agents and robots, adding expressivity to their motions. This versatility of hand gestures served as an incentive for their application in such domains as human-computer interaction (HCI) [207] and its related fields - human-robot interaction (HRI) [128] and human-agent interaction (HAI). In HRI, hand gestures are applied to socially assistive robots (SARs) because of the expressivity they add to robots' verbal and non-verbal communication with humans [170]. Besides, hand gestures are believed to ease the interaction between humans and robotic agents [142].

A considerable amount of research has been conducted on a data-driven generation of hand gestures, utilizing various databases and displaying a range of architectural choices [113, 194, 228]. For example, the earliest work by Chiu and Marsella [29] in 2011 made use of Hierarchical Factored Conditional Restricted Boltzmann machines (HFCRBMs) [30], whereas the most recent works resorted to models such as Long Short-Term Memory networks [85, 186] and a Variational Autoencoder (VAE) [111], to mention a few. Despite their purely communicative nature, sign language gestures are not covered in this survey as they rely solely and largely on a visual modality. Thus, in the paragraphs that follow, we cover the hand gestures that are characteristic of co-speech communication of information.

Chiu and Marsella [29] relied on Hierarchical Factored Conditional Restricted Boltzmann machines (HFCRBMs) [30] – an extension of Deep Belief Network [89] – to generate hand gestures that are tied to prosodic information. In particular, the gesture generator function learns the relationship between previous motion frames, audio features (inputs) and current motion frame (output) to generate hand gesture animations. The model was trained on motion capture and audio data from human conversation. Particularly, the motion capture data contained joint rotation vectors with 21 degree of freedom, whereas audio features used prosodic information such as pitch and intensity values. During the subjective evaluation, three animation types – Original, Generated, and Unmatched – were compared against each other in a user study. The results demonstrated the naturalness of the movements of generated gesture animations and the consistency of the motion dynamics with utterances.

Bozkurt et al. [17] presented a speaker-independent framework for joint analysis of hand gestures with continuous affect attributes, such as activation, valence, and dominance, and speech prosody using Hidden semi-Markov models (HSMMs) [230]. Moreover, during the synthesis step, prosody feature extraction and continuous affect attributes are followed by the HSMM-Viterbi algorithm. Gestures in motion capture data were represented by joint angles of arms and forearms. Consequently, the animation is generated via unit selection applied on a gesture pool with regard to a multi-objective cost function. Their system was trained on multimodal USC CreativeIT database [135]. Phrase-level gesture sequences for 1) affect and prosody feature fusion, 2) prosody only, and 3) affect only configurations were evaluated based on Canonical Correlation Analysis (CCA) scores [83] and symmetric Kullbeck-Leibler (KL) divergence. Their findings suggest that affect and prosody fusion provides the best correlation with the original gesture trajectories, and has the best gesture and gesture duration modeling. On the other hand, affect only configuration has the least kinetic energy difference with the original sequence. Subjective evaluations were planned for their future work.

Takeuchi et al. [186] used deep neural networks with Bi-directional Long Short-Term Memory (BLSTM) [232] to study the production of metaphoric hand gestures from speech features of audio. During the data pre-processing, the hand gestures were represented as rotations of bone joints. The network is composed of three non-recurrent layers, a BLSTM layer, and a final output layer. The first non-recurrent layer takes Mel-frequency Cepstral Coefficients (MFCCs) features of audio as input, while other non-recurrent layers take independent data. On the other hand, the final output layer takes the backward and forward recurrence units from the BLSTM layer as input. Thus, the model output - the vector of prediction - is represented in a BioVision Hierarchy (BVH) format. The objective evaluation, conducted by comparing

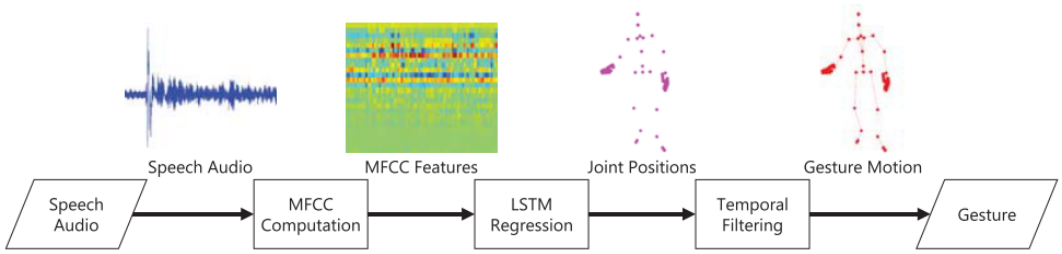

Fig. 3. The outline of the network architecture presented by Hasegawa et al. [85] consisting of five layers.

the final loss results from the proposed model with a simple Recurrent Neural Networks (RNN) implementation, resulted in significantly better performance of the proposed model. The subjective evaluation of the original, mismatched, and generated gestures demonstrated significantly lower ratings of the generated gestures than the former two (original and mismatched) in terms of naturalness, matching in timing, and context. This result, as the authors explain, might be affected by the gesture motion's frequent moving.

Hasegawa et al. [85] presented the BLSTM model integrating it with Bi-directional Recurrent Neural Networks (RNN) [75] to generate co-speech 3D metaphoric hand gestures from speech audio. Specifically, speech audio features were converted to mel frequency cepstral coefficients (MFCC) features and the joint positions of a whole body were used to represent the gestures. The network learns the relationship between speech and audio with backward and forward consistencies. Similar to the model proposed by Takeuchi et al. [186], the architecture consists of five layers shown in Figure 3. The objective evaluation was performed through Average Position Error (APE)[19][117], which displayed insignificant errors in the left and right wrists in terms of accuracy. Moreover, the user study revealed that the generated gestures among the three gesture conditions (original, mismatched, and generated) were perceived as significantly more natural but significantly less time and semantically consistent than original gestures.

Kucherenko et al. [112] presented a novel speech-input and gesture-output Deep Neural Network (DNN) framework consisting of two steps. First, the network learns the lower dimensional representation of human motion with a denoising autoencoder neural network. Then, an encoder network *SpeechE* learns a mapping between speech and a corresponding motion representation. Kucherenko et al. [112] applied representation learning on top of the DNN model to make learning from speech and speech-to-motion mapping easier. The objective evaluation compared the proposed network with the baseline BLSTM model presented in Hasegawa et al. [85] using Average Position Error (APE) [20] [117] and Motion Statistics[21] as metrics for the average distance between the generated and original motion as well as the average values and distributions of acceleration and jerk, respectively. The proposed model achieved better results compared to the baseline and demonstrated the plausibility of the generated gestures. A further validation of the results through a user study confirmed the model's performance in terms of producing natural gestures.

Ginosar et al. [70] presented a model based on Convolutional Neural Network with General Adversarial Network (CNN-GAN) and log-mel spectrogram input, which can predict and generate hand gestures from a large dataset of speech audio [70]. For gesture representation, the authors used skeletal keypoints corresponding to the neck, shoulders, elbows, wrists and hands, which were obtained through OpenPose [24]. The network learns to map speech to gesture

[19] **APE** compares the predicted positions with the original ones that accompany speech and calculates the Euclidean distance.
[20] Ibid., p. 10
[21] The average values and distributions of acceleration and jerk for the produced motion.

using L1 regression, while the adversarial discriminator D ensures that the produced motion is plausible. Using the L1 Regression Loss and percent of correct keypoints (PCK) [225] as objective evaluation metrics, it was discovered that the proposed model outperformed an RNN-based baseline [176] in gesture generation. Besides, the extent to which the produced gestures were convincing was measured through a perceptual study applying the percentage of the generated sequences, labelled as real, as a metric. The result of the comparison between fake (produced by an algorithm) and real pose sequences did not display any statistical significance.

Yoon et al. [228] deployed a Bi-directional Recurrent Neural Network (RNN) model consisting of an encoder and decoder for co-speech gesture generation from speech text input. More specifically, the encoder takes the input text, while the decoder RNN with pre- and post-linear layers generates gestures. The model was trained on the TED Gesture Dataset [228] to produce four common types of gestures - iconic, metaphoric, deictic, and beat gestures - from both trained and untrained speech texts. A gesture is represented as a sequence of human poses, namely, joint configurations of the upper-body. As for the speech text, it is represented as a sequence of words, and each word is encoded as a one-hot vector that indicates the word index in a dictionary. The results indicated that anthropomorphism and speech-gesture correlation were the most crucial factors for participants' perception of the generated gestures, as demonstrated in the subjective evaluation. The results also showed significance over the three baseline methods measured with BLEU [22] [149]. While the study used only speech text resulting in the weak coupling of the gestures with audio, it could be improved with audio input.

Ferstl et al. [63] attempted to map speech to 3D gestures through training networks with multiple adversaries to generate co-speech gestures. The authors extracted MFCC and pitch emphasis (F0) from the recorded speech and used upper-body joint positions to represent the gestures. The model architecture consists of a two-layer recurrent network composed of Long Short-Term Memory [90] cells and a feed-forward layer for input processing. Moreover, a Gated Recurrent Unit (GRU) [32] propagates the input for faster training purposes in producing joints. The novelty of the model lies in the training of the recurrent network with multiple generative adversaries instead of a standard regression loss. Drawing on the objective evaluation measured by the accuracy of the binary cross-entropy objective for each discriminator, the authors report the effectiveness of discriminators in solving a distinct sub-problem in the gesture generation task.

Tuyen et al. [194] employed a conditional extension of the Generative Adversarial Network (CGAN) [72] with an additional input condition. The GAN network includes convolutional Generator (**G**) and Discriminator (**D**) networks. Altogether, the model generates communicative gestures by synthesizing the verbal content of speech. Here, the gestures were represented as human joint configurations. The objective evaluation was carried out through covariance with temporal hierarchical construction [95]. Overall, the results illustrated the successful training of the model to imitate hand gestures that corresponded to the meaning of an utterance, which matched the iconic gestures by definition [134].

Lee et al. [118] introduced a temporal neural network, trained with Inverse Kinematics (IK) loss to generate finger motions and hand gestures taking upper body joint angles and audio as input from a multimodal *16.2-million-frame* (16.2M) dataset [118], created alongside the model. The audio features included frequency (e.g., pitch, jitter), energy, amplitude (e.g., shimmer, loudness), and spectral features. The IK was applied to LSTM [90], Variational Recurrent Neural Network (VRNN) [35], and Temporal Convolutional Network (TCN) [198] to incorporate kinematic structural knowledge. The ablation study results demonstrated the advantages of IK loss function contrary to joint angle loss,

[22] A method for automatic evaluation of machine translation.

whereas the subjective evaluation yielded positive results with respect to the proposed model and its capability to generate natural human-like finger gestures.

Table 3. Corpora and evaluation used in the hand gesture generation literature

	Corpus			Evaluation	
	Source	**Training**	**Test**	**Objective**	**Subjective**
Chiu and Marsella [29]	Conversational dataset [55]	38 seconds (1140 frames)	53 seconds (1591 frames)	N/A[23]	Motion-speech matching task
Bozkurt et al. [17]	USC CreativeIT database [135]	recordings of 15 actors [24]	recordings of 1 actor (2-10 minutes)	Canonical Correlation Analysis (CCA) [83]; symmetric Kullback-Leibler (KL) divergence [115]	N/A
Takeuchi et al. [186]	Gesture-speech dataset [187]	106.95 minutes (530 sentences)	9.69 minutes (59 sentences)	Comparison of final loss to the baseline RNN results	Questionnaire (7-point Likert scale)
Hasegawa et al. [85]	Gesture-speech dataset [187]	143 minutes[25] (767 sentences)	16 minutes[26] (90 sentences)	Average Position Error (APE) [117]	Questionnaire (naturalness, time consistency, and semantic consistency)
Kucherenko et al. [112]	Gesture-speech dataset [187]	171 minutes	20 minutes	Average Position Error (APE) [117]	Rating of statements on 7-point Likert-scale (naturalness, time consistency, and semantic consistency)
Ginosar et al. [70]	Person-specific video dataset [70]	115.2 hours	14.4 hours (2048 intervals)	L1 Regression Loss[27] and percent of correct keypoints (PCK) [224]	Questionnaire (real vs. fake), pairwise comparison
Yoon et al. [228]	TED Gesture Dataset [228]	52 hours	N/A [28]	N/A	Questionnaire (anthropomorphism by Godspeed, likeability, speech-gesture correlation)
Ferstl et al. [63]	Natural speech and 3D motion dataset [63]	3.75 hours (226 minutes)	6.5 minutes	Accuracy of the binary cross-entropy objective	N/A
Tuyen et al. [194]	KIT whole-body motion database [131]	20 optical markers in 3D	5, 136 usable annotation samples	Covariance with temporal hierarchical construction [95]	N/A
Lee et al. [118]	16.2-million-frame (16.2M) dataset [118]	120 minutes of multimodal data	N/A	MSE[29] [6]	Questionnaire (richness of motion, naturalness, personal motion characteristics, 5-point Likert scale)

Table 3 presents the summary of the corpora and evaluation metrics employed in the studies above. The majority of studies relied on both objective and subjective evaluation criteria, while a few studies either used objective [194] or subjective evaluation criteria [96, 228]. To sum up, the works reviewed here demonstrate the prevalence of speech input data among data modalities used for hand gesture generation. Model-wise, recent research [63, 85] shows a comprehensive exploration of recurrent networks to capture the dynamics of human motion, which excel at solving gesture generation tasks. That being said, an omnipresent limitation of such models lies in the dearth of gesture-rich datasets required to enable a robot to produce a wide range of hand gestures as opposed to certain predefined gestures produced with sparse datasets [29]. Interestingly, the training and test sets used in [29] seem arguable considering the

[25]Not applicable, ibid., p. 5
[26]Each recording lasts about 2-10 minutes [135]
[27]The authors used L1 regression loss as a quantitative evaluation metric to compare the model's performance against the baselines.
[28]Not applicable, ibid., p. 5
[29]As a quantitative measure, the authors computed MSE values.

training and test set sizes used in other works. Thus, the following section reviews the existing state-of-the-art on models that consider other body parts along with hands, hence outputting appropriate behaviours.

Summary: Hand Gestures

- Data-driven generative models for hand gestures aim to generate four types of gestures – beat, deictic, iconic and metaphoric – but struggle with the latter two as semantics are often poorly modelled.
- Hand gesture production relies on input which can consist of text, prosody, affect or contextual information, or a combination of some or all of these. Hand gestures are typically represented by joint rotations [29], joint angles of arms and fore-arms [17], rotations of bone joints [186], joint positions of a whole body [85], skeletal keypoints [70], human pose sequences [228], upper-body joint positions [63], joint configurations [194], upper-body and finger joints [118]. Speech and audio features are mostly represented as acoustic (e.g., MFCCs, pitch, jitter) [85, 112, 118], prosodic (e.g., pitch, intensity, confidence to pitch) [17, 29, 63, 112], phonemic features [186], verbal content of speech [194], and energy and amplitude [118].
- The generated gestures often look natural, but the match to the spoken content is not yet good enough. Generating semantically matched hand gestures remains a challenge.
- Two important limitations are the scope of datasets and the lack of diversity. Most studies use single-speaker datasets, with English being the dominant language across corpora. Interactive applications would benefit from dyadic or multiparty datasets. Cultural diversity and appropriateness would benefit from datasets from other languages and cultures.

