Multimodal analysis of the predictability of hand-gesture properties

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

Rajmund Nagy
KTH Royal Institute of Technology
Stockholm, Sweden
rajmundn@kth.se

Michael Neff
University of California
Davis, United States
mpneff@ucdavis.edu

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

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

ABSTRACT

Embodied conversational agents benefit from being able to accompany their speech with gestures. Although many data-driven approaches to gesture generation have been proposed in recent years, it is still unclear whether such systems can consistently generate gestures that convey meaning. We investigate which gesture properties (phase, category, and semantics) can be predicted from speech text and/or audio using contemporary deep learning. In extensive experiments, we show that gesture properties related to gesture meaning (semantics and category) are predictable from text features (time-aligned BERT embeddings) alone, but not from prosodic audio features, while rhythm-related gesture properties (phase) on the other hand can be predicted from either audio, text (with word-level timing information), or both. These results are encouraging as they indicate that it is possible to equip an embodied agent with content-wise meaningful co-speech gestures using a machine-learning model.

KEYWORDS

Non-verbal behavior, animation, gesture generation, virtual agents, iconic gestures

1 INTRODUCTION

Verbal and nonverbal communication are important and complementary components of embodied human communication. In human communication, speech is typically accompanied by *co-speech gestures* or *gesticulation*, performed by the hands, head, and occasionally the body. Automatically generating such co-speech gestures is an important task in character animation and human-agent interaction, because a substantial fraction of our communication takes place through co-speech gestures [22, 38]. Furthermore, gesticulation has also been shown to enhance interactions with embodied agents [4, 36], e.g., to help with learning tasks [4], and to lead to a higher sense of co-presence [51].

While early hand gesture-generation systems mainly relied on rule-based approaches [6, 24, 37, 40], data-driven gesture generation has become an important research area in recent years [1, 13, 26, 53, 54]. Both paradigms have advantages and disadvantages. Rule-based systems produce gestures with clear communicative function, but lack diversity and require much manual effort to design. Data-driven systems, on the other hand, need less manual work and are more flexible, since they can generalise and generate new gestures on the fly. They may also scale better to large datasets. However, despite several attempts [1, 26, 53], there have in our view been no convincing demonstrations of recent data-driven approaches consistently generating gestures with a clear semantic relation to the speech content. For example, in terms of subjective gesture appropriateness for the speech, no system in the 2020 GENEA gesture-generation challenge [27] surpassed a bottom line that simply paired the input speech audio with mismatched excerpts of training data motion, completely unrelated to the speech.

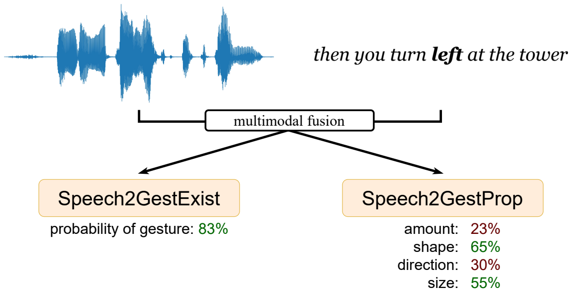

Figure 1: Overview of the problem we study.

It would be desirable to develop approaches that combine the strengths of both paradigms, enabling systems to be built from data yet produce gestures that fulfil a communicative function together with the speech. This has led us to investigate whether the communicative attributes of gesture can be modelled directly using recent data-driven methods.

The goal of this paper is to analyse to what extent modern deep-learning approaches are able to predict important communicative properties of hand gestures from the co-occurring speech. As such, this work should *not* be read as a machine-learning paper (our focus is not to propose new architectures or advance the numerical performance on some pre-existing benchmark), nor as a gesture-generation paper, since no gesture synthesis is performed. Instead, this is intended as a work on gesture analysis that studies the predictability of important gesture properties. Apart from being an interesting question in its own right, developing the ability to predict semantic aspects of gesticulation is a key element in driving future gesture-generation systems [28] to produce more meaningful and appropriate gesticulation. This work can therefore be seen as a continuation of recent efforts [13, 47, 57] towards imbuing data-driven systems with greater control over communicative function.

The specific contributions of our work are:

- We conduct extensive gesture-property prediction experiments on a direction-giving dataset with a high fraction of representational gestures, for which gesture properties have been extensively hand-annotated. Specifically, we predict 13 distinct property labels (8 relating to communicative function), which is significantly more than any prior work.
- We analyse which modalities of speech – audio and/or text – that are useful for predicting which gesture properties.
- We investigate how individual or general different gesture properties are, by looking at gesture-property prediction for both known and previously unseen speakers.

Despite the highly individual and stochastic nature of gestures, we find that numerous gesture properties can be predicted from speech, both for speakers inside and outside the training data. We also find speech text and audio to differ in their uses, where time-aligned text enables predicting gesture category, semantics, and phase, while prosodic audio features only help predict gesture phase. More information, including data and code, will be released on our project page at: svito-zar.github.io/speech2properties2gestures/

2 RELATED WORK

Since this paper considers the predictability of different properties of human gesticulation from multimodal representations of speech, our review of related work covers two aspects: first the prediction of various gesture properties, and then the use and combination of speech modalities for gesture generation. In general, the predictability of gesture properties has not been extensively studied, and most current gesture-generation systems do not integrate explicit gesture-property prediction, but there is nonetheless some prior work on predicting various gesture properties from speech.

2.1 Gesture presence/absence prediction

Ferstl et al. [14] used a statistical method based on speech prosody peaks to predict where a gesture should be placed. They set the timing so that gesture strokes were 55% complete at the pitch peak. Yunus et al. [56] predicted gesture presence and timing based on speech audio using a recurrent neural network (RNN). We also explore using a neural network for this, but we use a convolutional neural net instead of an RNN.

2.2 Gesture lexeme prediction

Many gesture synthesis approaches predict gesture lexemes, or tags, that encapsulate both gesture form and semantics. For example, a *cup* or *conduit* gesture involves a curved handshape, with the palm facing up and a forward motion of the hand from the speaker outward (gesture form) and is used to indicate an offering or conveyance (semantics). Systems of this type include [7, 9, 20, 30, 31, 37]. Some of these were rule-based, and predicted gesture semantics from input text based on a set of rules [7, 30, 31]. Other research applied statistical methods to learn probabilistic mappings from semantic concepts to gestures [16, 23]. Later, deep learning was applied to predict a fixed set of semantic gestures based on audio, text, and part-of-speech tags [9]. Our work, in contrast, does not consider a codified set of lexemes and instead predicts gesture properties that captures different elements of semantics, such as gesture categories and semantic gesture features.

