Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 1081–1088
Marseille, 11–16 May 2020

Creating a Corpus of Gestures and Predicting the Audience Response based on Gestures in Speeches of Donald Trump

Verena Ruf[1,2], **Costanza Navarretta**[1]
[1]University of Copenhagen, [2] Technical University Kaiserslautern
vruf@physik.uni-kl.de, costanza@hum.ku.dk

Abstract

Gestures are an important component of non–verbal communication. This has an increasing potential in human–computer interaction. For example, Navarretta (2017b) uses sequences of speech and pauses together with co–speech gestures produced by Barack Obama in order to predict audience response, such as applause. The aim of this study is to explore the role of speech pauses and gestures alone as predictors of audience reaction without other types of speech information. For this work, we created a corpus of speeches held by Donald Trump before and during his time as president between 2016 and 2019. The data were transcribed with pause information and co–speech gestures were annotated as well as audience responses. Gestures and long silent pauses of the duration of at least 0.5 seconds are the input of computational models to predict audience reaction. The results of this study indicate that especially head movements and facial expressions play an important role and they confirm that gestures can to some extent be used to predict audience reaction independently of speech.

Keywords: Multimodal Communication, Machine Learning, Audience Response

1. Introduction

Conscious or subconscious gesturing is part of non–verbal communication (Argyle, 2010). Therefore, gestures, such as hand gestures, head movements, facial expressions or body posture, which are connected to speech (Kendon, 2004; McNeill, 2015), are called co–speech gestures.

Politicians also use gestures and, because they frequently hold public speeches about their goals and plans, their speeches are often available online and have been objects of various studies. Navarretta (2017b) analyses two speeches by Barack Obama at the White House Correspondents' Association Dinner and finds that sequences of speech, pauses and these co–occurring gestures can be employed to predict audience response using machine learning. The content of the spoken sequences is not included in her study. The aim of this paper is to investigate whether long speech pauses and co–speech gestures alone contribute to the prediction of audience responses in speeches by a speaker different from Obama. Accordingly, the focus of this work lies on co–speech gestures and pauses. The findings could, for instance, be applied in future research concerning multimodal communication with robots or other communicative interfaces regarding, for example, the extent of gesturing necessary for eliciting a response from an interlocutor.

The paper is organised as follows: First, in section 2., background literature is discussed. Section 3. contains a description of the data, section 4. includes a qualitative analysis of a short extract and section 5. presents the computational models used in the prediction experiments. A discussion completes the paper (section 6.).

2. Background

The two main purposes of political speeches are to explain political decisions and to establish shared values (Charteris-Black, 2018; Longobardi, 2010) with the aim of convincing the audience (Longobardi, 2010). It is therefore important for the speakers that the audience pay attention to what is said. A sign of this is audience reaction, such as applause (Atkinson, 1984). Audience reaction has to be simultaneous and similar between people to be recognised as such (Atkinson, 1984). An example is clapping. Mann et al. (2013) found that on average the first person starts clapping 2.1 seconds after a presentation ends, with the last person starting to clap 2.93 seconds later. Applause usually starts slowly, which allows people to join in if they missed the beginning, and then dies down after approximately 7 – 9 seconds (Atkinson, 1984; Kurzon, 1996). Speakers use several techniques to gain applause, either consciously or unconsciously (Atkinson, 1984; Kurzon, 1996). These techniques include pauses, stressed words, or slowed speech (Atkinson, 1984; Kurzon, 1996).

Pauses can control speech pace, can be used to introduce new information, and help structure speeches (Esposito and Esposito, 2011; Hirschberg and Nakatani, 1998; Kurzon, 1996; Navarretta, 2017a). They are used in communication management and in some cases have the same function as gestures (Allwood et al., 2005a). For instance, pauses can be used to synchronise speech and gesture (Kendon, 2004; Loehr, 2007). Furthermore, they are indicators of applause as an increased use of pauses often precedes an audience reaction (Kurzon, 1996).

When using gestures in the context of speech, gesture and speech are synchronous and gestures often occur accompanying stressed syllables (Argyle, 2010). They are employed to enhance understanding, for example, in noisy environments (Kendon, 2004). Another purpose of gestures is conversation management, such as nodding to give feedback (Allwood, 2002; Allwood et al., 2005b; Allwood et al., 2007; Argyle, 2010) or gaining attention (Kimbara, 2014).