6 MULTIMODAL GESTURES

In this survey, we define multimodal gestures when referring to the multimodality of the output. In particular, we refer to the interpretation of multimodal output by Rojc et al. [160], who emphasized the importance of synchronisation of generated non-verbal gesture types (facial expressions, head, hands, and body) with verbal (speech audio or video) in an attempt to make the interaction more natural and fluent. Therefore, the generation of such multimodal outputs as *head and facial movements synchronized with speech* [26, 48, 58, 132] or body behaviours involving *shoulder and torso* along with *facial movements* [31, 49, 113] accompanied with speech will be discussed in this section.

An audiovisual model by Mariooryad and Busso [132] relied on three joint Dynamic Bayesian Networks (jDBNs) to generate facial gestures, involving head and eyebrow movements, by mapping the acoustic speech data from the IEMOCAP database [20] to Facial Animation Parameters [145]. The model was trained by adapting the algorithms used for HMM and FHMM [68]. Using the Canonical Correlation Analysis (CCA) [44, 83], the joint DBN model was compared to similar models used to synthesize head and eyebrow motions separately. Overall, the objective evaluation results revealed that the jDBN models can cope with speaker variability, while the subjective results showed an increase in the quality of jointly modeled eyebrow and head gestures as well as their naturalness.

Ding et al. [48] proposed an animation model of a virtual agent, based on a fully parameterized Hidden Markov Model (HMM), which produces head and eyebrow movements in synchronisation with speech. As an extension of the contextual HMM, in FPHMM [216], contextual variables control and parametrize the means, covariance matrices, transition probabilities as well as initial state distribution. The model was evaluated objectively and subjectively on the Biwi 3D AudioVisual Corpus of Affective Communication database [60], considering facial motion and speech features. An objective evaluation, compared with the baseline proposed by [132] using the Mean squared error (MSE) [6] demonstrated the best performance by the HMM-based joint model. Overall, the proposed model demonstrated an ability to capture the link between speech prosody and head and eyebrow motions. Subjectively, the perceptual questionnaire struggles to validate the objective evaluation as the results were marginally significant, showing quite identical performance in terms of expressiveness.

Ding et al. [49] presented a multimodal behaviour generation model based on the contextual Gaussian model and a Proportional-Derivative controller (PD). They leveraged the AVLaughter database [196] for producing multiple outputs (lip, jaw, head, eyebrow, torso and shoulder motions) synchronized with laughter audio. Using the pseudo-phonemes and speech features as input, motion synthesis was carried out in three steps: first, the lip and jaw motions were synthesized by a contextual Gaussian module (CGM); second, speech features were extracted for predicting head and eyebrow movements, consequently, torso and shoulder motions were synthesized from the previous step of synthesis by concatenation. The sophisticated subjective evaluation of the generated laughter and bodily behaviours, using a questionnaire adapted from [143] and Likert-scale rating, manifested users' preference for an agent which produces synchronized speech and laughter animations.

Chiu and Marsella [31] introduced a combined model to learn a twofold mapping: from speech to a gestural annotation using Conditional Random Fields (CRFs) and from gestural annotation to gesture motion by applying Gaussian Process Latent Variable Models (GPLVMs) [208]. The model was subjectively evaluated against the approach by [29], which used direct mapping. The subjective evaluation was followed up by an objective assessment to establish the performance of the model against support vector machines (SVMs) [42]. As a result, the proposed method performed significantly better in generating and coupling the gestures with speech, despite the hurdles of the inference model that requires temporal information.

Fan et al. [58] discussed the use of deep Bi-directional Long Short-Term Memory (DBLSTM) [232] to model the temporal and long-range dependencies of audio/visual stereo data for a photo-real talking head animation from audio, video, and text input. To train the network, the study used back-propagation through time algorithm (BPTT) [214, 215]. The study demonstrated the advantages of two BLSTM layers sitting on top of one feed-forward layer on the datasets. As a result of objective (RMSE [73, 162, 209] and CORR [215]) and subjective evaluation (A/B preference test [108]), the proposed deep BLSTM model showed higher performance compared with the previous HMM-based approach.

Li et al. [123] adopted a deep Bi-directional Long Short-Term Memory (DBLSTM) [232] recurrent neural network as a regression method to generate audiovisual animation of an expressive talking face. This method was devised to overcome the shortcomings of the previous state-of-the-art models in incorporating lip movements with emotional facial expressions. Thus, Li et al. [123] proposed five methods based on DBLSTM trained using a large corpus of neutral data and a smaller scale corpus of emotional data. Specifically, in method (a), the DBLSTM network is trained with emotional corpus only; method (b) and (c) capture neutral and emotional information simultaneously by training a single DBLSTM network; while method (d) and (e) capture neutral information by a separate DBLSTM network in addition to emotional DBLSTM. To evaluate the proposed approaches, the authors adopted root mean squared error (RMSE) between the predicted Facial Animation Parameters (FAP) and ground truth. This revealed how different regression models worked for different emotions. Notably, information from the neutral dataset was found more valuable for peaceful expressions (e.g., sadness) than exaggerated expressions (e.g., surprise and disgust). A further frame-wise comparison of RMSE values displayed the effectiveness of the proposed methods in modelling the interaction between emotional states, facial expressions and lip movements. Finally, the subjective evaluation results confirmed the effectiveness of using the neutral dataset as it can improve the performance of an expressive talking avatar.

Suwajanakorn et al. [183] used recurrent neural networks to learn the mapping from raw audio input (MFCC audio features) to lip landmarks (PCA), synthesizing lip textures and then merging them into the 3D face to output a realistic talking head with clear lip motions synced with the input audio. The network consisted of LSTM nodes and was trained using backpropagation through time with 100 time steps. When compared against AAM approach [41] and Face2Face algorithm [191] in an objective evaluation, the proposed method synthesized cleaner and more convincing lip movements.

Chung et al. [37] proposed an encoder-decoder CNN-based Speech2Vid model, taking still images and audio speech segments to output a video of the face, including lip synchronized with the audio. The architecture constitutes three modules, such as the audio encoder, identity encoder, and image decoder, which were trained together. Learning the joint embedding of the target face and speech segments is central to this approach in generating a talking face. Evaluations, conducted to qualitatively measure the quality using the alignment and the Poisson editing [150] techniques, determined the ability of Speech2Vid to generate videos of talking faces with certain identities.

Chen et al. [26] developed a method that takes speech audio and one lip image of a target identity as input and generates an output of multiple lip images with the accompanying speech audio. The model is designed by combining correlation networks with an audio encoder and an optical flow encoder, implemented on 3D RNN to mitigate delayed correlation problems. The generated lip movements were evaluated quantitatively and qualitatively on the GRID [40] corpus, LRW [36] and LDC [157] dataset, not used previously for training purposes, as well as with different metrics - LMD, CPBD [140], and Structural Similarity (SSIM) and Peak Signal-to-Noise Ratio (PSNR) [213]. The proposed model generated realistic lip movements and proved their robustness to view angles, lip shapes, and facial characteristics. However, the main limitations are bound to learning from a single image, which resulted in difficulties in capturing lip deformations.

Plappert et al. [153] introduced a model based on deep Recurrent Neural Networks (RNNs), and sequence-to-sequence learning [182], which learns a bi-directional mapping between whole-body motion and natural language. One model is fed the encoded motion sequences obtained from motion capture recordings during training, and the other is trained on natural language descriptions to generate whole-body motions. Based on the quantitative comparison with the baseline model, the language-to-motion model demonstrated the capability of generating proper human motion, achieving higher performance rates. The performance of the model was also measured by BLEU scores [149], which suggested minimal overfit and generalisation to previously unseen motions. The model showed a capability to generate whole body motions given proper descriptions in natural language.

Alexanderson et al. [5] adapted a deep learning-based MoGlow [87] for a probabilistic speech-driven model to output full-body gestures synced with speech. Particularly, the normalising flows were used the same way as GANs to generate output by a nonlinear transformation of latent noise variables. Thus, four models were trained on a speech-only condition, while the other four were conditioned on style control. The model was compared against three baselines taking the same speech representation as input: unidirectional LSTM [90], conditional variational autoencoder (CVAE) [77], and the audio-to-representation system (ARP) [112]. While the subjective evaluation of the style control experiment yielded significant results in favor of the MoGlow-based model for the human-likeness of the gesticulation, the model trained on speech only achieved better results compared to the second baseline.

Dahmani et al. [43] used a conditional generative model based on a variational auto-encoder (VAE) framework for expressive text-to-audiovisual speech synthesis. The proposed model learns from textual input, which provides the VAE with embedded representation to further capture emotion characteristics (Fig. 4). Although the experimental results showed a high recognition rate for almost all emotions in audiovisual animations, sadness and fear turned

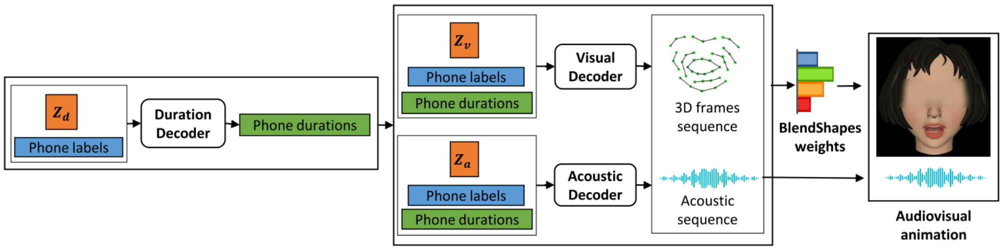

Fig. 4. The architecture of the audiovisual model for animation generation by Dahmani et al. [43].

out to be the hardest to recognize by participants. According to the authors, this was explained by the role of the upper part of the face, thus causing a potential limitation of the study. Overall, the model performed well in terms of producing nuances of emotions as well as generating emotions beyond those retrieved from the database as illustrated by subjective evaluation results.

Kucherenko et al. [113] presented a deep learning-based model that takes audio and text transcriptions as input data to generate arbitrary (metaphoric, iconic, and deictic) and semantically linked upper-body gestures together with speech for virtual agents. The model was evaluated on The Trinity Speech-Gesture Dataset [62] using the RMSE, acceleration and jerk, and acceleration histograms as objective metrics. A binomial test was used for the analysis of data obtained from the perceptual questionnaire and attention check. Altogether, the evaluations demonstrated a preference for the proposed model (no PCA) over the CNN-GAN model introduced by Ginosar et al. [70] in terms of human-likeness and speaker reflection. The evaluation results also highlighted the efficacy of the multiple modalities used to train the model.

Yoon et al. [227] discussed an end-to-end model that takes speech text, audio, and speaker identity to generate upper-body gestures, co-occurring with speech and its rhythm. The proposed method is based on Bi-directional GRU [32] along with recurrent neural networks used for encoding three different input modalities. The ablation study demonstrated that all three modalities had a positive effect on the generation of gestures. Overall, the proposed model performed well as identified by a novel objective evaluation metric called Fréchet Gesture Distance (FGD) [88], subjective user study and in comparison to other state-of-the-art models. Despite the superiority of the proposed model over baselines, the main disadvantage still remains the demand for a large dataset as the generated motion quality and upper-body gestures were limited to the dataset used in the study. Additionally, the gesture generation process lacks controllability. Other limitations regard the FGD, which made it atypical to analyze mixed measurements of motion quality and diversity.

Ahuja et al. [3] presented a Mixture-Model guided Style and Audio for Gesture Generation (Mix-StAGE) model which trains a single model for multiple speakers while learning unique style embeddings for each speaker's gestures in an end-to-end manner. A novelty of Mix-StAGE is to learn a mixture of generative models which allows for conditioning on the unique gesture style of each speaker. The model used a Temporal Convolution Network (TCN) module for both content and style encoders. It is trained on a custom-made dataset PoseAudio-Transcript-Style (PATS) designed specifically for this work. In the experimental study, the Mix-StAGE model was compared against existing baselines capable of generating similar co-speech gestures (i.e., single speaker models Speech2Gesture [70], CMix-GAN and multi-speaker models MUNIT [92], StAGE). The results of the objective evaluation revealed that the Mix-StAGE model significantly outperformed the state-of-the-art approaches for gesture generation and provided a path towards performing gesture style transfer across multiple speakers. Perceptual studies also showed that the generated animations by the proposed model were more natural whilst being able to retain or transfer style.

Wang et al. [210] introduced an integrated deep learning architecture for speech and gesture synthesis (ISG) model to synthesize two modalities in a single model, compatible with both social robots and embodied conversational agents (ECAs). The proposed model is adapted from Tacotron 2 [174] and Glow-TTS [102], with Tacotron 2 being auto-regressive and non-probabilistic and Glow-TTS being parallel and probabilistic, and takes text as input to generate speech and gesture. Subjective tests performed separately for each modality demonstrated that one of the proposed ISG models (ST-Tacotron2-ISG) performs comparably to the current state-of-the-art pipeline system while being faster and having much fewer parameters.

Huang et al. [93] proposed a fine-grained Audio-to-Video-to-Words framework, called AVWnet, which is deemed to produce videos of a talking face in a coarse-to-fine manner and maintain audio-lip motion consistency. The framework architecture consisted of tree-like architecture and a GAN-based [72] neural architecture for synthesizing realistic talking face frames directly from audio clips and an input image. The GAN framework is conditioned on image features to enable further fusion of facial features and audio information in generating the face video. Compared with the state-of-the-art approaches [27, 37], the performance of AWVnet excelled on all three adopted metrics and datasets as a result of objective evaluation. Metrics such as Structural Similarity Index (SSIM), Peak Signal-to-Noise Ratio (PSNR), and Landmark Distance Error (LMD) were used to evaluate the model objectively. A comparison of the proposed model with the model by Chen et al. [27] through perceptual user study revealed the former to be as good as the existing model.

Zhou et al. [236] presented a model that learns from disentangled audio-video representations to generate a talking face corresponding to speech. Both talking video and audio were used to train the Disentangled Audio-Visual System (DAVS). The DAVS network demonstrated several advantages over the previous baseline [36], which encompass the improvement of lip-reading performance, unification of audio-visual speech recognition and synchronisation in an end-to-end framework, and the achievement of a high-quality and temporally accurate talking face generation as a result of both subjective user study and effectiveness verification by Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity (SSIM) [213].

Sadoughi and Busso [166] demonstrated a Constrained Dynamic Bayesian Networks (CDBN) [132], to overcome the individual limitations of rule-based and data-driven approaches in gesture generation. The authors aimed to build a generative model to produce believable hand gestures along with head gestures with bimodal audio-speech and video data synchronisation. The model was evaluated by two objective metrics: canonical correlation analysis (CCA [21, 83]) and log-likelihood rate (LLR) [136]. Based on the results of the subjective evaluation, the CDBN model is perceived to generate more appropriate and natural gestures compared to baseline models. Overall, the hand gestures generated by the constrained model showed 85% accuracy for certain types of gestures.

Vougioukas et al. [206] discussed the GAN-based talking face generator, consisting of a temporal generator and multiple discriminators, which takes a single image and raw audio signals as input. The quality of the generated video output was evaluated on the GRID [40] corpus, TCD TIMIT [84] corpus, CREMA-D [23] and LRW [36] datasets by applying reconstruction (Peak Signal-to-Noise Ratio and Structural Similarity [213]), sharpness (cumulative probability blur detection (CPBD) measure [139]), content (average content distance (ACD) [193] and word error rate (WER)), and audio-visual synchrony metrics. When assessed subjectively, the results of the Turing test [30] showed naturalness of the generated faces. Moreover, compared to baselines [37, 183], the model demonstrated an ability to not only capture and maintain identity but generate facial expressions matching the speaker's tone and speech.