2.3 Gesture kinematics prediction

Ferstl et al. [13] considered predicting kinematic gesture properties (specifically velocity, initial acceleration, gesture size, arm swivel, and hand opening) from speech. They trained multiple recurrent neural networks to predict these gesture parameters from the speech audio signal, and found that some parameters, such as path length, were predicted more accurately than others, for example velocity. Instead of kinematics, we consider the predictability of gesture properties related to gesture semantics and phase.

2.4 Gesture phase and category prediction

Kendon [21] defined the following gesture phases: preparation, hold, stroke, and retraction. All phases are optional except for the stroke, which is the expressive phase of the gesture. It has been shown that gesture stroke is strongly correlated with pitch accentuation in speech [12, 18]. Furthermore, McNeill [38] defined different gesture categories, or dimensions, such as deictic, iconic, and metaphoric (all related to the spoken message) gestures and beat gestures (which are more strongly related to speech prosody and rhythm).

This paper investigates how well gesture phases (as defined by Kendon), gesture semantic meaning, and gesture categories (as defined by McNeill) can be predicted from speech audio and text in a data-driven manner. The most similar prior work is due to Yunus et al. [56, 57], where a restricted set of gesture phase and category were predicted based on acoustic features only. Our study differs in that we consider additional gesture properties and also study the effect of different speech modalities as input.

2.5 Effect of the speech input modality

Many data-driven systems have only considered a single speech modality – either audio recordings or text transcriptions thereof – as input to the gesture generation, e.g., [3, 25, 39, 54]. However, the field is now shifting to use both audio and text together [1, 9, 26, 53]. This is, among other things, based on recent ablation studies of end-to-end gesture-synthesis systems in [26, 53], that compared gesture generation models which used only one modality against models using both. These studies found that using both speech modalities (audio and text) improved the synthesised gestures. This paper delves further into the effects of the different input modalities, and addresses the question of which speech modalities that are useful for predicting particular properties of human gesticulation.

3 DATA

3.1 Corpus

There are two principal ways to obtain data for 3D gesture synthesis: optical motion capture [19, 29] and 3D pose estimation from videos [1, 53]. Among these datasets, almost all are monologues, with only [19] involving more than one person in interaction.

Our present work is aimed at modelling iconic gestures, which are rare in all the previously cited datasets. Despite their important role in enabling meaningful gesticulation, these gestures only occur occasionally during social conversations. Hence we decided to focus on a dataset that contains a large proportion of iconic gestures, the Bielefeld Speech and Gesture Alignment corpus (SaGA) [34]. This is the largest database we are aware of with detailed and accurate gesture-property annotations. Larger gesture databases exist, e.g., [1], but do not have the annotations necessary for our research. We believe the SaGA dataset is sufficiently large for our purposes, since it has been previously used for generating iconic gestures [3].

The SaGA dataset contains a total of 280 minutes of recordings of 25 different participants speaking and gesturing to an interlocutor. All recordings are in German. A key goal of SaGA was to capture a large number of iconic gestures. This was accomplished through a specific data collection procedure in which participants first saw a virtual reality bus tour and then described the route, and the prominent visual landmarks placed along that route, to another person. Both the navigation task and the landmarks provided natural visual grounding upon which iconic gestures are based. All participants followed the same route, thus maximising the degree of consistency between the recordings and simplifying the task of grounding gesture prediction in language by considering a tightly restricted semantic domain. Audio and video were recorded of each interaction [34] and every gesture was manually annotated according to a detailed labelling scheme. We use a subset of their annotation categories for our study, as described in Section 3.3.

Dataset partitioning. For our research, we ignored time-frames where the interlocutor was speaking (according to the annotation). This left 62,591 frames in total (at 5 fps), out of which 33,454 frames were annotated as containing a gesture. We held out three recordings (numbers 1, 13, and 16) for testing and used the remaining 22 recordings for training and cross-validation. We did not use any of these test recordings in our experiments and left them as a test set for future research, so that future models can be evaluated without data leakage from the experiments and findings in this paper.

We used two different data partitionings for cross-validation, to avoid tuning hyperparameters and evaluating on the exact same data splits. For choosing hyper-parameters, we do classical 10-fold cross-validation but only consider the average results across the first seven folds for hyperparameter selection. For evaluating the model, we do 20-fold cross-validation, set up such that every fold contains 5% of the data from each of the 22 subjects in the recordings we consider. Training and validation sequences never overlapped.

3.2 Speech modalities and their encoding

We used two different speech modalities from the dataset, each of which is described below.

Text. Each recording was transcribed in German. Transcriptions contain both the written form of every word spoken along with the timing (onset and offset) of the word. Text features were extracted by applying German DistilBERT [46], a simplified and compressed version of BERT [11] as implemented by HuggingFace [50], to the full sequence of words spoken by each the participant. The transcription does not contain punctuation or other sentence delimiters due to the spontaneous and continuous nature of the speech. The DistilBERT tokeniser produces one 768-dimensional feature vector (a.k.a. "embedding") per word-piece token. These were converted to a single feature vector per word by computing the arithmetic average of the feature vectors of all word pieces within that word. When predicting gesture properties, each vector was supplemented with one extra number about word timing, namely the time-difference from the word onset and the prediction target frame (negative for words starting before the target point and positive for future words). Text-based gesture-generation commonly uses timing information [16, 54] even though it cannot be derived from text alone.

Audio. We extracted the audio tracks from each video and converted them to mono waveforms with a 48 kHz sampling rate. We then used Parselmouth [17] to compute four prosodic features (fundamental frequency, energy, and their derivatives computed with finite differences) as the audio feature set of our experiments. These prosodic features are commonly used in speech emotion analysis as well as for gesture property prediction, e.g., [56]. We normalised pitch and intensity like in [8, 25]: the pitch values were adjusted by taking $log(x + 1) - 4$ and setting negative values to zero, and the intensity values were adjusted by taking $log(x) - 3$. The audio features were first extracted at 200 fps and then resampled to 5 fps by averaging, to match the resolution of the gesture annotations.

We also experimented with using spectrograms instead of prosodic features but found no difference between the two, except that prosody is more anonymous, enabling us to release audio features.