Gestures can be predicted via machine learning (Itauma et al., 2012; Mori et al., 2006). In these cases the results are often aimed at enabling the gesturing of robots (Itauma et al., 2012) or facilitating the imitation of gestures in real time (Mori et al., 2006). Furthermore, machine

learning is used in the context of predicting audience reaction (Navarretta, 2017b; Strapparava et al., 2010). Strapparava et al. (2010) predict if certain sentences can trigger applause and are accordingly particularly persuasive. Their results are promising. Navarretta (2017a) analyses humorous speeches by Barack Obama during the Annual White House Correspondents' Association Dinner in 2011 and 2016. Binary and trinary sequences of speech, pauses and applause together with Obama's co–speech gestures are then employed to predict applause with different machine learning models (Navarretta, 2017b). Her best results are f1–scores of 0.825 for an input of trinary sequences and employing a Naive Bayes and a Multilayer Perceptron model. Although her results indicate that series of events, more specifically speech, speech pauses and co–speech gestures, provide the best results, she also finds that co–speech gestures play a role in predicting applause in the humorous speeches by Obama.

The aim of this paper is to further investigate Navarretta (2017b)'s observation and to test it for other types of speeches and with a different speaker.

3. Data

For this study a corpus consisting of three speeches of Donald Trump between 2016 and 2019 is constructed. The speeches are not humorous as in (Navarretta, 2017b), but are political speeches.

Donald Trump was chosen because of the amount of research into his rhetoric after the 2016 presidential election. Moreover, Trump holds the same office as Barack Obama did in the speeches analysed in (Navarretta, 2017b) and the entertainment value of Trump's speeches is assumed to hold the attention of the audience similarly to humorous speeches (Charteris-Black, 2018; Kranish and Fisher, 2017).

The first speech is Donald Trump's rally speech in Toledo, Ohio, on 27 October, 2016[1]. Only the first 21 minutes of this speech were included in this study.

The second speech is Trump's Inaugural Address on 20 January, 2017. A transcript, as prepared for delivery, and the video can be seen under `https://www.whitehouse.gov/briefings-statements/the-inaugural-address/`. The actual speech and therefore the annotated part lasts approximately 17 minutes (from minute 28:30 – 45:30). A picture from the speech is shown in figure 1.

The last speech included in the corpus is the State of the Union Address which was held by Donald Trump on 5 February, 2019, at the Congress[2]. The speech lasts approximately 82 minutes. The whole annotated corpus has a duration of two hours.

Figure 1: A snapshot from the Inaugural Address Speech.

The speeches and long silent pauses (≥ 0.5 seconds) were transcribed in PRAAT (Boersma and Weenink, 2009) and annotated in the ANVIL tool (Kipp, 2001; Kipp, 2005) with annotations according to the MUMIN coding scheme v.3 (Paggio and Navarretta, 2008). The data consists of gesture annotations in different tracks: one track for speech, including pauses; one track for hand gestures; one for head movements and facial expressions; and one for changes in body posture. For each gesture the following information is provided: the physical form (for instance, smiling or hand movement to the right), the communicative function (feedback, turn management), the semiotic type, as well as its relation to speech. A track for audience response is added with choices of positive, negative, or neutral response. As Donald Trump clapped in several instances an attribute "Clapping" is added to the track for hand gestures. All features are shown in table 1.

To test whether the categories are assigned in a consistent way, an intercoder agreement experiment is conducted. In this a second coder independently annotated the Inaugural Speech. The intercoder agreement scores are calculated automatically in ANVIL. In table 2 the overall agreement (segmentation and classification) for the main identification of head movements, facial expressions, hand gestures and body postures is reported in terms of Cohen (1960)'s κ.

The intercoder agreement scores for both segmentation and classification of head movements, hand gestures and facial expression is high (κ 0.77 – 0.96). In particular, this is the case for hand gestures, since both coders identified the same gestures, but marked the start or end time of a gesture in a different frame. With respect to facial expressions, cases of disagreement are all the following: One coder identifies some of Trump's facial expressions as voluntary (displays) indicating that Trump wants to show that the subject is serious (the facial expressions are classified as *Scowl*), while the other coder does not mark them. The

[1]The video is available at `https://www.youtube.com/watch?v=BBPZIlj1Vf4`. The transcript of the speech was found under `https://factba.se/transcript/donald-trump-speech-toledo-oh-october-27-2016`. Since several remarks were not included in the transcript, they were added in the corpus.

[2]The speech is available on YouTube from the White House Channel under `https://www.youtube.com/watch?v=fpf1IYU0poY&t=3s`. The transcript can be found under `https://www.whitehouse.gov/briefings-statements/remarks-president-trump-state-union-address-2/`.