Sinha et al. [177] approached the generation of identity-preserving and audio-visually synchronized 2D facial animation through GAN, utilizing DeepSpeech features, given an audio input of speech, and facial landmarks from the benchmark corpora as GRID [40] and TCD-TIMIT [84]. Same objective evaluation metrics as in [26] were used in the study. Moreover, a qualitative evaluation compared the model with the state-of-the-art baselines of [26], [206], and [236]. These evaluations yielded overall positive results regarding identity preservation, superior image quality and texture clarity, and smooth audio-visual synchronisation.

Tables 4 and 5 summarize the state-of-the-art in multimodal gesture generation, concerning the corpora and evaluation metrics used. Even though studies emphasize objective evaluation as a challenging task, the existing literature shows effective and nuanced exploitation of objective metrics along with subjective ones. Note that objective metrics are often the same as the cost functions used to optimise the generative models, with authors assuming that optimising the cost functions equates with improving the model's performance. However, for now subjective measures remain the gold standard for assessing the quality of the generated behaviour and this is recognised across the field..

Summary: Multimodal Gestures

- Multimodal gesture generation creates an opportunity for a holistic approach to generating social behaviour, and improves over generating isolated behaviours (e.g., hand gestures, speech synthesis). Early demonstrations exist combining speech and hand gestures, and speech and body behaviours, to mention but a few.
- Future developments are expected to broaden the scope of multimodal gesture generation. Potential low-hanging fruit is using or predicting emotional states, e.g. from audio, to produce corresponding communicative behaviour [183], and moving towards gestures driven by semantic content [5, 113].
- In most multimodal generative systems, the different modalities are still considered in isolation. Building a flexible system that is able to jointly generate whole-body gestures, from and with verbal cues, remains a challenge [183, 227].

7 SPEECH SYNTHESIS

Speech is often a prime aspect of interactive communication, and in embodied systems often co-occurs with gestures. Recent years have seen active development of data-driven models for synthesizing speech from input text (Text-to-Speech (TTS) synthesis) using various deep learning models. Most speech synthesis approaches in the literature focused

[30] https://forms.gle/XDcZm8q5zbWmH7bD9
[31] The authors did not provide details on the sizes of training and test sets.
[32] Not applicable, ibid., p. 5
[33] The authors used the qualitative observation for evaluation.
[34] The type of qualitative metric used to measure the naturalness is not provided.
[35] This duration is an approximation.
[36] This duration is an approximation.
[37] In line with [112], the authors opted to use these metrics to measure the quality of the generated gestures.
[38] The exact duration for the training and test splits, other than that each sample contained a one-second video with the target word spoken, are not provided.
[39] https://forms.gle/XDcZm8q5zbWmH7bD9

Table 4. Corpora and evaluation used in the multimodal gesture generation literature

	Corpus			Evaluation	
	Source	**Training**	**Test**	**Objective**	**Subjective**
Mariooryad and Busso [132]	IEMOCAP database [20]	75% out of 418 utterances	25% out of 418 utterances	Canonical correlation analysis (CCA) [83]	Questionnaire (speaker-dependent and speaker-independent, 5-point Likert scale)
Ding et al. [48]	Biwi 3D AudioVisual Corpus of Affective Communication database [60]	80% out of 240 sequences	20% out of 240 sequences	MSE [6]	Questionnaire (5-point Likert scale)
Chiu and Marsella [31]	Audio and body motion perception dataset [55]	193 seconds	238 seconds	N/A	Questionnaire
Fan et al. [58]	Audio-visual database of a talking subject [58]	80% out of 81974 images (20000 images)	10% out of the total database	RMSE (shape) [209]; RMSE (texture) [162]; RMSE (appearance) [73]; CORR [215]	A/B preference test (naturalness) [108]
Ding et al. [49][31]	AVLaughterCycle database [196]	N/A	N/A	N/A	Questionnaires, riddles, smiles, laughs [143]
Li et al. [123]	eNTERFACE'05 emotion database [133]; Neutral dataset [123]	608 seconds (10.1 min) 1280 seconds (21.4 min)	24 seconds	Root mean squared error (RMSE)	Questionnaire (5-point Likert scale)
Suwajanakorn et al. [183]	Video addresses of Obama [183]	14 hours	3 hours	Consistency (with and without re-timing)	N/A [32]
Chung et al. [37][33]	VoxCeleb dataset [138] LRW dataset [36]	37.7 hours	0.5 hours	N/A	Image naturalness, movement naturalness [34]
Chen et al. [26]	GRID dataset [40] LDC dataset[157] LRW dataset [36]	37.5 hours 159.8 hours 6.4 hours	1.3 hours 7.8 hours 1.2 hours	LMD, CPBD [140], Structural Similarity (SSIM), Peak Signal-to-Noise Ratio (PSNR) [213]	N/A
Plappert et al. [153]	KIT Motion-Language Dataset [152]	80 % of the total dataset, (2 846 motion samples; 6 187 natural language annotations)	10% of the total dataset	BLEU scores [149]	N/A
Alexanderson et al. [5]	The Trinity Gesture Dataset [62]	20,665 samples of data	400 seconds	N/A	Cross-comparison rating, questionnaire
Dahmani et al. [43]	The ESTER database [76]	3h12[35] (1600 sentences) 4h8[36] (2400 sentences)	200 sentences 300 sentences	N/A	Preference test
Kucherenko et al. [113]	The Trinity Speech-Gesture dataset [62]	70 sequences of aligned text, audio and gestures per each training	20 minutes (50 segments of 10 seconds each)	Average values of RMSE, acceleration and jerk (rate of change of acceleration), and acceleration histograms[37]	Questionnaire, attention check
Yoon et al. [227]	TED Gesture Dataset [228]	97 hours (199,384 sequences/766 videos)	25,930 sequences	Fréchet Gesture Distance (FGD) [88]	Pairwise comparison
Wang et al. [210]	Trinity Speech-Gesture Dataset [62, 114]	10.6 minutes	N/A	N/A	Multiple Stimuli with Hidden Reference and Anchor interface (MUSHRA) [19], Mean Opinion Score (MOS), Questionnaire

on neutral speech, while some considered generating affective speech. In the next part, we will give an overview of some important and commonly used speech synthesis systems.

7.1 Neutral Speech Synthesis Systems

WaveNet: van den Oord et al. [197] discussed a system based on the PixelCNN decoders [199, 200]. The proposed model uses dilated causal convolutional layers to ensure that the conditional probability of an audio sample at a particular time step is not dependent on samples at future time steps (but only on previous time steps)[40]. Moreover, the model uses residual block and skip connections to accelerate convergence during the training of the network [86]. The results show

[40]In WaveNet, it is possible to condition the model on additional inputs like the speaker identity in case of a multi-speaker setting.

Table 5. Corpora and evaluation used in the multimodal gesture generation literature (continued)

	Corpus			Evaluation	
	Source	Training	Test	Objective	Subjective
Huang et al. [93]	GRID dataset [40] LRW dataset [36]	1000 video samples	50 video samples	Structural Similarity Index (SSIM), Peak Signal-to-Noise Ratio (PSNR) [213], and Landmark Distance Error (LMD) [26]	Questionnaire (5-point Likert scale)
Zhou et al. [236]	LRW dataset [36]	800 samples	50 samples [38]	PSNR and SSIM [213]	Questionnaire (true or false)
Sadoughi and Busso [166]	The MSP-Avatar corpus[168]	2 hours 58 minutes (74 sessions)	734.4s for affirmation, 1118.7s for negation, 1149.1s for question, 1582.5s for suggestion, 6111.7s for other	CCA [21, 83] and log-likelihood rate (LLR) [136]	Questionnaire (5-point Likert scale)
Vougioukas et al. [206]	GRID corpus [40] TCD-TIMIT corpus [84] CREMA-D dataset [23] LRW dataset [36]	26h4 9h1 9h7 36h3	8h31 1h2 0h68 1h9	PSNR, SSIM[213], cumulative probability blur detection (CPBD) [139], average content distance (ACD) [193], word error rate (WER) [107], Euclidean distances [38]	Online Turing test[39]
Sinha et al. [177]	GRID corpus [40] TCD TIMIT dataset [84]	26.4 hours 9.1 hours	8.31 hours 1.2 hours	PSNR, SSIM[213]; CPBD [139]; LMD [26]	Questionnaire (10-point Likert scale)

that the WaveNet speech synthesizer achieved a better Mean Opinion Score (MOS) [156] in terms of the naturalness of the generated speech samples than that of the LSTM-RNN-based statistical parametric speech synthesizer [231] and the HMM-driven unit selection concatenative speech synthesizer [71] in addition to higher subjective preference scores. This model was further improved to **Parallel WaveNet** [201] that can generate more than one audio sample at a time while keeping a similar quality to – but is largely faster than – the original WaveNet.

Tacotron: Wang et al. [211] presented a system based on a sequence-to-sequence (seq2seq) model [11, 182] with an encoder that encodes input character embeddings into context vectors, an attention-based decoder [11, 204] that turns the encoder final representation into a Mel-scale spectrogram, and a CBHG[41]-based post-processing net that converts spectrogram frames to waveforms using the Griffin-Lim reconstruction algorithm [78]. The results show that the Tacotron model achieved a better Mean Opinion Score (MOS) [156] in terms of speech naturalness than that of the parametric speech synthesis system [231], and a marginally lower score than that of the concatenative speech synthesis system [71], which is a promising result considering the audible artifacts produced by the Griffin-Lim synthesis approach. This opened the door to another improved version of the system; **Tacotron 2** [175], which is a combination of convolutional and recurrent neural networks and WaveNet vocoder (derived from the WaveNet architecture [197]). This model outperformed the parametric, concatenative, Tacotron (Griffin-Lim), and WaveNet text-to-speech systems in subjective evaluation.

Deep Voice: Arik et al. [8] discussed a system for speech synthesis, where each model of the system is based on an independently trained deep neural network. The main sub-models of the system have the following functions: segmenting voice for calculating phoneme boundaries, in the training pipeline only, using a recurrent architecture with connectionist temporal classification loss [74], in addition to converting grapheme (text)-to-phoneme using encoder

[41]CBHG is an efficient module for calculating sequence representation. It consists of a one-dimensional convolutional filters' bank, highway networks [181], and a Bi-directional Gated Recurrent Unit (GRU) net [34].

and decoder with Gated Recurrent Units (GRU) [32], predicting phoneme duration and fundamental frequency, and synthesizing audio based on WaveNet architecture [197] with a bi-directional Quasi-RNN (QRNN) conditioning network [18] in both the training and inference pipelines. The results show relatively lower (but promising) Mean Opinion Scores (MOS) [156] for the synthesized audio with respect to ground truth recordings. This opened the door to other improved/novel[42] multi-speaker versions of the system; **Deep Voice 2** [69] with a high quality of synthesized audio that outperforms that of the Deep Voice synthesis system, and **Deep Voice 3** [151] that outperforms Deep Voice 2 and Tacotron (Griffin-Lim), while it has a similar performance to Tacotron 2 in case both are using WaveNet vocoder.

VoiceLoop: Taigman et al. [185] introduced an approach for speech synthesis inspired by the working memory model; the phonological loop [10]. An input sentence (text) to the model is represented as a set of phonemes, where each phoneme is represented through an embedding vector. These vectors are weighted and summed to create a context vector using attention weights. The model uses a memory buffer, which is updated by a new, speaker-dependent, representation vector, at each time step, calculated with a shallow fully connected network that has as input: the context vector with speaker embedding, and both the output and buffer vectors at the previous time step. The output of the model is calculated through another network of the same architecture that has as input the buffer vector at the current time step with speaker embedding. The results show that the VoiceLoop model outperformed the Tacotron and Char2Wav [180] models in the Mean Opinion Scores (MOS) [156] – subjective evaluation – and Mel Cepstral Distortion (MCD) scores – objective evaluation – in single and multi-speaker speech synthesis.

WaveGlow: Prenger et al. [155] proposed a flow-based network capable of generating high-quality speech from mel-spectrograms. Following the examples of Glow [106] and WaveNet [197], the WaveGlow produces efficient and high-quality audio without the need for auto-regression. An experimental study is conducted to subjectively compare the proposed model against two baselines, such as the Griffin-Lim [79] algorithm and WaveNet [197], using the Mean Opinion Scores (MOS) [156] as a metric. The results showed that WaveGlow delivers audio quality as good as the best publicly available WaveNet implementation trained on the same dataset.

WaveGrad: Chen et al. [28] presented a conditional speech synthesis model of waveform samples that estimates the gradients of the data log-density as opposed to the density itself. It is non-autoregressive as it requires only a constant number of generation steps during inference. In particular, starting from Gaussian noise, gradient-based sampling is applied using as few as 6 iterations to achieve accurate audio. The experiments demonstrated that WaveGrad is capable of generating high-fidelity audio samples, outperforming adversarial non-autoregressive models [15, 116, 222, 223] in an objective evaluation and matching one of the best autoregressive baseline models [100] in terms of subjective naturalness.

7.2 Affective Speech Synthesis Systems

Lee et al. [120] introduced an altered version of Tacotron, injecting an emotional embedding e to attention RNN to generate speech with specifications of emotion and personality of a human. The model was trained and evaluated on two Korean emotional speech datasets – one from Acriil, the other from ETRI – the former containing speech, audio, emotional label pairs, while the latter containing a drama script. Through quantitative experiments, the authors identified two areas of improvement concerning attention alignment. First, due to the scarcity of the frame of a

[42]**Deep Voice 2** has a modified architecture with respect to **Deep Voice** through separating between the phoneme duration and frequency models and adding batch normalisation and residual connections in the convolutional layers in the segmentation model. **Deep Voice 3** is a novel fully convolutional attention-based speech synthesis system. It consists of an encoder that maps textual features to an internal representation, a decoder that maps the encoder representation to an audio representation, and a converter as a post-processing net. It is a fully convolutional system (unlike Tacotron), which makes computation and training very fast.

spectrogram, the authors opted to concatenate attention text to the attention RNN's input to achieve an alignment of the speech with pronunciation. Second, they applied residual connections to the Convolution Bank + Highway + bi-GRU (CBHG) module [119] for a sharper and clearer attention alignment. Overall, the results showed that the quality of the generated speech was highly correlated with the sharpness of the attention alignment, despite the limited emotional representation in the speech.

Um et al. [195] developed a text-to-speech system based on the intra-category distance that generates emotional speech and controls the intensity of emotion representation. In doing so, they first proposed an inter-to-intra distance ratio algorithm to enable the inclusion of a wider range of emotions simultaneously and enhance their clarity utilizing the ratio between intra- and inter-cluster embedding vectors. Then an interpolation technique was introduced to control the intensity of the emotions effectively. During training, the global style token Tacotron (GST-Tacotron) model [212] was used as a baseline, taking a large number of neutral utterances as input. The effectiveness of the method was assessed subjectively using Mean Opinion Score (MOS) tests [156] in terms of the quality of the synthesized speech, while the preference test measured the expressiveness of sadness, anger, and happiness against the mean-based method. As a result, the proposed approach outperformed the conventional mean-based method in both criteria.

Byun and Lee [22] proposed a multi-conditional emotional speech synthesizer through the Tacotron [211] model by providing it with an emotional embedding from a multiple-speaker Korean emotional speech database [22]. For the Tacotron to synthesize multi-conditional speech, the authors injected the embedding vector into the Decoder RNN, which enables the generation of mel-spectrogram frames. In addition, the Attention module of the Tacotron was trained using both the emotional speech dataset and a large set of speech data for TTS. The extent to which the model was emotionally expressive and clear was evaluated by the Mean Opinion Score (MOS) test [156] in a subjective study, which resulted in the superiority of the proposed method of emotional speech synthesis generating four emotions as output: happiness, anger, neutrality and sadness.