3.3 Gesture properties and their encoding

The SaGA corpus contains detailed annotations of the properties of the gestures in the recordings. We made use of the following gesture properties in our experiments: *R.G.Left Semantic*, *R.G.Right Semantic*, *R.G.Left Phrase*, *R.G.Right Phrase*, *R.G.Left.Phase*, *R.G.Right.Phase*. The *Semantic* property indicates which semantic information is contained in the gesture. *Phrase* indicates gesture category. *Phases* are sub-units of gestures that indicate if the hands are preparing to gesture, meaning is currently being conveyed, the hands are retracting, etc. For details about the data collection and annotation scheme we refer the reader to Lücking et al. [35] and Bergmann and Kopp [2]. To simplify modelling, we merged the features for the left and right hand into a single feature using a per-frame logical OR. Each feature was encoded into a vector of binary values, which is one-hot for *Phase* since phases are mutually exclusive.

Note that the work in this paper does not make use of the videos captured during the SaGA corpus recordings, only transcriptions, gesture annotations, and anonymous audio features (prosody) derived from those recordings.

Gesture-property representations. We encoded gesture properties at a rate of five frames per second (5 fps). As described in Section 4.2, our system first predicts if a gesture is needed and then what kind of gesture it should be. For the latter gesture-property prediction task, we only consider time-frames where a gesture was present in the data, i.e., frames where any of the annotations we considered were present and nonzero. This amounted to 33k out of 62k total frames. We list the gesture-property labels we considered and the number of frames they were present at in Figure 2. As can be seen, most of the gesture-property labels only apply to a small fraction of the gesture-containing frames in the data.

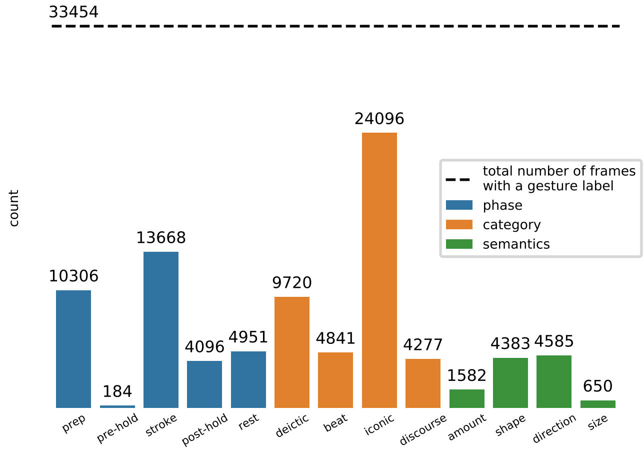

Figure 2: The frequency of each gesture-property label in the SaGA dataset. Note that frequencies may sum to more than 33,454 since most categories are not mutually exclusive.

We encoded the gesture properties as binary vectors. For this, we first created an ordering of the different labels relevant to each property. For example, for *Gesture Category* we ordered the different possible labels as follows {1: ‘deictic’, 2: ‘beat’, 3: ‘iconic’, 4: ‘discourse’}. A frame with Category annotation “beat-iconic” would then be encoded by the vector $[0, 1, 1, 0]^T$. As the example shows, gesture categories are not mutually exclusive, and several labels can be present simultaneously. The same applies to gesture semantics labels. Gesture phase, on the contrary, is exclusive – only one label can be applicable at a time – and we take this mutual exclusivity of gesture phases into account during modelling and evaluation.

4 EXPERIMENTAL SETUP

This section describes the experimental setup for our experiments on predicting gesture properties from speech text and audio.

4.1 Problem formulation

We frame the problem of gesture-property prediction as follows: given a sequence of speech features $\boldsymbol{s} = [s_t]_{t=1:T}$ the task is to generate a sequence of corresponding binary gesture properties $\hat{\boldsymbol{p}} = [\hat{p}_t]_{t=1:T}$. Here, $t = 1 : T$ denotes indexing into a sequence of vectors for integer t in 1 to T. Each speech segment s_t is represented by several different features, specifically acoustic features (e.g., prosody), semantic features (e.g., word embeddings), or a combination of the two.

4.2 Gesture-property prediction model

Our gesture-property prediction model (Figure 1) consists of two components that take speech audio and text as input: *Speech2Gest-Exist*, which predicts the probability of making a gesture, and *Speech2GestProp*, which predicts the probabilities of different labels for a given gesture property. Such hierarchical models have been successful on other sequence-prediction tasks such as text-to-speech intonation generation, where first predicting the presence or absence of voicing, and then predicting voicing frequency, worked better than predicting the two aspects jointly at once [49].

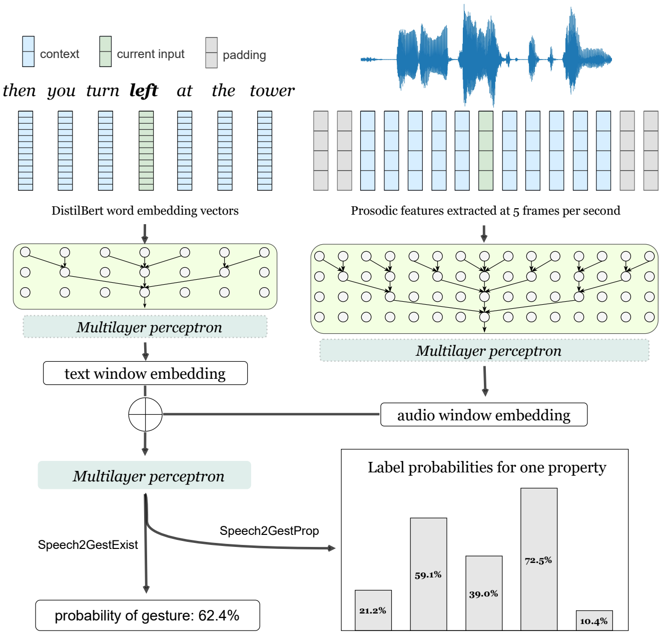

Figure 3: The shared multimodal architecture of our two networks. First, the two modalities are independently encoded using dilated temporal CNNs, using zero-padding as necessary. Then, the two encodings are concatenated and fed into an MLP decoder, which returns the final output.

Detailed model specification. Dilated CNNs [55] are a widely-used neural-network architecture for sequence modelling, used in WaveNet [48] and WaveGlow [44] and recently also adapted to human motion modelling [15]. Following the developments in [44, 48], we evaluated the effect of residual and skip connections, different activation functions and weight initialisation schemes. We also experimented with no dilation and with a simple MLP [45]. The experimental results are publicly available on FigShare: doi.org/10.6084/m9.figshare.15134295. Seeing that neither of these modifications provided any benefits, we use the dilated convolution architecture depicted in Figure 3 for all other experiments.