Type	Features
Feedback	Give, Elicit, Understand, NonUnderstand, Accept, NonAccept
Turns	Take, Accept, Yield, Elicit, Complete, Hold
Inf.Structure	True, False
Relationtospeech	Addition, Reinforcement, Substitution, Contradiction, Other
Semiotic Type	IndexDeictic, IndexNonDeictic, Iconic, Symbolic, IconicandIndexNon-deictic, SymbolicandIndexNon-deictic
Face	Smile, Laughter, Scowl, FaceOther, EyebrowsFrown, EyebrowsRaise, EyebrowsLifted, BrowsOther, EyesX–Open, EyesCloseBoth, EyesCloseOne, EyesCloseRepeat, EyesOther, GazeForward, GazeBackward, GazeUp, GazeDown, GazeSide, GazeDirectionOther, GazeToInterlocutor, GazeAwayFromInterlocutor, OpenMouth, CloseMouth, LipsCornersUp, LipsCornersDown, LipsProtruded, LipsRetracted, LipsOther, Nod, Jerk, HeadBackward, HeadForward, Tilt, SideTurn, Waggle, HeadOther, HeadRepeatedSingle, HeadRepeated
Hand gestures	SingleHand, BothHands, PalmOpen, PalmClosed, PalmOther, PalmUp, PalmDown, PalmSide, PalmPosOther, IndexExtended, ThumbExtended, AllFingersExtendend, FingersOther, AmplitudeCentre, AmplitudePeriphery, AmplitudeOther, Trajectory for left and right hand: Forward, Backward, Side, Up, Down, Complex, Other, RepeatedSingle, Repeated, VisibleClapping, AudioOnlyClapping
Body posture	Forward, Backward, Up, Down, Side, Other
Audience	Positive, Negative, Both

Table 1: Coding features.

cases of disagreement for head movements are due to their segmentation (start/end frames are not always exactly the same) and classification disagreement between the types *Waggle* and *HeadOther*. The lower agreement for Body Posture ($\kappa = 0.596$) is exclusively due to the fact that one coder judges many turning body movement by Trump as feedback eliciting or giving signals since they occur before or after the audience's response, while the other coder does not judge them to be communicative. Only the annotations which both annotators agree upon are included in the experiments. Body postures are excluded since they are not frequent. It must also be noted that the agreement scores for some of the sub–categories of hand gestures, such as SemioticType and PalmPosition, are lower than those obtained when classifying the general category, but these scores are still good (κ between $0.70-0.83$). The same level of agreement is also obtained for the subcategories of head movements.

Gesture type	Cohen's κ
Head Movement	0.769
Facial Expression	0.847
Hand Gestures	0.964
Body Posture	0.596

Table 2: Intercoder Agreement Scores for the Gestures in the Inaugural Speech

The annotations of the three speeches are exported from ANVIL. The corpus contains 2250 gestures, including 709 head movements, 104 facial expressions, 1296 hand gestures, and 239 changes in body posture. 98 of the facial expressions co–occur with head movements. There are 206 instances of audience reaction and 1218 silent pauses $\geq$ 0.5 seconds. The frequencies (occurrences per second) of gestures, audience responses and pauses for the individual speeches can be seen in table 3. Hand gestures and head movements are the most frequently produced gestures, while Trump very seldom moves his body or shows a facial expression.

Type	Rally speech in Toledo, Ohio	Inaugural Address	State of the Union Address 2019
Hand gestures	0.35	0.26	0.11
Head movements	0.10	0.04	0.11
Facial expressions	0	0.01	0.01
Body posture	0.003	0.004	0.04
Pauses	0.18	0.12	0.17
Audience response	0.05	0.03	0.02

Table 3: Frequencies of gestures, pauses, and audience response during the speeches included in the corpus.

4. Qualitative Analysis

A transcript from a short qualitative analysis of an extract of the State of the Union Address (from minute 45:33 – 45:56) can be seen in example 1. Gesture preparation was marked by $\sim$, strokes by $*$, holds after strokes by $\underline{*}$, and retractions by $-\cdot-$. Head gestures were labelled as 'hg' and hand ges-

tures with forearm movements as ‘fg’. Almost all strokes occurred on the stressed syllables of the concurrent words.

Ex. 1.