Li et al. [122] introduced a novel reference-based approach for emotional speech synthesis based on Tacotron to synthesize speech with neutral and six basic emotions [52]. Specifically, the model integrates four losses such as the basic Tacotron MSE loss, two emotion classification losses and the style loss [67, 98]. As input, the model takes the Chinese test first converted into a character sequence, then, CBHG module [119] converts a pre-net output into the final encoder representation, and finally, the mel-spectrogram is transformed using the CBHG post-net to obtain a linear spectrogram. The model's ability to transfer emotion was evaluated through ablation studies, while the emotion strength control was measured by strength ordering test against the RA-Tacotron [237] in a subjective evaluation. It was observable from the results that the speech synthesized with the proposed method was more accurate and expressive, displaying less emotion confusion.

Lei et al. [121] proposed a fine-grained emotion transfer (FET), control, and prediction approach for expressive speech synthesis that shares architecture with Tacotron [211] and Tacotron2 [175], generating mel-spectrogram through a CBHG-based text encoder and an attention-based auto-regressive acoustic decoder. As regards emotion expression, emotional information is learned from the input text in emotion transfer, reference audio in emotion control, and manual labels in emotion prediction. To control the emotion category, the authors adopted the emotional embeddings, which is further treated as the global render of speech in the seq2seq model for emotion transfer. The emotion prediction, on the other hand, learns directly from the phoneme sequences without any reference audio or labels. Finally, the FET was compared subjectively with the GST model [212] and the utterance-level emotion transfer model (UET) [237], trained by ground-truth mel-spectrogram, using mel-cepstral distortion (MCD) [110] and A/B preference test [108] as metrics. For objective evaluation, Dynamic Time Warping (DTW) [137] was adopted to evaluate the predicted features and target features. The FET model demonstrated better performance compared to the baselines in terms of coarse emotional expressions and its flexibility in synthesizing the emotional speech with the six basic emotions as happiness, anger, fear, sadness, disgust and surprise [52].

Liu et al. [126] proposed a novel training strategy for Tacotron-based speech synthesis which does not require prosody annotation for training. Instead, the model unifies frame and style reconstruction loss. It is then implemented on speech emotion recognition (SER) and used as a style descriptor for extracting high-level prosody representations. The proposed strategy is called Tacotron-PL due to the use of perception loss (PL) [98] for style reconstruction loss. In a comparative study, there were five Tacotron-based text-to-speech systems developed, including baseline Tacotron and its four variants with the proposed Tacotron-PL among them. Three different evaluation metrics were used for an objective performance evaluation with regard to spectral modeling, F0 modeling, duration modeling, and deep style features. Subjective evaluations are conducted through Mean Opinion Score (MOS) [156], A/B preference tests [108], and Best Worst Scaling (BWS) [65]. By outperforming the other baselines, Tacotron-PL demonstrated the advantages of the proposed training strategy in terms of expressiveness and feasibility in synthesizing four emotional categories including sad, happy, angry and neutral.

Wu et al. [220] integrated two descriptors – Capsule Network (CapNet) and Residual Error Network (RENet) – for a sequence-to-sequence (seq2seq) architecture of an end-to-end emotive speech synthesizer which synthesizes speech with anger, happiness, sadness and other emotions. CapNet is employed for speech emotion recognition (SER) by outputting a set of probabilities that correspond to the emotions, while RENet is considered advantageous for deriving latent emotive representations. Unlike the existing methods, this method utilizes an utterance exemplar for emotion specification. Specifically, exemplary descriptors are integrated into the seq2seq to control the synthesis. Thus, this work proposed five E-TTS systems based on categorical descriptors - emotion code vector *(EC-TTS)*, various emotions *(EP-TTS)*, logit-based descriptor *(EL-TTS)* from SER, and automatically derived descriptor - *EA-TTS* and *EAli-TTS* from RENet. An experimental study evaluated the emotion similarity and speech quality objectively by calculating the mean squared error (MSE) [6] and subjectively through mean opinion scores (MOS) test [156] on an audio-book corpus from the 2011 Blizzard Challenge [104]. Among the two baselines (Tacotron [211] and GST-Tacotron [212]) and five proposed E-TTS systems (EC-TTS, EP-TTS, EL-TTS, EA-TTS, and EAli-TTS), the E-TTS systems performed significantly better than the baselines, while EA-TTS achieved the best performance in emotion similarity.

Annotated here are the advanced versions of the speech synthesis systems both for neutral and affective speech, primarily based on Tacotron [211], the performance and quality of which were proven through objective and subjective measures (See Table 6 for details) and benchmarking against the state-of-the-art models. Nonetheless, a few shortcomings have been encountered during training. For instance, Lee et al. [120] pointed out the scarcity of the emotional representations in speech as a significant limitation. It can also be observed from Table 6 that the subjective evaluations prevail compared to the objective evaluations.

[43] Dataset sizes are **not available**
[44] Not applicable, ibid., p. 5
[45] This is an approximation based on the details provided in the article, where authors each file lasting from two to three hours for each of the four actors.
[46] As a quantitative measure, the authors computed MSE values.

Table 6. Corpora and evaluation used in the speech synthesis literature

	Corpus			Evaluation	
	Source	**Training**	**Test**	**Objective**	**Subjective**
van den Oord et al. [197]	CSTR VCTK corpus [221]	32 audio clips (7,860 timesteps)	N/A[43]	N/A[44]	Mean Opinion Score (MOS) [156]
Wang et al. [211]	North American English dataset [211]	24.6 hours of speech	4.1 minutes (1% of the training data)	N/A	MOS [156]
Arik et al. [8]	English speech database [8]; Blizzard Challenge dataset [154]	20 hours (13,079 utterances); 20.5 hours (9,741 utterances)	N/A	N/A	MOS [156]
Taigman et al. [185]	CSTR VCTK corpus [203] LJ database [97] The Nancy corpus [104] English audiobook [154]	N/A	N/A	Mel-cepstral distortion (MCD) [110]	MOS [156]
Lee et al. [120]	Korean speech dataset from Acriil	21 hours	N/A	N/A	MOS [156]
Um et al. [195]	Korean male voice database	3.79 hours (2,965 utterances)	N/A	N/A	MOS [156]
Byun and Lee [22]	Korean Single Speaker Speech Dataset (KSS Dataset) [1]	8-10 hours [45] (18,324 audio files)	100 audio files (3-10 seconds each)	N/A	MOS [156]
Li et al. [122]	Emotional Speech Corpus [237]	14 hours	70 sentences (10 per emotion)	N/A	Strength ordering test
Lei et al. [121]	Emotional Speech Corpus [237]	14 hours	210 sentences (30 per emotion)	Dynamic Time Warping (DTW) [137], Mel-cepstral distortion (MCD) [110]	A/B preference test
Liu et al. [126]	IEMOCAP database [20] LJ database [97]	10039 utterances 24 hours	N/A	MCD [110], Root Mean Squared Error (RMSE), Frame Disturbance (FD), Dynamic Time Warping (DTW) [137]	MOS [156], A/B preference test [108], Best Worst Scaling (BWS) [65]
Wu et al. [220]	IEMOCAP database [20], The English audiobook [104]	8 speaker sessions 4.79 hours	1 speaker session 0.35 hours	Mean squared error (MSE)[46]	MOS [156]
Chen et al. [28]	Proprietary speech dataset [28], LJ database [97]	385 hours, 23 hours	1,000 sentences	Log-mel spectrogram mean squared error metrics (LS-MSE), MCD [110], F_0 Frame Error (FFE) [33]	Listening test (5-point MOS scale) [156]

Summary: Speech Synthesis

- Speech production, known as text-to-speech synthesis, has benefited considerably from data-driven approaches, and is the most mature data-driven social behaviour available, with some artificial speech being almost indistinguishable from human speech.
- Commercial vendors have invested considerably in data-driven models, which far outperform academic products especially for neutral speech. Still, there is considerable spread in quality between languages.
- Most speech synthesis engines are unable to adaptively overlay affect and emotion, with most voices sounding neutral. This, currently, is a limitation for the field of Human-Robot Interaction (HRI), which calls for rich affective speech.
- Last but not least, it is noteworthy to mention that the high fidelity of artificial speech might not always suit the needs of HRI: studies [22, 185] suggest that a human-like voice might not fit the robotic appearance and that a more robotic voice might be more appropriate to the context of interaction.

8 OUTLOOK

It is clear that data-driven methods relying on connectionist architectures are an important and perhaps definitive answer to the question of how to generate human-like communicative behaviour. Never before have models produced such rich and varied behaviour without the need for explicit programming. However, there are a number of challenges that still face the relatively young field of data-driven behaviour generation.

Multimodal behaviour generation. Most models take a single signal and map it onto a modality: text to speech, emotion to facial expression, speech to gesture. However, in human-to-human communication all modalities are intertwined: emotion colours speech and gestures, gestures have an impact on speech, context influences eye gaze, etcetera. The fact that communication is a highly interdependent process is glossed over in current data-driven generation methods, for obvious reasons. Still, in future systems we would expect more modalities to be taken into consideration. In the speech generation community, for example, emotion has long been the subject of study, and research systems are able to generate speech modulated by emotion. However, the flipside to this is that for a data-driven approach more data will be needed. Already the amount of data required to train systems is expensive to collect for two connected modalities, adding other modalities is likely to increase the size of the required training data exponentially. How this will be overcome is as yet unclear.

Dyadic and multiparty communication. The large majority of data-driven models do not take the receiver into account. Instead they are trained to produce communicative behaviour as if it would concern a monologue in which the receiver of the message does not respond. In human-to-human communication, most interactions are multiparty interactions and our communicative behaviour is finely tuned to the reactions and responses of others. We watch for signals showing understand or misunderstanding, monitor for affective responses and are sensitive to bids for turn-taking. All these elements are largely missing from current data-driven methods, as they are exclusively trained on data that does not take into account the interactive nature of communication. Again, it seems likely that more data could resolve this problem, but at the same time collecting this data comes at a great cost and might be beyond the means of most R&D labs.

Measuring quality of generated behaviour. Assessing the quality of generated behaviour relies on objective and subjective measures. Objective measures are the workhorse of data-driven methods, as they form the cost function against which the models are optimised. Unfortunately, these objective measures only weakly correlate with subjective measures (see for example [114]). Subjective measures, during which people (or simulated subjective raters) judge the quality of the generated behaviour, remain the gold standard in evaluation. However, using human raters is expensive and time consuming and as such subjective measures cannot be used during training when many millions of evaluations are needed to drive the model ever closer to generating behaviour that is human-like. Recent work on gesture generation showed how subjective measures still are better for measuring the quality of models, and that objective measures often fall short as they only optimise a quantitative metric which is often a poor representation of qualitative assessment [217, 219]. Simulated subjective raters might be a way forward, as in GAN models in which one part of the model is trained to discriminate between artificial and human-like output, pushing the generated behaviour ever closer to being indistinguishable from human behaviour. Another challenge is the lack of common standards to evaluate models. Sometimes this is informed by the need to evaluate very specific elements of the generated behaviour, or because the accepted standard has outlived its usefulness. Benchmarks often form the focus of intense research investment and are often reached in just a few years, at which point they become useless as a target to aim for. Challenges, where different models are pitted against each other, have proven useful in this context – co-speech gestures for example have benefited from a series of challenges pushing the field, but also pushing the way in which models are evaluated [114, 229].

Common datasets and evaluation methods. From the survey it appears that there are few common datasets on which models are trained and evaluated. Researchers and engineers prefer taking a pragmatic approach when chosing data to train and evaluate against. Factors such as availability, easy-of-use, feature availability, cost and appropriateness for the task at hand are deemed important and are often used as a reason to not use datasets which have been used by others. One corollary is that the field would benefit from agreed datasets and evaluation standards, something which happens for some modalities (such as speech synthesis) and is slowly being adopted for other modalities (such as gesture generation [114]).

Semantics of multimodal communication. Communication serves to change the mind of others. As such, any communicative act carries semantics. However, this is usually glossed over in data-driven models. In some cases, this is not too much of a problem. Speech generation, for example, generates speech from text. Text has a well-agreed notation and speech generation maps this orthography to sound. However, speech generation is largely context-free and the production of human-like speech is possible without requiring much access to the semantics of the text and without access to the internal affective state of the agent. For exceptions to this the context of the neighbouring text is sufficient to disambiguate the required speech sounds. For example, disambiguating "bass" as a fish (/bas/) or a musical instrument (/beɪs/) can often be done by relying on other words nearby. Other modalities are different in that what they convey is tightly linked with affect, emotion and semantics of the message. Current data-driven methods do not have access to these, and while the models can with sufficient data pick up semantic correlations, the training cost at which this comes is prohibitive.

Fine tuning models. One promising benefit of data-driven neural models is the potential for fine-tuning (also known as transfer learning) of a pre-trained model. In this, a model is first trained using a large amount of data and then later training continues often on a smaller dataset so that the pre-trained model is more relevant for a specific task. While few behaviour generation models have been made available for fine-tuning, the practice is already well established in other fields, such as Large Language Models, where models can be relatively easily fine-tuned for other language-based generative tasks (e.g., [233]).

Hardware does not match the dynamics of software generated behaviour. Most social robots rely on actuation technology, such as electric motors and planetary gears, which do not offer the velocity, acceleration and jerk typically seen in the human body. This leads to multimodal social behaviour that appears unnaturally slow. Some solutions exist: some robots, such as Keepon, rely on simpler, smaller and lighter bodies which allow low-cost actuators to generate high-velocity dynamics. Others, such as EngineeredArts' Ameca or RoboThespian animatronic robots, rely on alternative actuation technology, often using pneumatics, to produce high-velocity animations matching human dynamics. However, human-like dynamics are for the moment still out of scope for most commercial and research social robots.

Despite these challenges, data-driven methods for the time being look to be the way forward. But to achieve near-human multimodal behaviour, a number of important obstacles will need to be overcome. One striking observation is that a developing child does not have access to thousands or perhaps millions of hours of training opportunities. Instead, children learn to interact multimodally through a combination of observation and online learning, and innate biases and constraints. This combination allows them to become skilled multimodal communicators in just a short few years. Perhaps future data-driven models should, instead of taking a *tabula rasa* approach, also start with biases and constraints to make the training process more efficient.

9 CONCLUSION

In this survey paper, we review different data-driven approaches, in the related literature, for behaviour generation covering speech, gestures, facial expressions, and body behaviour. The paper discusses the findings of different deep learning-based systems for behaviour generation and reflects on a road map for future research in this area at the intersection of both the Human-Robot Interaction (HRI) and Human-Agent Interaction (HAI) communities. We conclude that there are still challenges facing the efforts towards generating credible human-like multimodal behaviours, like the size of the available data sets for training the systems, generating affective behaviours, and evaluating measures of the generated behaviours.

The objective of this survey was to show the current state-of-the-art of behaviour generation approaches, and highlights successes in behaviour generation (e.g., speech synthesis that has come on in leap and bounds, based on the availability of transcribed data and sophisticated artificial neural models) but also areas in which improvement can be made (to stay with speech synthesis, one important limitation is that it still only generates neutral sounding speech). While we tried to be comprehensive, we have not covered all possible modalities. Eye gaze, for example, while important in face-to-face interaction between people and robots [2] is not covered as a separate modality in this review, as eye gaze behaviour has received little attention in data-driven behaviour generation. Still, given the ongoing success of data-driven generative methods, no modality will be untouched by it.