The model inputs are sequences of audio frames and transcribed spoken words in a sliding window centred on the current time frame. Based on findings regarding the temporal synchrony between speech and gesture [33, 43], we consider the current, three past, and three future word-token feature vectors[1] and the current and five past and five future audio frames (i.e., 1 s to either side). By sliding these windows over the input speech-feature sequences, we can make predictions for the selected gesture properties frame by frame for all times t in the sequence. (For this paper, we only considered frames sufficiently far from sequence edges for all model inputs to be well defined, to avoid edge effects.) This setup makes

[1] In our experiments, these features were timing information and DistilBERT contextual word embeddings, which depend on the entire sentence each word belongs to, so our system predictions can be informed also by words outside this 7-token window.

use of future speech, which is standard in gesture generation and rarely considered a limitation since most applications do not depend on live speech. For example, the utterance-based TTS systems used by many social robots and virtual agents require the entire utterance text to be available before audio synthesis can begin.

As illustrated in Figure 3, speech audio and text are first encoded into the intermediate *text window embedding* and *audio window embedding* representations using two separate neural networks, each of them containing several layers of dilated convolution. The two embeddings are then concatenated and passed into a simple fully-connected neural network (MLP). At the final layer, we map the values onto the unit interval $[0, 1]$, since the output should indicate the probability that each relevant gesture property is present. For that, a sigmoid output nonlinearity is applied to *Gesture Category* and *Gesture Semantics* and a softmax output nonlinearity is applied to the *Gesture Phase* outputs. The softmax is used since different phase labels, unlike the other property categories, are mutually exclusive. From a probabilistic perspective, the use of a sigmoid for each binary property corresponds to the assumption that each property is statistically independent of the others, given the input features. This is a common modelling assumption for binary variables that are not mutually exclusive.

Since any given gesture property is present in just a fraction of the time frames, any data used to train for gesture-property predictors will be highly imbalanced. To mitigate this, we experimented with upsampling underrepresented classes to balance the data and also considered several different loss functions: not only the standard *cross-entropy loss*, $CE(p_t) = -log(p_t)$ (where p_t stands for the model probability of the correct class at time t), but also the *focal loss* [32], which was developed to address the rarity of positive labels in common datasets, and a class-balancing version of the focal loss from [10]. Each loss function is aggregated for sequences and minibatches by summing over constituent frames.

Hyperparameters. For each experiment and each model in Section 5, we conducted a separate hyperparameter search using random search [5]. Each random search consisted of 100 runs. For each run, we randomly sampled all the key hyperparameters over a predefined range for each value and trained the model for a fixed number of epochs dependent on the task. Specifically, we varied: hidden dimensionality, number of layers, kernel size, dropout, and output embedding dimensionality for each encoder; hidden dimensionality, number of layers, and dropout for the decoder; learning rate, batch size, and other optimisation parameters.

We selected the best hyperparameters based on the average Macro F_1 [52] score over seven folds during 10-fold cross-validation, and used these settings to compute the results reported in Section 5. Hyperparameters for all models in the paper are publicly available on FigShare: doi.org/10.6084/m9.figshare.15134076.

4.3 Baseline systems

For the majority of the properties we predict, no previous baseline systems or benchmark performance exist. Instead, our main starting point for baselining is the finding from [27] that no gesture-generation system beat a mismatched bottom line that paired speech with unrelated training-data motion. Inspired by this, we create and compare against a number of simple bottom-line systems that similarly have no dependence on the input speech. These include two constant-output systems (*AlwaysZero* and *AlwaysOne*), and two systems based on random output, either uniformly random (system *UniformRandom*) or random draws with the same distribution as the a-priori class abundances in the training data (system *InformedRandom*). Any system can be said to be *better than chance* if it surpasses all four of these bottom lines. Moreover, any time that happens, we say that the corresponding property is *predictable* from the given input features. (This is very different from being perfectly predictable, which arguably is an unrealistic goal for problems that involve human behaviour.) Even a minor improvement over chance predictability could add important communicative value to generated gestures, since current state-of-the-art gesture generation is no more appropriate than random gestures [27].

4.4 Evaluation metrics

It is well known that standard classification accuracy (one minus the error rate) does not capture overall system performance well when the data is highly unbalanced, since it may then be possible to achieve high accuracy by always predicting the majority class, regardless of the input features of the given instance. Instead, we use the F_1 score as our main performance indicator. This measure is the harmonic mean of precision and recall, and is a popular evaluation measure for classification of unbalanced classes. More specifically, we use the Macro F_1 score [42], which is simply the arithmetic average of F_1 scores for all possible, mutually exclusive classes c: Macro $F_1 = \frac{1}{C}\sum_{c=1}^{C} F_1(c)$.

Note that since phase labels are mutually exclusive, while other gesture property labels are not, phase is evaluated differently. For gesture categories and semantics we calculate separate Macro F_1 score for each label, since they are not mutually exclusive and are treated as independent. For the gesture phase, on the contrary, we evaluate only F_1 scores for each label, and not the Macro F_1 score, which averages over all possible labels.

To get a better impression of generalisation ability on our limited dataset, we used cross-validation. For each of our experiments, we report the mean and standard deviation of our performance measure across 20 cross-validation folds. These folds were set up such that every fold contained 5% of the data from each of the 22 people in the recordings we considered. This means that the cross-validation quantifies *within-person generalisation performance*, although we also looked at *across-person generalisation* by holding out one individual at a time (see Section 5.5).

5 RESULTS AND DISCUSSION

We conducted several experiments, first comparing different performance metrics, and then evaluating 1) how well we can predict gesture presence, 2) which modalities are essential for predicting which gesture properties, 3) how well predictions generalise to new speakers, and more. In this section, we report and discuss the results of these experiments.

In each experiment, we vary one aspect while keeping everything else the same. Our default settings are:

- using both speech modalities, instead of only audio or text;
- evaluating generalisation within known speakers, instead of generalisation to new speakers;

	Accuracy	Precision	Recall	F_1 for 1	F_1 for 0	Macro F_1
AlwaysZero	**85%** ± **10%**	0% ± 0%	0% ± 0%	0% ± 0%	**92%** ± **6%**	46% ± 3%
AlwaysOne	15% ± 10%	15% ± 10%	**100%** ± **0%**	25% ± 14%	0% ± 0%	13% ± 7%
UniformRandom	50% ± 3%	15% ± 10%	51% ± 4%	21% ± 11%	62% ± 6%	42% ± 4%
InformedRandom	76% ± 8%	17% ± 11%	14% ± 3%	14% ± 5%	**86%** ± **2%**	50% ± 3%
our result	**85%** ± **9%**	**51%** ± **18%**	39% ± 17%	**41%** ± **14%**	**89%** ± **4%**	**65%** ± **6%**

Table 1: A comparison between various evaluation metrics for gesture phase prediction for the gesture semantic property "shape". The baselines are *italicised*; "our result" refers to our multimodal dilated CNN (BothModalities).