Tonight I am also asking you to pass the United States (0.54)
Reciprocal Trade Act (0.76)
∼∼ * * * * * * * * * * −·−
[——— hg1 ———]

so that if another country places an unfair tariff (0.68)
∼∼∼∼∼∼∼∼∼ * * * * | * * * | ∼ * * * * | * * * * * *
[———————————— fg1 ————————————]

on an American product (1.15)
∼∼∼ * * * * * | * * * · -
[———- fg2 ———]
∼∼ * * * * * * · -
[—— hg2 ——]

we can charge them the exact same tariff
∼∼∼∼∼∼∼ * * * * ∼∼ * * | * * * | * *
[——— fg3 ———][——— fg4 ———]

on the exact same product (0.65)
∼∼∼ * * * | * * * | * * * * * * * *
[————————- fg5 ————————-]

that they sell to us (Applause: 10.12)
* * * * * * * * * * * * −· − ·−
[——— fg6 ———]
∼∼∼∼ * * | * * *
[——— hg3 ———]

Example 1: Gestures in an extract of the State of the Union Address (from minute 45:33 – 45:56).

In this part of the talk, Donald Trump spoke about the United States Reciprocal Trade Act and associated tariffs. The camera moved to a close–up view of Donald Trump during his pronunciation of *asking*. The first gesture, hg1, occurred when Donald Trump uttered *Reciprocal Trade Act*. He started to tilt his head to the left when saying *Reciprocal* and then held his head in this position for the remainder of the phrase. A comparable head tilt can be seen in figure 1. The gesture’s exact attributes are shown in table 4.
In the following pause Donald Trump moved his head back to rest position. This gesture was therefore used to structure the discourse as it occurred during the first mention of the topic.
After he moved his head back to rest position, Donald Trump prepared a hand gesture by raising his right hand with his thumb and index finger forming a ring (fg1), which was a typical gesture in this corpus. This was encoded as a

Figure 2: A snapshot of a beat gesture during the State of the Union Address.

Attributes	Type
HeadMovement	Tilt
InformationStructure	InfoStructure
SemioticType	IndexNon-deictic
Reinforcement	*Reciprocal*, *Trade*, *Act*

Table 4: Gesture attributes and types of hg1.

separate gesture in the corpus (see table 5). As it is clearly a preparatory gesture, it was combined with the following gesture for the qualitative analysis.

Attributes	Type
Handedness	SingleHand
Palm	PalmClosed
PalmPos	PalmPosOther
TrajectoryRightHand	RightHandUp
SemioticType	IndexNon-deictic
Reinforcement	*so*, *that*

Table 5: Gesture attributes and types of the preparation of fg1.

Then the fingers extended. According to Kendon (2004) the opening of a ring–position is often followed by specific aspects of a topic, as was the case here. The words *country*, *places*, *unfair*, and *tariff* were emphasised by moving the hand up and down in beat gestures with the downward strokes occurring during the stressed syllables. The gesture’s attributes can be seen in table 6.
The first two gestural strokes were very similar, subsequently Donald Trump moved his hand slightly to his centre and turned a little when saying *an* to prepare for the next two beats. The first of these was a rather small movement, which unlike the others did not occur on the stressed syllable of the word, and was followed by a large down– and side–movement of the right hand, where the position was

Attributes	Type
Handedness	SingleHand
Palm	PalmOpen
PalmPos	PalmSide
TrajectoryRightHand	RightHandComplex
HandRepetition	Repeated
SemioticType	IndexNon-deictic
Reinforcement	*country*, *places*, *unfair*, *tariff*

Table 6: Gesture attributes and types of fg1.

held during the pause and retracted afterwards. The difference in amplitude made it clear that tariffs were the main reason for the Trade Act.

Donald Trump then explained his point further by stating that these tariffs were placed on American products. In fg2 he prepared his next gesture by shifting his right hand back towards the centre of his body. Then he moved his hand laterally to the side with his palm facing downwards. This movement was repeated twice. A depiction from this gesture can be seen in figure 2. However, the second beat additionally included a downward movement. Afterwards, Donald Trump returned the hand to the rest position gripping the lectern. Kendon (2004) classified this as an 'open hand prone' gesture, which usually indicates the interruption or stopping of an action. In this case, *unfair tariff[s] on [...] American product[s]* should end.

When Donald Trump said *product*, he tilted his head to the left again and held it in this position during the pause (hg2). This second tilt differed from the first in hg1 because Donald Trump did not tilt his head completely to the left. Instead, his rest position during the passage was a right tilt of the head and the left tilt in this case was a movement of the head in such a way that the neck was straight and the gaze pointed forwards. This was classified as a gesture used to make clear that tariffs were being placed on products as the head was moved back to the rest position of a right tilt. The implication of the statement was that American products would be more expensive in countries with tariffs, which would be bad for American companies.