10 APPENDICES

A SEARCH KEYWORDS

Table 7. Examples of keywords used in the search query across databases.

Web of Science
TS[47]=face AND TS=generation AND TS=data-driven AND PY=(2014-2020)
TS=facial AND TS=generation AND TS=data-driven AND PY [48]=(2014-2020)
TS=hand gesture AND TS=generation AND TS=data-driven AND PY=(2014-2020)
Scopus
TITLE-ABS-KEY[49] (facial AND behaviour AND generation) AND TITLE-ABS-KEY (data-driven) AND PUBYEAR[50] >2014[51]
TITLE-ABS-KEY (face AND behaviour AND generation) AND TITLE-ABS-KEY (data-driven) AND PUBYEAR >2014
TITLE-ABS-KEY (face AND gesture AND generation) AND TITLE-ABS-KEY (data-driven) AND PUBYEAR >2014
TITLE-ABS-KEY (facial AND expression AND data-driven AND generation) AND PUBYEAR >2014
TITLE-ABS-KEY (lip AND motion AND generation) AND PUBYEAR >2014
TITLE-ABS-KEY (data AND lip AND motion AND generation) AND PUBYEAR >2014
TITLE-ABS-KEY (hand AND gesture AND generation) AND TITLE-ABS-KEY (data-driven) AND PUBYEAR >2014
TITLE-ABS-KEY (hand AND gesture AND generation) AND PUBYEAR >2014 AND (LIMIT-TO (PUBYEAR , 2020) OR LIMIT-TO (PUBYEAR , 2019) OR LIMIT-TO (PUBYEAR , 2018) OR LIMIT-TO (PUBYEAR , 2017) OR LIMIT-TO (PUBYEAR , 2016) OR LIMIT-TO (PUBYEAR , 2015))
TITLE-ABS-KEY (body AND action AND generation AND human AND data) AND PUBYEAR >2014 AND (LIMIT-TO (DOCTYPE , "ar") OR LIMIT-TO (DOCTYPE , "cp")) AND (LIMIT-TO (SUBJAREA , "COMP") OR LIMIT-TO (SUBJAREA , "ENGI")) LIMIT-TO (DOCTYPE , "ar") OR LIMIT-TO (DOCTYPE, "cp")) AND (LIMIT-TO (SUBJAREA, "COMP") OR LIMIT-TO (SUBJAREA, "ENGI"))
TITLE-ABS-KEY (multi-modal AND gesture AND generation) AND PUBYEAR >2014
TITLE-ABS-KEY (multi-modal AND gesture AND generation) AND PUBYEAR >2014 AND (LIMIT-TO (DOCTYPE[52], "cp"[53]) OR LIMIT-TO (OCTYPE , "ar"[54]))
TITLE-ABS-KEY (head AND gesture AND generation) AND PUBYEAR >2014
ACM
AllField[55]:(face) AND AllField:(data-driven) AND AllField:(generation) AND AllField:(visual prosody) AND [Publication Date: (01/01/2014 TO 12/31/2020)]
[All: data-driven hand gesture generation] AND [Publication Date: (01/01/2014 TO 12/31/2020)]
IEEE
((“All Metadata":facial) AND “All Metadata":generation) AND “All Metadata":data-driven) Year range: 2014-2020
((“All Metadata":face) AND “All Metadata":generation) AND “All Metadata":data-driven) Filter for year range = 2014-2020 Filter: journals
((“All Metadata":fac*) AND “All Metadata":generation) AND “All Metadata":data-driven) Year range=2014-2020

REFERENCES

[1] [n. d.]. KSS Dataset: Korean Single Speaker Speech Dataset. 6

[2] Henny Admoni and Brian Scassellati. 2017. Social eye gaze in human-robot interaction: a review. *Journal of Human-Robot Interaction* 6, 1 (2017), 25–63. 9

[3] Chaitanya Ahuja, Dong Won Lee, Yukiko I. Nakano, and Louis-Philippe Morency. 2020. Style transfer for co-speech gesture animation: A multi-speaker conditional-mixture approach. In *Computer Vision – ECCV 2020*, Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (Eds.). Springer International Publishing, Cham, 248–265. 6

[4] Niki Aifanti, Christos Papachristou, and Anastasios Delopoulos. 2010. The MUG facial expression database. In *11th International Workshop on Image Analysis for Multimedia Interactive Services WIAMIS 10*. IEEE, Desenzano del Garda, Italy, 1–4. 4.2, 2

[5] Simon Alexanderson, Gustav Eje Henter, Taras Kucherenko, and Jonas Beskow. 2020. Style-controllable speech-driven gesture synthesis using normalising flows. *Computer Graphics Forum* 39, 2 (2020), 487–496. https://doi.org/10.1111/cgf.13946 6, 6, 4

[6] D.M. Allen. 1971. Mean square error of prediction as a criterion for selecting variables. *Technometrics* 13, 3 (1971), 469–475. 3, 1, 2, 3, 6, 4, 7.2

[7] Jens Allwood. 1998. Cooperation and flexibility in multimodal communication. In *International Conference on Cooperative Multimodal Communication*. Springer, Berlin, Heidelberg, 113–124. 1

[8] S. Arik, M. Chrzanowski, A. Coates, G. Diamos, A. Gibiansky, Y. Kang, X. Li, J. Miller, J. Raiman, S. Sengupta, and M. Shoeybi. 2017. Deep Voice: Real-time neural text-to-speech. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*. JMLR.org, Sydney, NSW, Australia, 195–204. 7.1, 6

[9] Martin Arjovsky, Soumith Chintala, and Léon Bottou. 2017. Wasserstein generative adversarial networks. In *International Conference on Machine Learning (ICML'17)*. JMLR.org, Sydney, NSW, Australia, 214–223. 4.2

[10] A. D. Baddeley. 1986. *Working memory*. Oxford University Press, Oxford, UK. 7.1

[11] D. Bahdanau, K. Cho, and Y. Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*. San Diego, CA, USA. 7.1

[12] Sajal Chandra Banik, Chandima Dedduwa Pathiranage, Keigo Watanabe, and Kiyotaka Izumi. 2007. Behavior generation through interaction in an emotionally intelligent robot system. In *2007 International Conference on Industrial and Information Systems*. IEEE, 517–522. 1

[13] Christoph Bartneck, Tony Belpaeme, Friederike Eyssel, Takayuki Kanda, Merel Keijsers, and Selma Šabanović. 2020. *Human-robot interaction: An introduction*. Cambridge University Press, Cambridge. 1

[14] Sanjay Bilakhia, Stavros Petridis, Anton Nijholt, and Maja Pantic. 2015. The MAHNOB mimicry database: A database of naturalistic human interactions. *Pattern Recognition Letters* 66 (2015), 52–61. https://doi.org/10.1016/j.patrec.2015.03.005 Pattern Recognition in Human Computer Interaction. 2

[15] Mikołaj Bińkowski, Jeff Donahue, Sander Dieleman, Aidan Clark, Erich Elsen, Norman Casagrande, Luis C Cobo, and Karen Simonyan. 2019. High fidelity speech synthesis with adversarial networks. *arXiv preprint arXiv:1909.11646* (2019). 7.1

[16] Ali Borji. 2019. Pros and cons of GAN evaluation measures. *Computer Vision and Image Understanding* 179 (2019), 41–65. https://doi.org/10.1016/j.cviu.2018.10.009 4.2

[17] Elif Bozkurt, Engin Erzin, and Yücel Yemez. 2015. Affect-expressive hand gestures synthesis and animation. In *2015 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, Istanbul, Turkey, 1–6. https://doi.org/10.1109/ICME.2015.7177478 5, 3, 5

[18] J. Bradbury, S. Merity, C. Xiong, and R. Socher. 2017. Quasi-recurrent neural networks. In *Proceedings of the 5th International Conference on Learning Representations (ICLR)*. Toulon, France. 7.1

[19] ITUR BS. 2015. "Method for the subjective assessment of intermediate quality level of audio systems,". *International Telecommunication Union, Geneva, Switzerland* (2015). 4

[20] C. Busso, M. Bulut, C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. Chang, S. Lee, and S. Narayanan. 2008. IEMOCAP: Interactive emotional dyadic motion capture database. *Journal of Language Resources and Evaluation* 42, 4 (2008), 335–359. 1, 2, 2, 6, 4, 6

[21] C. Busso and S. Narayanan. 2007. Interrelation between speech and facial gestures in emotional utterances: A single subject study. *IEEE Transactions on Audio, Speech, and Language Processing* 15, 8 (2007), 2331–2347. 6, 5

[22] S-W. Byun and S-P Lee. 2021. Design of a multi-condition emotional speech synthesizer. *Applied Science* 11, 3 (2021). https://doi.org/10.3390/app11031144 7.2, 6, 7.2

[23] Houwei Cao, David G. Cooper, Michael K. Keutmann, Ruben C. Gur, Ani Nenkova, and Ragini Verma. 2014. CREMA-D: Crowd-sourced emotional multimodal actors dataset. *IEEE Transactions on Affective Computing* 5, 4 (2014), 377–390. https://doi.org/10.1109/TAFFC.2014.2336244 6, 5

[47] TS=Title Search

[48] PY=Publication Year

[49] Search by Title-Abstract-Keyword

[50] Publication Year

[51] Behaviour gave the same result

[52] Document type

[53] Conference paper

[54] Article

[55] Search by all fields

[24] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2017. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7291–7299. 5

[25] C. Chen, L.B. Hensel, Y. Duan, R.A.A. Ince, O.G.B. Garrod, J. Beskow, R.E. Jack, and P.G. Schyns. 2019. Equipping social robots with culturally-sensitive facial expressions of emotion using data-driven methods. In *2019 14th IEEE International Conference on Automatic Face Gesture Recognition (FG 2019)*. Institute of Electrical and Electronics Engineers Inc., Jack, R.E.; Institute of Neuroscience and Psychology, United Kingdom; email: rachael.jack@glasgow.ac.uk. https://doi.org/10.1109/FG.2019.8756570 4, 4.2

[26] Lele Chen, Zhiheng Li, Ross K Maddox, Zhiyao Duan, and Chenliang Xu. 2018. Lip movements generation at a glance. In *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer International Publishing, Cham, 538–553. 6, 6, 4, 5

[27] Lele Chen, Ross K Maddox, Zhiyao Duan, and Chenliang Xu. 2019. Hierarchical cross-modal talking face generation with dynamic pixel-wise loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, Rochester, USA, 7832–7841. https://doi.org/10.1109/CVPR.2019.00802 6

[28] Nanxin Chen, Yu Zhang, Heiga Zen, Ron J. Weiss, Mohammad Norouzi, and William Chan. 2021. WaveGrad: Estimating gradients for waveform generation. *ArXiv* abs/2009.00713 (2021). 7.1, 6

[29] 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*. Springer, Playa Vista, USA, 127–140. 5, 3, 5, 6

[30] Chung-Cheng Chiu and Stacy Marsella. 2011. A style controller for generating virtual human behaviors. In *The 10th International Conference on Autonomous Agents and Multiagent Systems - Volume 3* (Taipei, Taiwan) *(AAMAS '11)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 1023–1030. 5

[31] Chung-Cheng Chiu and Stacy Marsella. 2014. Gesture generation with low-dimensional embeddings. In *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 781–788. 6, 4

[32] Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Doha, Qatar, 1724–1734. https://doi.org/10.3115/v1/D14-1179 5, 6, 7.1

[33] Wei Chu and Abeer Alwan. 2009. Reducing f0 frame error of f0 tracking algorithms under noisy conditions with an unvoiced/voiced classification frontend. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 3969–3972. 6

[34] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio. 2014. Empirical evaluation of Gated Recurrent Neural Networks on sequence modeling. In *Proceedings of the Deep Learning and Representation Learning Workshop at the 28th International Conference on Neural Information Processing Systems (NIPS)*. Montreal, Canada. 41

[35] Junyoung Chung, Kyle Kastner, Laurent Dinh, Kratarth Goel, Aaron C Courville, and Yoshua Bengio. 2015. A recurrent latent variable model for sequential data. In *Advances in Neural Information Processing Systems*, C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett (Eds.), Vol. 28. Curran Associates, Inc., Cambridge, MA, USA. https://proceedings.neurips.cc/paper/2015/file/b618c3210e934362ac261db280128c22-Paper.pdf 5

[36] J.S. Chung and A. Zisserman. 2017. Lip reading in the wild. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 10112 LNCS (2017), 87–103. https://doi.org/10.1007/978-3-319-54184-6_6 6, 6, 4, 5

[37] Joon Son Chung, Amir Jamaludin, and Andrew Zisserman. 2017. You said that?. In *British Machine Vision Conference*. 6, 6, 4

[38] Joon Son Chung and Andrew Zisserman. 2016. Out of time: Automated lip sync in the wild. In *Asian conference on computer vision*. Springer, 251–263. 5

[39] Michael M Cohen, Rashid Clark, and Dominic W Massaro. 2001. Animated speech: Research progress and applications. In *AVSP 2001-International Conference on Auditory-Visual Speech Processing*. AVSP 2001, Santa Cruz, CA, USA. 4.2

[40] M. Cooke, J. Barker, S. Cunningham, and Xu Shao. 2006. An audio-visual corpus for speech perception and automatic speech recognition. *The Journal of the Acoustical Society of America* 120 5 Pt 1 (2006), 2421–4. 6, 6, 4, 5

[41] Timothy F. Cootes, Gareth J. Edwards, and Christopher J. Taylor. 2001. Active appearance models. *IEEE Transactions on pattern analysis and machine intelligence* 23, 6 (2001), 681–685. 6

[42] Corinna Cortes and Vladimir Vapnik. 1995. Support-Vector Networks. *Mach. Learn.* 20, 3 (Sept. 1995), 273–297. https://doi.org/10.1023/A:1022627411411 6

[43] Sara Dahmani, Vincent Colotte, Valérian Girard, and Slim Ouni. 2019. Conditional variational auto-encoder for text-driven expressive audiovisual speech synthesis. In *INTERSPEECH 2019-20th Annual Conference of the International Speech Communication Association*. 6, 4, 4

[44] Catherine Dehon, Peter Filzmoser, and Christophe Croux. 2000. Robust methods for canonical correlation analysis. In *Data Analysis, Classification, and Related Methods*, Henk A. L. Kiers, Jean-Paul Rasson, Patrick J. F. Groenen, and Martin Schader (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 321–326. 6

[45] C. Ding, L. Xie, and P. Zhu. 2014. Head motion synthesis from speech using deep neural networks. *Multimedia Tools and Applications* 74, 22 (2014). 3, 1, 3

[46] C. Ding, P. Zhu, and L. Xie. 2015. BLSTM neural networks for speech driven head motion synthesis. In *Proceedings of the 16th Conference of the International Speech Communication Association (INTERSPEECH)*. INTERSPEECH 2015, Dresden, Germany. 3, 1, 3, 2

[47] C. Ding, P. Zhu, L. Xie, D. Jiang, and Z. Fu. 2014. Speech-Driven head motion synthesis using neural networks. In *Proceedings of the 15th Conference of the International Speech Communication Association (INTERSPEECH)*. Singapore. 3