- training individual models for each gesture property, instead of training a single model of all properties simultaneously.

5.1 Comparison of evaluation metrics

In order to put the evaluation metric used into context, Table 1 reports the accuracy, precision, recall, F_1 and Macro F_1 scores for predicting the presence/absence of (as an example) the gesture semantic property label "shape".

Overall, Macro F_1 is the most preferable evaluation metric. Accuracy is misleading because it can be very high for primitive baselines (such as AlwaysZero) simply because one class is dominant over the other. Using only precision or recall is not sufficient, as each focuses only on either false negatives or false positives. As the F_1 score for label presence is the harmonic mean between precision and recall, it tends to be closer to the lower of the two values (see, e.g., UniformRandom). However, the F_1 score is not symmetric and strongly focuses on true positives. The Macro F_1 score is computed as a simple arithmetic average between the F_1 scores for label presence and for label absence. This metric has the added advantage that chance performance is 50% even for imbalanced data, and it is the metric we choose to report in the rest of this section. The maximum achievable score, on the other hand, is likely significantly below 100%, since human gesticulation is highly stochastic.

Interestingly, our model achieves much higher precision than recall and a much lower F_1 score for one than for zero. The accuracy and recall are more similar to AlwaysZero than AlwaysOne, suggesting that the model may be overpredicting the majority class, as is typical of cross-entropy-based classifiers on imbalanced data.

5.2 Predicting gesture presence and timing

The first question we considered was whether or not it is possible to predict when to make a gesture (i.e., predict the presence or absence of a gesture from the speech features in our dataset). The best model found by our hyperparameter search achieved a 70% ± 3.7% Macro F_1 score on this binary classification task. This is better than chance (see Figure 4) and agrees with results from previous work on another dataset in a different language (English) [56].

5.3 Experiments on machine-learning setup

We experimented with two different approaches for dealing with data imbalance, namely upsampling and special loss functions, as described in Section 4.2. Our experiments found no benefits to either upsampling or the special loss functions in terms of Macro F_1 score: the results were slightly better for some features and slightly worse

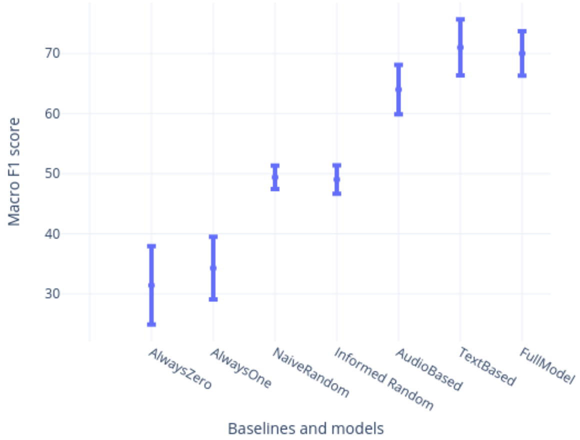

Figure 4: Macro F_1 score for gesture presence prediction. It can be seen that gesture presence is predictable regardless of the input modality used.

for the others, but there was no major difference. We therefore use the conventional cross-entropy loss without any upsampling for all other experiments in this paper.

We also experimented with training a single model to predict all the gesture properties at once, versus training individual models for each gesture property. We found that using individual models performed better, possibly due to different tasks benefiting from different hyperparameter choices, hence we model each gesture property individually in the rest of our experiments.

5.4 Evaluating text and audio contributions

This experiment analysed the importance of the two input speech modalities – text and audio (prosodic features) – for predicting the gesture properties under study. As can be seen from the results in Table 2, the text features were informative and predict gesture category and gesture semantic content better than chance. Audio features, in contrast, did not improve on the best bottom line predictions, and so did not allow better-than-chance prediction on their own. This provides evidence that audio alone is insufficient for predicting the semantics of iconic gestures. It also indicates that even the category of gesture cannot be predicted from audio alone. Moreover, the combination of audio and text did not in general perform better than text on its own. Text appears to be a necessary input for both gesture category and semantics. This has strong implications for the need to include text in machine learning gesture models in general. This finding can be explained by semantic information, which gesture category strongly depends on, being challenging to obtain directly from the audio. Whether or not an iconic gesture is appropriate will generally depend on the semantic information in the speech.

A different story emerges for gesture phase prediction, where both audio and text were helpful. In particular, gesture stroke could be predicted noticeably better than informed random sampling using either modality, with the trend being especially clear for audio. The utility of text may relate to the relationship between the

	gesture category [Macro F_1]				**gesture semantics** [Macro F_1]				**gesture phase** [F_1]				
label	deictic	beat	iconic	discourse	amount	shape	direction	size	pre-hold	post-hold	stroke	retr	prep
relative frequency	29.05%	14.47%	72.03%	12.78%	4.7%	13.1%	13.7%	1.9%	0.6%	12.2%	40.9%	14.8%	30.8%
AlwaysOne	20% ± 6%	13% ± 5%	43% ± 2%	11% ± 3%	4% ± 3%	13% ± 7%	10% ± 4%	3% ± 2%	–	–	–	–	–
AlwaysZero	43% ± 3%	46% ± 2%	20% ± 5%	46% ± 11%	49% ± 7%	46% ± 3%	47% ± 1.5%	49% ± 0.6%	–	–	–	–	–
UniformRandom	47% ± 5%	42% ± 3%	46% ± 2%	42% ± 2%	37% ± 2%	42% ± 4%	41% ± 3%	36% ± 2%	1.5% ± 2%	14% ± 4%	23% ± 3%	14% ± 5%	20% ± 3%
InformedRandom	50% ± 2%	50% ± 2%	50% ± 1.5%	50% ± 2%	49% ± 1%	49% ± 2%	49% ± 2%	50% ± 1%	1.3% ± 4%	12% ± 4%	42% ± 4%	14% ± 5%	30% ± 3%
AudioOnly	50% ± 3%	46% ± 2%	53% ± 5%	50% ± 2%	49% ± 2%	49% ± 4%	50% ± 2%	49% ± 1%	0%	5% ± 3%	**53% ± 8%**	12% ± 5%	**40% ± 4%**
TextWithTiming	**60% ± 6%**	54% ± 3%	**64% ± 7%**	**58% ± 8%**	**62% ± 11%**	**65% ± 8%**	**60% ± 7%**	57% ± 9%	1% ± 2%	21% ± 12%	**51% ± 12%**	25% ± 9%	**42% ± 7%**
TextNoTiming	**61% ± 6%**	56% ± 5%	**63% ± 7%**	**60% ± 8%**	**68% ± 8%**	**66% ± 7%**	**61% ± 9%**	59% ± 12%	0% ± 0%	18% ± 11%	**54% ± 11%**	19% ± 7%	**39% ± 7%**
BothModalities	**60% ± 6%**	53% ± 6%	**63% ± 5%**	**59% ± 7%**	**63% ± 8%**	**65% ± 6%**	**62% ± 8%**	**59% ± 9%**	0.5% ± 1.3%	23% ± 12%	**47% ± 10%**	25% ± 5%	**45% ± 6%**