[48] Yu Ding, Catherine Pelachaud, and Thierry Artieres. 2013. Modeling multimodal behaviors from speech prosody. In *International Workshop on Intelligent Virtual Agents*. Springer, Berlin Heidelberg, 217–228. 6, 4

[49] Yu Ding, Ken Prepin, Jing Huang, Catherine Pelachaud, and Thierry Artières. 2014. Laughter animation synthesis. In *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 773–780. 6, 4

[50] P Ekman. 1976. Measuring facial movement. *Environmental Psychology and Nonverbal Behavior* 1 (1976), 56–75. 4

[51] P. Ekman and W. V. Friesen. 1978. *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Consulting Psychologists Press, CA, USA. 4, 4.1

[52] P. Ekman, W. V. Friesen, and P. Ellsworth. 1982. What emotion categories or dimensions can observers judge from facial behavior? In *Emotion in the Human Face*, P. Ekman (Ed.). Cambridge University Press, NY, USA, 39–55. 4.2, 4.2, 7.2

[53] Paul Ekman and Dacher Keltner. 1997. Universal facial expressions of emotion. *Segerstrale U, P. Molnar P, eds. Nonverbal communication: Where nature meets culture* (1997), 27–46. 4, 4.1

[54] Kevin El Haddad. 2017. Nonverbal conversation expressions processing for human-agent interactions. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, Mons, Belgium, 601–605. 1

[55] Cathy Ennis, Rachel McDonnell, and Carol O'Sullivan. 2010. Seeing is believing: Body motion dominates in multisensory conversations. *ACM Trans. Graph.* 29, 4, Article 91 (July 2010), 9 pages. https://doi.org/10.1145/1778765.1778828 3, 4

[56] Irfan A. Essa and Alex Paul Pentland. 1997. Coding, analysis, interpretation, and recognition of facial expressions. *IEEE transactions on pattern analysis and machine intelligence* 19, 7 (1997), 757–763. 2

[57] Florian Eyben, Klaus R Scherer, Björn W Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Y Devillers, Julien Epps, Petri Laukka, Shrikanth S Narayanan, et al. 2015. The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *IEEE transactions on affective computing* 7, 2 (2015), 190–202. 4.2

[58] Bo Fan, Lijuan Wang, Frank K Soong, and Lei Xie. 2015. Photo-real talking head with deep bidirectional LSTM. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, Beijing, China, 4884–4888. 6, 4

[59] Bo Fan, Lei Xie, Shan Yang, Lijuan Wang, and Frank K Soong. 2016. A deep bidirectional LSTM approach for video-realistic talking head. *Multimedia Tools and Applications* 75, 9 (2016), 5287–5309. 4.2

[60] G. Fanelli, J. Gall, H. Romsdorfer, T. Weise, and L. Van Gool. 2010. A 3D audio-visual corpus of affective communication. *Trans. Multi.* 12, 6 (Oct. 2010), 591–598. https://doi.org/10.1109/TMM.2010.2052239 6, 4

[61] Mireille Fares. 2020. Towards multimodal human-like characteristics and expressive visual prosody in virtual agents. In *Proceedings of the 2020 International Conference on Multimodal Interaction* (Virtual Event, Netherlands) *(ICMI '20)*. Association for Computing Machinery, New York, NY, USA, 743–747. https://doi.org/10.1145/3382507.3421155 4, 4.2

[62] Ylva Ferstl and Rachel McDonnell. 2018. Investigating the use of recurrent motion modelling for speech gesture generation. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents*. Association for Computing Machinery, New York, NY, USA, 93–98. https://doi.org/10.1145/3267851.3267898 1, 6, 4

[63] Ylva Ferstl, Michael Neff, and Rachel McDonnell. 2019. Multi-objective adversarial gesture generation. In *Motion, Interaction and Games*. Association for Computing Machinery, New York, NY, USA, 1–10. 5, 3, 5

[64] R Fletcher. 1987. Practical optimization methods. *Chichester: John Wiley and Sons* (1987). 1, 2

[65] Terry N Flynn, Jordan J Louviere, Tim J Peters, and Joanna Coast. 2007. Best–worst scaling: what it can do for health care research and how to do it. *Journal of health economics* 26, 1 (2007), 171–189. 7.2, 6

[66] Shengli Fu, R. Gutierrez-Osuna, A. Esposito, P.K. Kakumanu, and O.N. Garcia. 2005. Audio/visual mapping with cross-modal hidden Markov models. *IEEE Transactions on Multimedia* 7, 2 (2005), 243–252. https://doi.org/10.1109/TMM.2005.843341 4.1

[67] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. 2015. A neural algorithm of artistic style. *arXiv preprint arXiv:1508.06576* (2015). 7.2

[68] Zoubin Ghahramani and Michael I. Jordan. 1997. Factorial Hidden Markov Models. *Mach. Learn.* 29, 2–3 (Nov. 1997), 245–273. https://doi.org/10.1023/A:1007425814087 6

[69] A. Gibiansky, S. Arik, G. Diamos, J. Miller, K. Peng, W. Ping, J. Raiman, and Y. Zhou. 2017. Deep Voice 2: Multi-speaker neural text-to-speech. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS)*. Long Beach, CA, USA, 2962–2970. 7.1

[70] Shiry Ginosar, Amir Bar, Gefen Kohavi, Caroline Chan, Andrew Owens, and Jitendra Malik. 2019. Learning individual styles of conversational gesture. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE Computer Society, Berkeley, 3497–3506. https://doi.org/10.1109/CVPR.2019.00361 5, 3, 5, 6

[71] X. Gonzalvo, S. Tazari, C. Chan, M. Becker, A. Gutkin, and H. Silen. 2016. Recent advances in Google real-time HMM-driven unit selection synthesizer. In *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*. San Francisco, USA. 7.1

[72] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. 2014. Generative adversarial nets. In *Proceedings of the 27th International Conference on Neural Information Processing Systems (NIPS): Advances in Neural Information Processing Systems*. Montreal, Canada. 3, 4, 4.2, 2, 5, 6

[73] Alex Graves. 2012. Supervised sequence labelling. In *Supervised sequence labelling with recurrent neural networks*. Springer, 5–13. 6, 4

[74] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd International Conference on Machine Learning* (Pittsburgh, Pennsylvania, USA) *(ICML '06)*. Association for Computing Machinery, New York, NY, USA, 369–376. https://doi.org/10.1145/1143844.1143891 7.1

[75] Alex Graves and Jürgen Schmidhuber. 2005. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks* 18, 5 (2005), 602–610. https://doi.org/10.1016/j.neunet.2005.06.042 IJCNN 2005. 5

[76] G. Gravier, J-F. Bonastre, E. Geoffrois, S. Galliano, K. McTait, and K. Choukri. 2004. The ESTER evaluation campaign for the rich transcription of French broadcast news. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*. European Language Resources Association (ELRA), Lisbon, Portugal. http://www.lrec-conf.org/proceedings/lrec2004/pdf/672.pdf 4

[77] D. Greenwood, S. Laycock, and I. Matthews. 2017. Predicting head pose from speech with a conditional variational autoencoder. In *Proceedings of the 18th Conference of the International Speech Communication Association (INTERSPEECH)*. Stockholm, Sweden. 3, 1, 6, 6

[78] D. Griffin and J. Lim. 1984. Signal estimation from modified short-time Fourier transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing* 32, 2 (1984), 236–243. 7.1

[79] Daniel Griffin and Jae Lim. 1984. Signal estimation from modified short-time Fourier transform. *IEEE Transactions on acoustics, speech, and signal processing* 32, 2 (1984), 236–243. 7.1

[80] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. 2017. Improved training of Wasserstein GANs. In *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2017/file/892c3b1c6dccd52936e27cbd0ff683d6-Paper.pdf 4.2, 2

[81] K. Haag and H. Shimodaira. 2015. The University of Edinburgh speaker personality and MoCap dataset. In *Proceedings of the Facial Analysis and Animation (FAA)*. Vienna Austria. 1

[82] K. Haag and H. Shimodaira. 2016. Bidirectional LSTM networks employing stacked bottleneck features for expressive speech-driven head motion synthesis. In *Proceedings of the 16th International Conference on Intelligent Virtual Agents (IVA)*. Springer International Publishing, Los Angeles, USA. https://doi.org/10.1007/978-3-319-47665-0_18 3, 1, 3

[83] D. Hardoon, S. Szedmak, and J. Shawe-Taylor. 2004. Canonical correlation analysis: An overview with application to learning methods. *Neural Computation* 16, 12 (2004), 2639–2664. 3, 1, 5, 3, 6, 6, 4, 5

[84] Naomi Harte and Eoin Gillen. 2015. TCD-TIMIT: An audio-visual corpus of continuous speech. *IEEE Transactions on Multimedia* 17, 5 (2015), 603–615. https://doi.org/10.1109/TMM.2015.2407694 6, 5

[85] Dai Hasegawa, Naoshi Kaneko, Shinichi Shirakawa, Hiroshi Sakuta, and Kazuhiko Sumi. 2018. Evaluation of speech-to-gesture generation using bi-directional LSTM network. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents*. Association for Computing Machinery, New York, NY, USA, 79–86. 1, 5, 3, 5, 3, 5

[86] K. He, X. Zhang, S. Ren, and J. Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, (CVPR)*. IEEE, NV, USA, 770–778. https://doi.org/10.1109/CVPR.2016.90 7.1

[87] Gustav Eje Henter, Simon Alexanderson, and Jonas Beskow. 2020. MoGlow: Probabilistic and controllable motion synthesis using normalising flows. *ACM Trans. Graph.* 39, 6, Article 236 (Nov. 2020), 14 pages. https://doi.org/10.1145/3414685.3417836 4.1, 6

[88] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (Long Beach, California, USA) *(NIPS'17)*. Curran Associates Inc., Red Hook, NY, USA, 6629–6640. 6, 4

[89] G. E. Hinton, S. Osindero, and Y-W Teh. 2006. A fast learning algorithm for Deep Belief Nets. *Neural Computation* 18, 7 (2006), 1527–1554. 3, 5

[90] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780. 4, 4.2, 5, 6

[91] G. Hofer and H. Shimodaira. 2007. Automatic head motion prediction from speech data. In *Proceedings of the 8th Conference of the International Speech Communication Association (INTERSPEECH)*, Vol. 2. Antwerp, Belgium, 722–725. https://doi.org/10.21437/Interspeech.2007-299 3

[92] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. 2018. Multimodal unsupervised image-to-image translation. In *Proceedings of the European conference on computer vision (ECCV)*. 172–189. 6

[93] X. Huang, M. Wang, and M. Gong. 2021. Fine-grained talking face generation with video reinterpretation. *Visual Computer* 37, 1 (2021), 95–105. https://doi.org/10.1007/s00371-020-01982-7 6, 5

[94] Y. Huang and S. M. Khan. 2017. DyadGAN: Generating facial expressions in dyadic interactions. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, Princeton, NJ, 2259–2266. https://doi.org/10.1109/CVPRW.2017.280 4.2, 4.2, 2

[95] Mohamed E Hussein, Marwan Torki, Mohammad A Gowayyed, and Motaz El-Saban. 2013. Human action recognition using a temporal hierarchy of covariance descriptors on 3d joint locations. In *Twenty-third international joint conference on artificial intelligence (IJCAI '13)*. AAAI Press, Beijing, China, 2466–2472. 5, 3

[96] Carlos T Ishi, Daichi Machiyashiki, Ryusuke Mikata, and Hiroshi Ishiguro. 2018. A speech-driven hand gesture generation method and evaluation in android robots. *IEEE Robotics and Automation Letters* 3, 4 (2018), 3757–3764. 5

[97] Keith Ito and Linda Johnson. 2017. The LJ speech dataset. https://keithito.com/LJ-Speech-Dataset/. 6

[98] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. 2016. Perceptual losses for real-time style transfer and super-resolution. In *Computer Vision – ECCV 2016*. Springer International Publishing, Amsterdam, The Netherlands, 694–711. 7.2

[99] Patrik Jonell, Taras Kucherenko, Gustav Eje Henter, and Jonas Beskow. 2020. Let's face it: probabilistic multi-modal interlocutor-aware generation of facial gestures in dyadic settings. In *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents* (Virtual Event, Scotland,

UK) *(IVA '20)*. Association for Computing Machinery, New York, NY, USA, Article 31, 8 pages. https://doi.org/10.1145/3383652.3423911 4.1, 2

[100] Nal Kalchbrenner, Erich Elsen, Karen Simonyan, Seb Noury, Norman Casagrande, Edward Lockhart, Florian Stimberg, Aaron Oord, Sander Dieleman, and Koray Kavukcuoglu. 2018. Efficient neural audio synthesis. In *International Conference on Machine Learning*. PMLR, 2410–2419. 7.1

[101] Tero Karras, Timo Aila, Samuli Laine, Antti Herva, and Jaakko Lehtinen. 2017. Audio-driven facial animation by joint end-to-end learning of pose and emotion. *ACM Transactions on Graphics (TOG)* 36, 4 (2017), 1–12. 2, 4.2, 2

[102] Jaehyeon Kim, Sungwon Kim, Jungil Kong, and Sungroh Yoon. 2020. Glow-TTS: A generative flow for text-to-speech via monotonic alignment search. *Advances in Neural Information Processing Systems* 33 (2020), 8067–8077. 6

[103] Taehwan Kim, Yisong Yue, Sarah Taylor, and Iain Matthews. 2015. A decision tree framework for spatiotemporal sequence prediction. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Sydney, NSW, Australia) *(KDD '15)*. Association for Computing Machinery, New York, NY, USA, 577–586. https://doi.org/10.1145/2783258.2783356 4.1

[104] S. King and Vasilis Karaiskos. 2011. The Blizzard challenge 2011. 7.2, 6

[105] Durk P Kingma and Prafulla Dhariwal. 2018. Glow: Generative flow with invertible 1x1 convolutions. In *Advances in Neural Information Processing Systems*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Eds.), Vol. 31. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2018/file/d139db6a236200b21cc7f752979132d0-Paper.pdf 4.1

[106] Durk P Kingma and Prafulla Dhariwal. 2018. Glow: Generative flow with invertible 1x1 convolutions. *Advances in neural information processing systems* 31 (2018). 7.1

[107] Dietrich Klakow and Jochen Peters. 2002. Testing the correlation of word error rate and perplexity. *Speech Communication* 38, 1-2 (2002), 19–28. 5

[108] Ron Kohavi and Roger Longbotham. 2017. Online Controlled Experiments and A/B Testing. *Encyclopedia of machine learning and data mining* 7, 8 (2017), 922–929. 1, 2, 6, 4, 7.2, 6

[109] S. Kopp and I. Wachsmuth. 2004. Synthesizing multimodal utterances for conversational agents. *Computer Animation and Virtual Worlds* 15, 1 (2004), 39–52. 1

[110] Robert Kubichek. 1993. Mel-cepstral distance measure for objective speech quality assessment. In *Proceedings of IEEE pacific rim conference on communications computers and signal processing*, Vol. 1. IEEE, 125–128. 7.2, 6

[111] Taras Kucherenko. 2018. Data driven non-verbal behavior generation for humanoid robots. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction* (Boulder, CO, USA) *(ICMI '18)*. Association for Computing Machinery, New York, NY, USA, 520–523. https://doi.org/10.1145/3242969.3264970 1, 5

[112] 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*. Association for Computing Machinery, New York, NY, USA, 97–104. https://doi.org/10.1145/3308532.3329472 5, 3, 5, 6, 37