Table 2: Gesture-property prediction scores for all baselines, and our trained predictors using text, audio, or both modalities. Baselines are *italicised*; bold, coloured numbers indicate that the given label is found to be predictable as defined in Sec. 4.3.

verbal message and prosodic prominence – prosody is predictable from text, as TTS systems show (although their input is graphemic symbols rather than semantic word vectors). Our default text features (used by *TextWithTiming* and *BothModalities*) also include word timing information relative to the current frame, but this does not seem essential since a system without these numbers (*TextNoTiming*) performed equally well. Overall, these results confirm previous findings that gesture stroke can be predicted from speech, but sheds new light on which modality is needed for such prediction: while previously speech audio was mainly used, time-aligned speech text could also be helpful on its own.

5.5 Generalising within and across speakers

Next, we evaluated gesture property prediction performance when generalising to novel speakers, versus the performance on speakers present in the training data.

Table 3 compares prediction results of a hold-one-speaker-out cross-validation strategy (*BetweenSpeaker*) to our default cross-validation strategy where speakers are present in both training and test sets (*WithinSpeaker*), and when speakers are present in both sets and the model has access to speaker IDs encoded as one-hot vectors (*WithinSpeaker$_{ID}$*).

We can observe that performance drops a little when we evaluate on completely novel speakers, but the drop is not substantial and the performance stays better than chance. Given that gesture behaviour is highly idiosyncratic across individuals, one might a priori have expected a large drop in prediction performance. It is reassuring that this drop is quite modest and suggests that gesture property predictors generalise relatively well to new speakers.

Conversely, we observe no notable difference between *WithinSpeaker* and *WithinSpeaker$_{ID}$*, indicating that predictions did not benefit from knowing the speaker label. One reason for this could be that each speaker only has 10 min of data, which may not allow learning personalised correlations. As above, another contributing factor could be that gesture behaviour is very stochastic overall.

We have also investigated the generalisation gap between training and validation performance. As can be seen in our results on FigShare, doi.org/10.6084/m9.figshare.15134796, the model overfits by double digits for gesture semantics and category, but not for gesture phase. It may be that gesture-phase prediction is ambiguous and difficult even on training data with the features we used.

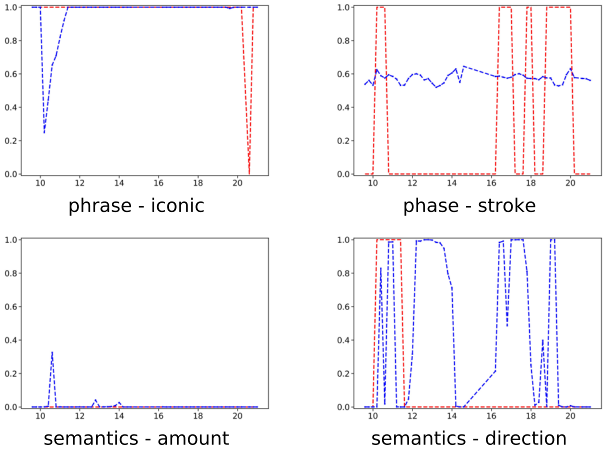

Figure 5: Example output sequences from gesture label prediction. The x-axes are time (in frames) and the y-axes label probability. True labels are red, predictions blue.

5.6 Some prediction examples

Figure 5 shows several example sequences of per-frame predicted probabilities for the presence of various gesture property labels on a held-out speech sequence. We can see both good and bad performance. In general, predicted sequences are somewhat jagged, suggesting that future predictors could benefit from temporal smoothing or a more explicit model of labels over time. We note that the predictions for gesture semantics and category are surprisingly confident, given that cross-entropy tends to promote cautious models that favour predicting numbers close to the a-priori class probabilities in the absence of compelling evidence to the contrary.

5.7 Do all speakers have similar predictability?

There is a large difference in prediction performance for different speakers, as seen in the high standard deviation in most of the tables above. We show one example of the performance variation

	gesture category [Macro F_1]				**gesture semantics** [Macro F_1]				**gesture phase** [F_1]				
label	deictic	beat	iconic	discourse	amount	shape	direction	size	pre-hold	post-hold	stroke	retr	prep
BetweenSpeaker	57% ± 5%	49% ± 3%	55% ± 5%	56% ± 6%	60% ± 7%	63% ± 8%	62% ± 7%	59% ± 12%	1% ± 1.5%	14% ± 8%	35% ± 8%	25% ± 8%	43% ± 6%
WithinSpeaker	60% ± 6%	53% ± 6%	63% ± 5%	59% ± 7%	63% ± 8%	65% ± 6%	62% ± 8%	59% ± 9%	0.5% ± 1.3%	23% ± 12%	47% ± 10%	25% ± 5%	45% ± 6%
$\text{WithinSpeaker}_{\text{ID}}$	60% ± 6%	55% ± 5%	63% ± 7%	60% ± 9%	66% ± 9%	66% ± 7%	62% ± 8%	59% ± 10%	0% ± 0%	21% ± 10%	55% ± 9%	24% ± 9%	42% ± 5%

Table 3: A comparison of prediction results for different cross-validation strategies. For detailed information see Section 5.5.

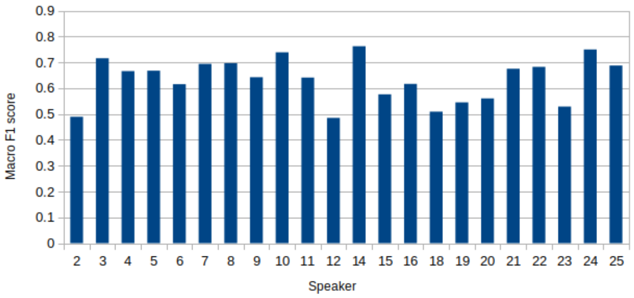

Figure 6: Macro F_1 score for predicting the "shape" label across the 22 speakers considered. Note the high variation.

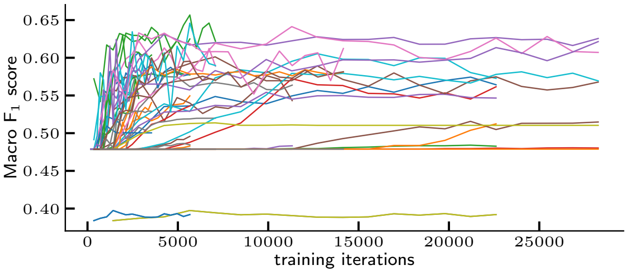

Figure 7: Training evolution of semantics prediction average Macro F_1 score for 100 different hyperparameter settings. The total number of epochs was fixed, but the number of iterations differ since the batch size was changed as part of the hyperparameter search.

between speakers in Figure 6. We see that how well we predict the gesture semantic label "shape" changes substantially from speaker to speaker, indicating that not all speakers are equally predictable, although predictions are better than chance in nearly all cases.

5.8 On the effect of hyperparameters

Aside from the large variation between different speakers, we also observe great performance variation depending on model hyperparameters. Figure 7 shows the Macro F_1 score for predicting gesture semantics for 100 different hyperparameter runs. We can see that results vary greatly depending on hyperparameters. The variation depending on the hyperparameters is much more notable than the difference between many conditions in our experiments, indicating that the model used is sensitive to the hyperparameter settings.

6 CONCLUSIONS

We have studied the extent to which 13 different gesture-property labels – mainly ones of relevance to communicative gestures – can be predicted from speech. Numerous experiments on a direction-giving dataset show that the gesture properties we considered, such as gesture categories and gesture phases, can be predicted from speech with Macro F_1 scores better than chance. Predicting gesture properties for speakers outside the training data was only slightly more challenging, suggesting that gesture-property prediction may generalise well.

Another central finding is that, for predicting gesture properties such as gesture category and gesture semantics, all that must be known is the time-aligned text transcript, while for others, specifically phase, prosodic audio features can be used in place of text.

Our ≈10% advantage over the chance baselines should be viewed in the context that human gestures are highly stochastic, and that state-of-the-art data-driven gesture synthesis does not compare favourably to random gesticulation [27]. Leveraging our gesture-property predictions to achieve semantically appropriate gestures even a fraction of the time could thus add important communicative value, and the prediction models we identify are well-suited for integration into modern data-driven gesture generation systems.

6.1 Future work

The present study opens up several directions for future research:

First, there are many alternative design choices left to explore, e.g., more detailed audio features, other word embeddings, different data-processing frame rates, and evaluating on a metric that (like human perception [41]) is less sensitive to small timing shifts.

Second, it would be interesting to perform a similar study on other datasets, e.g., in different languages or from situations other than direction giving. Also, while the SaGA dataset we used is the largest one we know that has been annotated at this high level of detail, larger datasets are also of interest as they become available.

Last, we suggest integrating gesture-property prediction into modern gesture-generation models, following [28], with the goal of enabling more appropriate and meaningful gesture synthesis.

ACKNOWLEDGMENTS

The authors are grateful to Stefan Kopp for providing the SaGA dataset and fruitful discussions about it, and to Olga Abramov for advising on the dataset gesture-property processing. This work was partially supported by the Swedish Foundation for Strategic Research Grant No. RIT15-0107 and by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation.

REFERENCES

[1] Chaitanya Ahuja, Dong Won Lee, Ryo Ishii, and Louis-Philippe Morency. 2020. No Gestures Left Behind: Learning Relationships between Spoken Language and Freeform Gestures. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*. 1884–1895.

[2] Kirsten Bergmann and Stefan Kopp. 2006. *Verbal or visual?: How information is distributed across speech and gesture in spatial dialog*. Universität Potsdam.

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

[4] Kirsten Bergmann and Manuela Macedonia. 2013. A virtual agent as vocabulary trainer: iconic gestures help to improve learners' memory performance. In *International Workshop on Intelligent Virtual Agents*. Springer, 139–148.

[5] James Bergstra and Yoshua Bengio. 2012. Random search for hyper-parameter optimization. *Journal of Machine Learning Research* 13, 2 (2012).

[6] Justine Cassell, Catherine Pelachaud, Norman Badler, Mark Steedman, Brett Achorn, Tripp Becket, Brett Douville, Scott Prevost, and Matthew Stone. 1994. Animated conversation: rule-based generation of facial expression, gesture & spoken intonation for multiple conversational agents. In *Proceedings of the 21st annual conference on Computer graphics and interactive techniques*. 413–420.

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

[8] 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, 127–140.

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

[10] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. 2019. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9268–9277.

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics* (2018).

[12] Núria Esteve-Gibert and Pilar Prieto. 2013. Prosodic structure shapes the temporal realization of intonation and manual gesture movements. (2013).

[13] Ylva Ferstl, Michael Neff, and Rachel McDonnell. 2020. Understanding the predictability of gesture parameters from speech and their perceptual importance. In *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents*. 1–8.

[14] Ylva Ferstl, Michael Neff, and Rachel McDonnell. 2021. ExpressGesture: Expressive gesture generation from speech through database matching. *Computer Animation and Virtual Worlds* (2021), e2016.

[15] Shuaiying Hou, Weiwei Xu, Jinxiang Chai, Congyi Wang, Wenlin Zhuang, Yu Chen, Hujun Bao, and Yangang Wang. 2021. A Causal Convolutional Neural Network for Motion Modeling and Synthesis. *arXiv preprint arXiv:2101.12276* (2021).

[16] 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.

[17] Yannick Jadoul, Bill Thompson, and Bart De Boer. 2018. Introducing parselmouth: A python interface to praat. *Journal of Phonetics* 71 (2018), 1–15.

[18] Stefanie Jannedy and Norma Mendoza-Denton. 2005. *Structuring information through gesture and intonation*. Universitätsbibliothek Johann Christian Senckenberg.

[19] Hanbyul Joo, Tomas Simon, Mina Cikara, and Yaser Sheikh. 2019. Towards social artificial intelligence: Nonverbal social signal prediction in a triadic interaction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 10873–10883.

[20] Sri Rama Kartheek Kappagantula, Nicoletta Adamo-Villani, Meng-Lin Wu, and Voicu Popescu. 2020. Automatic Deictic Gestures for Animated Pedagogical Agents. *IEEE Transactions on Learning Technologies* 13, 1 (2020), 1–13.

[21] Adam Kendon. 1980. Gesticulation and speech: Two aspects of the process of utterance. In *The relationship of verbal and nonverbal communication*. 207–228.