[113] Taras Kucherenko, Patrik Jonell, Sanne van Waveren, Gustav Eje Henter, Simon Alexandersson, Iolanda Leite, and Hedvig Kjellström. 2020. Gesticulator: A Framework for semantically-aware speech-driven gesture generation. In *Proceedings of the ACM International Conference on Multimodal Interaction*. Association for Computing Machinery, New York, NY, USA, 242–250. https://doi.org/10.1145/3382507.3418815 5, 6, 6, 6, 4

[114] Taras Kucherenko, Patrik Jonell, Youngwoo Yoon, Pieter Wolfert, and Gustav Eje Henter. 2021. A large, crowdsourced evaluation of gesture generation systems on common data: The GENEA Challenge 2020. In *26th international conference on intelligent user interfaces*. 11–21. 4, 8, 8

[115] S. Kullback and R. A. Leibler. 1951. On Information and Sufficiency. *The Annals of Mathematical Statistics* 22, 1 (1951), 79 – 86. https://doi.org/10.1214/aoms/1177729694 3

[116] Kundan Kumar, Rithesh Kumar, Thibault de Boissiere, Lucas Gestin, Wei Zhen Teoh, Jose Sotelo, Alexandre de Brébisson, Yoshua Bengio, and Aaron C Courville. 2019. Melgan: Generative adversarial networks for conditional waveform synthesis. *Advances in neural information processing systems* 32 (2019). 7.1

[117] Jonathan Lam, Bill Kapralos, Kc Collins, Andrew Hogue, and Kamen Kanev. 2010. Amplitude panning-based sound system for a horizontal surface computer: A user-based study. (10 2010). https://doi.org/10.1109/HAVE.2010.5623999 5, 5, 3

[118] Gilwoo Lee, Zhiwei Deng, Shugao Ma, Takaaki Shiratori, Siddhartha S Srinivasa, and Yaser Sheikh. 2019. Talking with hands 16.2 M: A large-scale dataset of synchronized body-finger motion and audio for conversational motion analysis and synthesis. In *ICCV*. IEEE, Seoul, South Korea, 763–772. 5, 3, 5

[119] Jason Lee, Kyunghyun Cho, and Thomas Hofmann. 2017. Fully character-level neural machine translation without explicit segmentation. *Transactions of the Association for Computational Linguistics* 5 (2017), 365–378. 7.2

[120] Y. Lee, A. Rabiee, and S-Y Lee. 2017. Emotional end-to-end neural speech synthesizer. In *Proceedings of the International Conference on Neural Information Processing Systems (NIPS)*. Long Beach, CA, USA. 7.2, 7.2, 6

[121] Y. Lei, S. Yang, and L. Xie. 2021. Fine-grained emotion strength transfer, control and prediction for emotional speech synthesis. In *2021 IEEE Spoken Language Technology Workshop, SLT 2021 - Proceedings*. 423–430. https://doi.org/10.1109/SLT48900.2021.9383524 7.2, 6

[122] T. Li, S. Yang, L. Xue, and L. Xie. 2021. Controllable emotion transfer for end-to-end speech synthesis. In *2021 12th International Symposium on Chinese Spoken Language Processing, ISCSLP 2021*. https://doi.org/10.1109/ISCSLP49672.2021.9362069 7.2, 6

[123] Xu Li, Zhiyong Wu, Helen M Meng, Jia Jia, Xiaoyan Lou, and Lianhong Cai. 2016. Expressive speech-driven talking avatar synthesis with DBLSTM using limited amount of emotional bimodal data.. In *Interspeech*. 1477–1481. 6, 4

[124] Kyle Lindgren, Niveditha Kalavakonda, David E Caballero, Kevin Huang, and Blake Hannaford. 2018. Learned hand gesture classification through synthetically generated training samples. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 3937–3942. 1

[125] Phoebe Liu, Dylan F Glas, Takayuki Kanda, and Hiroshi Ishiguro. 2016. Data-driven HRI: Learning social behaviors by example from human–human interaction. *IEEE Transactions on Robotics* 32, 4 (2016), 988–1008. 1

[126] R. Liu, B. Sisman, G.L. Gao, and H. Li. 2021. Expressive TTS training with frame and style reconstruction loss. *IEEE/ACM Transactions on Audio Speech and Language Processing* (2021). https://doi.org/10.1109/TASLP.2021.3076369 7.2, 6

[127] Yu Liu, Gelareh Mohammadi, Yang Song, and Wafa Johal. 2021. Speech-based gesture generation for robots and embodied agents: A scoping review. In *Proceedings of the 9th International Conference on Human-Agent Interaction*. 31–38. 1

[128] Manja Lohse, Reinier Rothuis, Jorge Gallego-Pérez, Daphne E Karreman, and Vanessa Evers. 2014. Robot gestures make difficult tasks easier: The impact of gestures on perceived workload and task performance. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1459–1466. 5

[129] Patrick Lucey, Jeffrey F Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews. 2010. The extended Cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression. In *2010 ieee computer society conference on computer vision and pattern recognition-workshops*. IEEE, Pittsburgh, PA, USA, 94–101. 4.2, 2

[130] Maurizio Mancini, Beatrice Biancardi, Florian Pecune, Giovanna Varni, Yu Ding, Catherine Pelachaud, Gualtiero Volpe, and Antonio Camurri. 2017. Implementing and evaluating a laughing virtual character. *ACM TRANSACTIONS ON INTERNET TECHNOLOGY* 17, 1, SI (MAR 2017). https://doi.org/{10.1145/2998571} 4

[131] Christian Mandery, Omer Terlemez, Martin Do, Nikolaus Vahrenkamp, and Tamim Asfour. 2015. The KIT whole-body human motion database. In *2015 International Conference on Advanced Robotics (ICAR)*. IEEE, Germany, 329–336. https://doi.org/10.1109/ICAR.2015.7251476 3

[132] 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. 3, 4, 6, 6, 4

[133] Olivier Martin, Irene Kotsia, Benoit Macq, and Ioannis Pitas. 2006. The eNTERFACE'05 audio-visual emotion database. In *22nd International Conference on Data Engineering Workshops (ICDEW'06)*. IEEE, 8–8. 4

[134] David McNeill. 1992. *Hand and mind: What gestures reveal about thought.* University of Chicago press. 5, 5

[135] Angeliki Metallinou, Chi-Chun Lee, Carlos Busso, Sharon Carnicke, and Shrikanth Narayanan. 2010. The USC creativeit database: A multimodal database of theatrical improvisation. In *Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality 18 May 2010*. 5, 3, 26

[136] Robert C Moore. 2004. On log-likelihood-ratios and the significance of rare events. In *Proceedings of the 2004 conference on empirical methods in natural language processing*. 333–340. 6, 5

[137] Meinard Müller. 2007. Dynamic time warping. *Information retrieval for music and motion* (2007), 69–84. 7.2, 6

[138] A. Nagrani, J. S. Chung, and A. Zisserman. 2017. VoxCeleb: A large-scale speaker identification dataset. In *INTERSPEECH*. Oxford, UK. 4

[139] Niranjan D. Narvekar and Lina J. Karam. 2009. A no-reference perceptual image sharpness metric based on a cumulative probability of blur detection. In *2009 International Workshop on Quality of Multimedia Experience*. IEEE, Tempe, AZ, USA, 87–91. https://doi.org/10.1109/QOMEX.2009.5246972 6, 5

[140] Niranjan D Narvekar and Lina J Karam. 2011. A no-reference image blur metric based on the cumulative probability of blur detection (CPBD). *IEEE Transactions on Image Processing* 20, 9 (2011), 2678–2683. 6, 4

[141] M. Neff, M. Kipp, I. Albrecht, and H. P. Seidel. 2008. Gesture modeling and animation based on a probabilistic recreation of speaker style. *ACM Transactions on Graphics* 27, 1 (2008), 1–24. 1

[142] Joshua R New, Erion Hasanbelliu, and Mario Aguilar. 2003. Facilitating user interaction with complex systems via hand gesture recognition. In *Proceedings of the 2003 Southeastern ACM Conference, Savannah, GA*. 5

[143] Magalie Ochs and Catherine Pelachaud. 2012. Model of the perception of smiling virtual character. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems - Volume 1* (Valencia, Spain) *(AAMAS '12)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 87–94. 6, 4

[144] Naima Otberdout, Mohammed Daoudi, Anis Kacem, Lahoucine Ballihi, and Stefano Berretti. 2020. Dynamic facial expression generation on Hilbert hypersphere with conditional Wasserstein generative adversarial nets. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2020), 1–1. https://doi.org/10.1109/TPAMI.2020.3002500 4.2, 15, 2

[145] Algirdas Pakstas, Robert Forchheimer, and Igor S. Pandzic. 2002. *MPEG-4 facial animation: The standard, implementation and applications.* John Wiley & Sons, Inc., USA. 6

[146] Maja Pantic. 2009. Machine analysis of facial behaviour: Naturalistic and dynamic behaviour. *Philosophical Transactions of the Royal Society B: Biological Sciences* 364, 1535 (2009), 3505–3513. 4

[147] Maja Pantic, Roderick Cowie, Francesca D'Errico, Dirk Heylen, Marc Mehu, Catherine Pelachaud, Isabella Poggi, Marc Schroeder, and Alessandro Vinciarelli. 2011. Social signal processing: The research agenda. In *Visual analysis of humans*. Springer, 511–538. 1

[148] Maja Pantic and Leon J. M. Rothkrantz. 2000. Automatic analysis of facial expressions: The state of the art. *IEEE Transactions on pattern analysis and machine intelligence* 22, 12 (2000), 1424–1445. 2

[149] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, 311–318. https://doi.org/10.3115/1073083.1073135 5, 6, 4

[150] Patrick Pérez, Michel Gangnet, and Andrew Blake. 2003. Poisson image editing. In *ACM SIGGRAPH 2003 Papers*. Association for Computing Machinery, New York, NY, USA, 313–318. https://doi.org/10.1145/1201775.882269 6

[151] W. Ping, K. Peng, A. Gibiansky, S. Arik, A. Kannan, S. Narang, J. Raiman, and J. Miller. 2018. Deep Voice 3: Scaling text-to-speech with convolutional sequence learning. In *Proceedings of the 6rd International Conference on Learning Representations (ICLR)*. International Conference on Learning Representations, ICLR, Vancouver, Canada. 7.1

[152] Matthias Plappert, Christian Mandery, and Tamim Asfour. 2016. The KIT motion-language dataset. *Big Data* 4, 4 (Dec 2016), 236–252. https://doi.org/10.1089/big.2016.0028 4

[153] M. Plappert, C. Mandery, and T. Asfour. 2018. Learning a bidirectional mapping between human whole-body motion and natural language using deep recurrent neural networks. *Robotics and Autonomous Systems* 109 (2018), 13–26. https://doi.org/10.1016/j.robot.2018.07.006 6, 4

[154] Kishore Prahallad, Anandaswarup Vadapalli, Naresh Elluru, Gautam Mantena, Bhargav Pulugundla, Peri Bhaskararao, Hema A Murthy, Simon King, Vasilis Karaiskos, and Alan W Black. 2013. The blizzard challenge 2013–Indian language task. In *Blizzard challenge workshop*, Vol. 2013. 6

[155] Ryan Prenger, Rafael Valle, and Bryan Catanzaro. 2019. Waveglow: A Flow-based generative network for speech synthesis. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 3617–3621. https://doi.org/10.1109/ICASSP.2019.8683143 7.1

[156] F. Ribeiro, D. Florencio, C. Zhang, and M. Seltzer. 2011. CROWDMOS: An approach for crowdsourcing mean opinion score studies. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, Prague, Czech Republic, 2416–2419. https://doi.org/10.1109/ICASSP.2011.5946971 1, 7.1, 7.2, 6

[157] Carolyn Richie, Sarah Warburton, and Megan Carter. 2009. *Audiovisual database of spoken American English*. Linguistic Data Consortium, Philadelphia. 6, 4

[158] K. Richmond, P. Hoole, and S. King. 2011. Announcing the electromagnetic articulography (Day 1) subset of the MNGU0 articulatory corpus. In *Proceedings of the 12th Conference of the International Speech Communication Association (INTERSPEECH)*. Interspeech 2011, Florence, Italy. 1, 2

[159] J. L. Rodgers and W. A. Nicewander. 1988. Thirteen ways to look at the correlation coefficient. *The American Statistician* 42, 1 (1988), 59–66. 3, 1

[160] Matej Rojc, Izidor Mlakar, and Zdravko Kačič. 2017. The TTS-driven affective embodied conversational agent EVA, based on a novel conversational-behavior generation algorithm. *Engineering Applications of Artificial Intelligence* 57 (2017), 80–104. 1, 6

[161] E. L. Rosenberg and P. Ekman. 1997. *What the face reveals: Basic and applied studies of spontaneous expression using the facial action coding system (FACs)*. Oxford University Press, New York. 4.1, 7

[162] Sam Roweis. 1997. EM algorithms for PCA and SPCA. *Advances in neural information processing systems* 10 (1997). 6, 4

[163] 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*. Springer International Publishing, Cham, 389–402. 4.2, 2

[164] Najmeh Sadoughi and Carlos Busso. 2018. Expressive speech-driven lip movements with multitask learning. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, Richardson, TX, USA, 409–415. 4, 4.2, 4.2, 2

[165] N. Sadoughi and C. Busso. 2018. Novel realizations of speech-driven head movements with generative adversarial networks. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, Calgary, Canada, 6169–6173. https://doi.org/10.1109/ICASSP.2018.8461967 3, 1, 3, 2

[166] Najmeh Sadoughi and Carlos Busso. 2019. Speech-driven animation with meaningful behaviors. *Speech Communication* 110 (2019), 90–100. 6, 5

[167] N. Sadoughi and C. Busso. 2019. Speech-driven expressive talking lips with conditional sequential generative adversarial networks. *IEEE Transactions on Affective Computing* (2019), 1–1. https://doi.org/10.1109/TAFFC.2019.2916031 4.2, 2, 4.2

[168] Najmeh Sadoughi, Yang Liu, and Carlos Busso. 2015. MSP-AVATAR corpus: Motion capture recordings to study the role of discourse functions in the design of intelligent virtual agents. In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, Vol. 7. IEEE, Dallas, USA, 1–6. 5

[169] Masaki Saito, Eiichi Matsumoto, and Shunta Saito. 2017. Temporal generative adversarial nets with singular value clipping. In *2017 IEEE International Conference on Computer Vision (ICCV)*, Vol. 2017-October. Institute of Electrical and Electronics Engineers Inc., Japan, 2849–2858. https://doi.org/10.1109/ICCV.2017.308 4.2

[170] Maha Salem and Kerstin Dautenhahn. 2017. Social signal processing in social robotics. *Social signal processing* (2017), 317. 1, 5

[171] Philip Schatz. 2011. *Forced-Choice Test*. Springer New York, New York, NY, 1067–1067. https://doi.org/10.1007/978-0-387-79948-3_183 2

[172] M. Schuster and K. K. Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing* 45, 11 (1997), 2673–2681. 3

[173] Iulian Serban, Ryan Lowe, Peter Henderson, Laurent Charlin, and Joelle Pineau. 2018. A survey of available corpora for building data-driven dialogue systems. *ArXiv* abs/1512.05742 (2018). 1

[174] Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, et al. 2018. Natural TTS synthesis by conditioning WaveNet on Mel spectrogram predictions. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 4779–4783. 6

[175] J. Shen, R. Pang, R-J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, RJ. Skerry-Ryan, R. Saurous, Y. Agiomyrgiannakis, and Y. Wu. 2018. Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Calgary, Canada. 7.1, 7.2