[22] Adam Kendon. 2004. *Gesture: Visible action as utterance*. Cambridge University Press.

[23] Michael Kipp. 2005. *Gesture generation by imitation: From human behavior to computer character animation*. Universal-Publishers.

[24] Stefan Kopp and Ipke Wachsmuth. 2004. Synthesizing multimodal utterances for conversational agents. *Computer animation and virtual worlds* 15, 1 (2004), 39–52.

[25] Taras Kucherenko, Dai Hasegawa, Naoshi Kaneko, Gustav Eje Henter, and Hedvig Kjellström. 2021. Moving fast and slow: Analysis of representations and post-processing in speech-driven automatic gesture generation. *International Journal of Human-Computer Interaction* 37, 14 (2021), 1300–1316.

[26] Taras Kucherenko, Patrik Jonell, Sanne van Waveren, Gustav Eje Henter, Simon Alexanderson, 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*.

[27] 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* (College Station, TX, USA) *(IUI '21)*. Association for Computing Machinery, New York, NY, USA, 11–21.

[28] Taras Kucherenko, Rajmund Nagy, Patrik Jonell, Michael Neff, Hedvig Kjellström, and Gustav Eje Henter. 2021. Speech2Properties2Gestures: Gesture-Property Prediction as a Tool for Generating Representational Gestures from Speech. In *Proceedings of the 21st ACM International Conference on Intelligent Virtual Agents*.

[29] 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*. 763–772.

[30] Jina Lee and Stacy Marsella. 2006. Nonverbal behavior generator for embodied conversational agents. In *International Workshop on Intelligent Virtual Agents*. Springer, 243–255.

[31] Margaux Lhommet and Stacy C Marsella. 2013. Gesture with meaning. In *International Workshop on Intelligent Virtual Agents*. Springer, 303–312.

[32] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*. 2980–2988.

[33] Daniel P. Loehr. 2012. Temporal, structural, and pragmatic synchrony between intonation and gesture. *Laboratory Phonology* 3, 1 (2012), 71–89.

[34] Andy Lücking, Kirsten Bergman, Florian Hahn, Stefan Kopp, and Hannes Rieser. 2013. Data-based analysis of speech and gesture: The Bielefeld Speech and Gesture Alignment Corpus (SaGA) and its applications. *Journal on Multimodal User Interfaces* 7, 1 (2013), 5–18.

[35] Andy Lücking, Kirsten Bergmann, Florian Hahn, Stefan Kopp, and Hannes Rieser. 2010. The Bielefeld speech and gesture alignment corpus (SaGA). In *LREC 2010 workshop: Multimodal corpora–advances in capturing, coding and analyzing multimodality*.

[36] Pengcheng Luo, Victor Ng-Thow-Hing, and Michael Neff. 2013. An examination of whether people prefer agents whose gestures mimic their own. In *International Workshop on Intelligent Virtual Agents*. Springer, 229–238.

[37] Stacy Marsella, Yuyu Xu, Margaux Lhommet, Andrew Feng, Stefan Scherer, and Ari Shapiro. 2013. Virtual character performance from speech. In *Proceedings of the 12th ACM SIGGRAPH/Eurographics Symposium on Computer Animation*. 25–35.

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

[39] Michael Neff, Michael Kipp, Irene Albrecht, and Hans-Peter Seidel. 2008. Gesture modeling and animation based on a probabilistic re-creation of speaker style. *ACM Transactions on Graphics* (2008).

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

[41] Jens Nirme, Magnus Haake, Agneta Gulz, and Marianne Gullberg. 2019. Motion capture-based animated characters for the study of speech–gesture integration. *Behavior research methods* (2019), 1–16.

[42] Juri Opitz and Sebastian Burst. 2019. Macro f1 and macro f1. *arXiv preprint arXiv:1911.03347* (2019).

[43] Wim Pouw and James A. Dixon. 2019. Quantifying gesture-speech synchrony. In *Gesture and Speech in Interaction Workshop*.

[44] Ryan Prenger, Rafael Valle, and Bryan Catanzaro. 2019. WaveGlow: A flow-based generative network for speech synthesis. In *Proc. ICASSP (ICASSP'19)*. IEEE Signal Processing Society, Piscataway, NJ, USA, 3617–3621.

[45] Frank Rosenblatt. 1958. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review* 65, 6 (1958), 386.

[46] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. In *Advances in Neural Information Processing Systems (NeurIPS)*.

[47] Carolyn Saund, Andrei Bîrlădeanu, and Stacy Marsella. 2021. CMCF: An Architecture for Realtime Gesture Generation by Clustering Gestures by Motion and Communicative Function. In *Proceedings of the 20th InternationalConference on Autonomous Agents and Multiagent Systems*.

[48] Aäron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. 2016. WaveNet: A generative model for raw audio. arXiv:1609.03499

[49] Xin Wang, Shinji Takaki, and Junichi Yamagishi. 2018. Autoregressive Neural F0 Model for Statistical Parametric Speech Synthesis. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 26, 8 (2018), 1406–1419.

[50] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu,

Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, Online, 38–45. https://www.aclweb.org/anthology/2020.emnlp-demos.6

[51] Yanxiang Wu, Sabarish V Babu, Rowan Armstrong, Jeffrey W Bertrand, Jun Luo, Tania Roy, Shaundra B Daily, Lauren Cairco Dukes, Larry F Hodges, and Tracy Fasolino. 2014. Effects of virtual human animation on emotion contagion in simulated inter-personal experiences. *IEEE Transactions on Visualization and Computer Graphics* 20, 4 (2014), 626–635.

[52] Yiming Yang and Xin Liu. 1999. A re-examination of text categorization methods. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*. 42–49.

[53] 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 Trans. Graph* 39 (2020), 222:1–222:16.

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

[55] Fisher Yu and Vladlen Koltun. 2016. Multi-scale context aggregation by dilated convolutions. In *Proc. ICLR*.

[56] Fajrian Yunus, Chloé Clavel, and Catherine Pelachaud. 2019. Gesture Class Prediction by Recurrent Neural Network and Attention Mechanism. In *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents* (Paris, France) *(IVA '19)*. Association for Computing Machinery, New York, NY, USA, 233–235.

[57] Fajrian Yunus, Chloé Clavel, and Catherine Pelachaud. 2021. Sequence-to-Sequence Predictive Model: From Prosody to Communicative Gestures. In *Digital Human Modeling and Applications in Health, Safety, Ergonomics and Risk Management. Human Body, Motion and Behavior*. Springer International Publishing, Cham, 355–374.