[176] Eli Shlizerman, Lucio Dery, Hayden Schoen, and Ira Kemelmacher-Shlizerman. 2018. Audio to body dynamics. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, Washington, 7574–7583. https://doi.org/10.1109/CVPR.2018.00790 5

[177] S. Sinha, S. Biswas, and B. Bhowmick. 2020. Identity-preserving realistic talking face generation. In *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE, Glasgow, UK. https://doi.org/10.1109/IJCNN48605.2020.9206665 6, 5

[178] P. Smolensky. 1986. Information processing in dynamical systems: Foundations of harmony theory. In *Parallel Distributed Processing: Explorations in the Microstructure of Cognition: Foundations*, D. E. Rumelhart and J. L. McClelland (Eds.). MIT Press, Cambridge, USA, 194–281. 3

[179] K. Sohn, X. Yan, and H. Lee. 2015. Learning structured output representation using deep conditional generative models. In *Proceedings of the 28th International Conference on Neural Information Processing Systems (NIPS): Advances in Neural Information Processing Systems*. Curran Associates, Inc., Montreal, Canada. 3

[180] J. Sotelo, S. Mehri, K. Kumar, J. Santos, K. Kastner, A. Courville, and Y. Bengio. 2017. Char2Wav: End-to-end speech synthesis. In *ICLR workshop track*. 7.1

[181] R-K. Srivastava, K. Greff, and J. Schmidhuber. 2015. Highway networks. In *Proceedings of the Deep Learning Workshop at the 32nd International Conference on Machine Learning (ICML)*. Lille, France. 41

[182] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2* (Montreal, Canada) *(NIPS'14)*. MIT Press, Cambridge, MA, USA, 3104–3112. 6, 7.1

[183] 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), 1–13. 6, 6, 6, 4

[184] V. Sze, Y-H. Chen, T-J. Yang, and J. S. Emer. 2017. Efficient processing of Deep Neural Networks: A tutorial and survey. *Proc. IEEE* 105, 12 (2017), 2295–2329. 3

[185] Y. Taigman, L. Wolf, A. Polyak, and E. Nachmani. 2018. VoiceLoop: Voice fitting and synthesis via a phonological loop. In *Proceedings of the 6rd International Conference on Learning Representations (ICLR)*. Vancouver, Canada. 7.1, 6, 7.2

[186] 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 *Proceedings of the 5th International Conference on Human Agent Interaction*. 365–369. 5, 3, 5

[187] 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*. Springer International Publishing, Cham, 198–202. 3

[188] Sarah Taylor, Akihiro Kato, Iain Matthews, and Ben Milner. 2016. Audio-to-visual speech conversion using deep neural networks. In *Interspeech 2016*. International Speech and Communication Association, 1482–1486. https://doi.org/10.21437/Interspeech.2016-483 4.1, 4.2, 2

[189] Sarah L. Taylor, Moshe Mahler, Barry-John Theobald, and Iain Matthews. 2012. Dynamic units of visual speech. In *Proceedings of the ACM SIGGRAPH/Eurographics Symposium on Computer Animation* (Lausanne, Switzerland) *(SCA '12)*. Eurographics Association, Goslar, DEU, 275–284. 4.1, 2

[190] Rafael Luiz Testa, Cleber Gimenez Correa, Ariane Machado-Lima, and Fatima L. S. Nunes. 2019. Synthesis of facial expressions in photographs: characteristics, approaches, and challenges. *ACM COMPUTING SURVEYS* 51, 6 (FEB 2019). https://doi.org/{10.1145/3292652} 4

[191] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. 2016. Face2face: Real-time face capture and reenactment of RGB videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2387–2395. 6

[192] George Trigeorgis, Fabien Ringeval, Raymond Brueckner, Erik Marchi, Mihalis A Nicolaou, Björn Schuller, and Stefanos Zafeiriou. 2016. Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network. In *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 5200–5204. 2

[193] Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, and Jan Kautz. 2018. MoCoGAN: Decomposing motion and content for video generation. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1526–1535. https://doi.org/10.1109/CVPR.2018.00165 4.2, 6, 5

[194] Nguyen Tan Viet Tuyen, Armagan Elibol, and Nak Young Chong. 2020. Learning from humans to generate communicative gestures for social robots. In *2020 17th International Conference on Ubiquitous Robots (UR)*. IEEE, 284–289. 5, 5, 3, 5

[195] S-Y Um, S. Oh, K. Byun, I. Jang, C. Ahn, and H-G Kang. 2020. Emotional speech synthesis with rich and granularized control. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Barcelona, Spain. 7.2, 6

[196] Jérôme Urbain, Radoslaw Niewiadomski, Elisabetta Bevacqua, Thierry Dutoit, Alexis Moinet, Catherine Pelachaud, Benjamin Picart, Joëlle Tilmanne, and Johannes Wagner. 2010. AVlaughter cycle. *Journal on Multimodal User Interfaces* 4, 1 (2010), 47–58. 6, 4

[197] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, A. Vinyals, O.and Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu. 2016. WaveNet: A generative model for raw audio. *CoRR* (2016). 1, 7.1, 6

[198] Aäron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alexander Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. 2016. WaveNet: A generative model for raw audio. In *Arxiv*. https://arxiv.org/abs/1609.03499 5

[199] A. van den Oord, N. Kalchbrenner, and K. Kavukcuoglu. 2016. Pixel recurrent neural networks. In *Proceedings of the 33rd International Conference on Machine Learning Research (PMLR)*. NY, USA. 7.1

[200] A. van den Oord, N. Kalchbrenner, O. Vinyals, L. Espeholt, A. Graves, and K. Kavukcuoglu. 2016. Conditional image generation with PixelCNN decoders. In *Proceedings of the 30th International Conference on Neural Information Processing Systems (NIPS)*. Barcelona, Spain. 7.1

[201] Aaron van den Oord, Yazhe Li, Igor Babuschkin, Karen Simonyan, Oriol Vinyals, Koray Kavukcuoglu, George Driessche, Edward Lockhart, Luis Cobo, Florian Stimberg, et al. 2018. Parallel WaveNet: Fast high-fidelity speech synthesis. In *International conference on machine learning*. PMLR, 3918–3926. 7.1

[202] Stef van der Struijk, Hung-Hsuan Huang, Maryam Sadat Mirzaei, and Toyoaki Nishida. 2018. FACSvatar: An open source modular framework for real-time FACS based facial animation. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents* (Sydney, NSW, Australia) *(IVA '18)*. Association for Computing Machinery, New York, NY, USA, 159–164. https://doi.org/10.1145/3267851.3267918 4.1, 4.2, 2

[203] Christophe Veaux, Junichi Yamagishi, Kirsten MacDonald, et al. 2017. Superseded-CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit. (2017). 6

[204] O. Vinyals, L. Kaiser, T. Koo, S. Petrov, I. Sutskever, and G. Hinton. 2015. Grammar as a foreign language. In *Proceedings of the the 29th International Conference on Neural Information Processing Systems (NIPS)*. Montreal, Canada. 7.1

[205] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. 2016. Generating videos with scene dynamics. In *Advances in Neural Information Processing Systems*, D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett (Eds.), Vol. 29. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2016/file/04025959b191f8f9de3f924f0940515f-Paper.pdf 4.2

[206] Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. 2019. Realistic speech-driven facial animation with gans. *International Journal of Computer Vision* (2019), 1–16. 4.1, 6, 5

[207] Tijana Vuletic, Alex Duffy, Laura Hay, Chris McTeague, Gerard Campbell, and Madeleine Grealy. 2019. Systematic literature review of hand gestures used in human computer interaction interfaces. *International Journal of Human-Computer Studies* 129 (2019), 74–94. 5

[208] Jack M. Wang, David J. Fleet, and Aaron Hertzmann. 2008. Gaussian process dynamical models for human motion. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 30, 2 (2008), 283–298. https://doi.org/10.1109/TPAMI.2007.1167 6

[209] Qiang Wang, Weiwei Zhang, Xiaoou Tang, and Heung-Yeung Shum. 2006. Real-time Bayesian 3D pose tracking. *IEEE Transactions on Circuits and Systems for Video Technology* 16, 12 (2006), 1533–1541. 6, 4

[210] Siyang Wang, Simon Alexanderson, Joakim Gustafson, Jonas Beskow, Gustav Eje Henter, and Éva Székely. 2021. Integrated speech and gesture synthesis. In *Proceedings of the 2021 International Conference on Multimodal Interaction*. 177–185. 6, 4

[211] Y. Wang, R. J. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. Le, Y. Agiomyrgiannakis, R. Clark, and R. A. Saurous. 2017. Tacotron: Towards end-to-end speech synthesis. In *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*. 1, 7.1, 7.2, 7.2, 6

[212] Yuxuan Wang, Daisy Stanton, Yu Zhang, RJ-Skerry Ryan, Eric Battenberg, Joel Shor, Ying Xiao, Ye Jia, Fei Ren, and Rif A. Saurous. 2018. Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis. In *Proceedings of the 35th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 80)*, Jennifer Dy and Andreas Krause (Eds.). PMLR, 5180–5189. http://proceedings.mlr.press/v80/wang18h.html 7.2

[213] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. 2004. Image quality assessment: From error visibility to structural similarity. *IEEE transactions on image processing* 13, 4 (2004), 600–612. 4.2, 2, 6, 6, 4, 5

[214] Paul J. Werbos. 1990. Consistency of HDP applied to a simple reinforcement learning problem. *Neural Networks* 3, 2 (1990), 179–189. https://doi.org/10.1016/0893-6080(90)90088-3 6

[215] Ronald J Williams and David Zipser. 1995. Gradient-based learning algorithms for recurrent networks and their computational complexity. *Backpropagation: Theory, architectures, and applications* 433 (1995), 17. 6, 4

[216] A.D. Wilson and A.F. Bobick. 1999. Parametric hidden Markov models for gesture recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 21, 9 (1999), 884–900. https://doi.org/10.1109/34.790429 6

[217] Pieter Wolfert, Jeffrey M Girard, Taras Kucherenko, and Tony Belpaeme. 2021. To rate or not to rate: Investigating evaluation methods for generated co-speech gestures. In *Proceedings of the ACM International Conference on Multimodal Interaction*. 8

[218] Pieter Wolfert, Taras Kucherenko, Hedvig Kjellström, and Tony Belpaeme. 2019. Should beat gestures be learned or designed?: A benchmarking user study. In *ICDL-EPIROB 2019 Workshop on Naturalistic Non-Verbal and Affective Human-Robot Interactions*. IEEE conference proceedings. 1

[219] Pieter Wolfert, Nicole Robinson, and Tony Belpaeme. 2022. A review of evaluation practices of gesture generation in embodied conversational agents. *IEEE Transactions on Human-Machine Systems* (2022). 8

[220] X. Wu, Y. Cao, H. Lu, S. Liu, S. Kang, Z. Wu, X. Liu, and H. Meng. 2021. Exemplar-based emotive speech synthesis. *IEEE/ACM Transactions on Audio Speech and Language Processing* 29 (2021), 874–886. https://doi.org/10.1109/TASLP.2021.3052688 7.2, 6

[221] Junichi Yamagishi. 2012. English multi-speaker corpus for CSTR voice cloning toolkit. *URL http://homepages. inf. ed. ac. uk/jyamagis/-page3/page58/page58. html* (2012). 6

[222] Ryuichi Yamamoto, Eunwoo Song, and Jae-Min Kim. 2020. Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6199–6203. 7.1

[223] Geng Yang, Shan Yang, Kai Liu, Peng Fang, Wei Chen, and Lei Xie. 2021. Multi-band MelGAN: Faster waveform generation for high-quality text-to-speech. In *2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 492–498. 7.1

[224] Yi Yang and Deva Ramanan. 2012. Articulated human detection with flexible mixtures of parts. *IEEE transactions on pattern analysis and machine intelligence* 35, 12 (2012), 2878–2890. 3

[225] Yi Yang and Deva Ramanan. 2013. Articulated human detection with flexible mixtures of parts. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35, 12 (2013), 2878–2890. https://doi.org/10.1109/TPAMI.2012.261 5

[226] Zijie Ye, Haozhe Wu, and Jia Jia. 2021. Human motion modeling with deep learning: A survey. *AI Open* (2021). 1

[227] Youngwoo Yoon, Bok Cha, Joo-Haeng Lee, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee. 2020. Speech gesture generation from the trimodal context of text, audio, and speaker identity. *ACM Transactions on Graphics (TOG)* 39, 6 (2020), 1–16. 6, 6, 4

[228] Y. Yoon, W-R Ko, M. Jang, J. Lee, J. Kim, and G. Lee. 2019. Robots learn social skills: End-to-end learning of co-speech gesture generation for humanoid robots. In *Proceedings of the International Conference on Robotics and Automation (ICRA)*. IEEE, Montreal, QC, Canada, 4303–4309. 1, 5, 5,

3, 5, 4

[229] Youngwoo Yoon, Pieter Wolfert, Taras Kucherenko, Carla Viegas, Teodor Nikolov, Mihail Tsakov, and Gustav Eje Henter. 2022. The GENEA Challenge 2022: A large evaluation of data-driven co-speech gesture generation. In *Proceedings of the 2022 International Conference on Multimodal Interaction*. 736–747. 8

[230] Shun-Zheng Yu. 2010. Hidden semi-Markov models. *Artificial Intelligence* 174, 2 (2010), 215–243. https://doi.org/10.1016/j.artint.2009.11.011 Special Review Issue. 5

[231] H. Zen, Y. Agiomyrgiannakis, N. Egberts, F. Henderson, and P. Szczepaniak. 2016. Fast, compact, and high quality LSTM-RNN based statistical parametric speech synthesizers for mobile devices. In *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*. San Francisco, USA. 7.1

[232] Shu Zhang, Dequan Zheng, Xinchen Hu, and Ming Yang. 2015. Bidirectional long short-term memory networks for relation classification. In *Proceedings of the 29th Pacific Asia conference on language, information and computation*. 73–78. 4.2, 5, 6

[233] Tianyi Zhang, Felix Wu, Arzoo Katiyar, Kilian Q Weinberger, and Yoav Artzi. 2020. Revisiting few-sample BERT fine-tuning. *arXiv preprint arXiv:2006.05987* (2020). 8

[234] Guoying Zhao, Xiaohua Huang, Matti Taini, Stan Z Li, and Matti PietikäInen. 2011. Facial expression recognition from near-infrared videos. *Image and Vision Computing* 29, 9 (2011), 607–619. 4.2, 2

[235] Long Zhao, Xi Peng, Yu Tian, Mubbasir Kapadia, and Dimitris Metaxas. 2018. Learning to forecast and refine residual motion for image-to-video generation. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 4.2, 2

[236] H. Zhou, Y. Liu, Z. Liu, P. Luo, and X. Wang. 2019. Talking face generation by adversarially disentangled audio-visual representation, In AAAI Conference on Artificial Intelligence (AAAI). *33rd AAAI Conference on Artificial Intelligence, AAAI 2019, 31st Innovative Applications of Artificial Intelligence Conference, IAAI 2019 and the 9th AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019* (2019), 9299–9306. 6, 5

[237] Xiaolian Zhu, Shan Yang, Geng Yang, and Lei Xie. 2019. Controlling emotion strength with relative attribute for end-to-end speech synthesis. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. 192–199. https://doi.org/10.1109/ASRU46091.2019.9003829 7.2, 6