Based on this, Donald Trump emphasised the main point of the Reciprocal Trade Act by a series of beat gestures (fg3 – fg5). The preparation was the same as in fg1: raising the right hand in an open fist position. Donald Trump then made the first beat gesture with the hand in this position while uttering *them*, emphasising a retaliation against the countries hurting the U.S.A., a reference to *another country* mentioned before while the hand was in the same position. After this, Donald Trump opened the hand and made three beat gestures with an open hand and extended fingers followed together with the words *exact*, *same*, and *tariff*. During this series of beats, the amplitude of the gestures expanded with the maximum amplitude reached while uttering *tariff*, highlighting the main point again. Afterwards, Donald Trump raised the hand to shoulder height while the fingers formed a kind of L–shape, with middle–, ring– and little finger extended at a 90° angle. The index finger was also held at this angle, but was not extended. Subsequently, Donald Trump moved the hand up and down two times concurrently with the words *exact* and *same*. During the pronunciation of *product*, in addition to moving the hand downwards he also moved it to the side, therefore adding emphasis to this word as before. This position was held in the following pause.

Donald Trump then moved both arms to his side, opening them widely, with the hand in an open position, palms facing to the audience and fingers extended in fg6. During this he said that they sell. Afterwards he held this position while completing the sentence. Furthermore, he moved his head and torso forwards and backwards two times during this phrase on the words *sell* and *to us* with the forward strokes occurring during *sell* and *us* (hg3). This combination of hand gestures and head movements was therefore used to emphasise *sell* and highlight the disparity between other countries selling to Americans, but impeding America's trade with them. A sense of belonging was stressed with *us*, which not only included the present audience, but all Americans.

After the argument was completed, applause set in. Donald Trump then retracted his hand gesture by placing his right hand on the lectern and relaxing his left arm completely. He subsequently stepped back and let go of the lectern.

The use of gestures in this extract exemplified gesture use in the speeches included in the corpus. It illustrated that Donald Trump mainly used gestures to emphasise a point, mostly with beat gestures, or to connect different parts of a sentence, for instance, by using the same gesture for associated words.

5. Prediction Experiments

The aim of our prediction experiments was to investigate to what extent information about long silent pauses and gestures can predict audience response in the three speeches by Donald Trump. Long silent pauses were chosen as a speech feature enabling the comparison with other research.

Pyhton's scikit–learn package was used to program the models (Pedregosa et al., 2011). The overall structure of the machine learning part of the code was based on Brownlee (2016).

As the amount of negative audience reactions was very small and all types of audience reaction in the corpus included clapping, no distinction between positive and negative audience reaction was made. For each speech the gestures correlated with audience reaction occurring in the range of 5 seconds before it started and 5 seconds after the start of the applause were labelled as leading to an audience reaction. 5 seconds were chosen because the duration between the end of a presentation and the last person to start clapping is 5.03 seconds on average according to previous studies, such as Mann et al. (2013). This also accounted for the fact that Donald Trump sometimes continued speaking after the applause started and applause increased gradually as well as possible errors due to manual annotation. For example, the annotator might not have heard the exact start of the applause and only recognised it after the whole audience joined.

Overlap was defined as co–occurring gestures or pauses. Overlapping gestures for each speech were found by com-

paring start– and end–times and subsequently grouping gestures which co–occurred. For features that occurred in two overlapping gestures the mean–value was used. Mixed types that could not be computed were disposed of by first converting all features to strings and then encoding them with labels using the LabelEncoder. The data were split into a training part consisting of 80 % of the data and a testing part consisting of 20 % of the data. The models were then trained using 10–fold cross–validation. This process split the data into subsets (in this case 10) that are approximately equally sized and trained the model on all subsets but one (Burkov, 2019; Theodoridis, 2015). The last subset was then used for testing (Burkov, 2019; Theodoridis, 2015). The final parameters of the model were set as the average of the models trained during cross–validation (Burkov, 2019; Theodoridis, 2015).
The sklearn stratified Dummy–Classifier was chosen as baseline estimator as it predicts labels based on their distribution.

5.1. Model Evaluation

First, correlations between different kinds of gestures and audience reaction were tested for each speech included in the corpus. As the data were not normally distributed, Spearman's correlation was used. In the following only significant correlations were reported. There were significant correlations between hand gestures, head movements with facial expressions and audience response. In the rally speech the correlations for both hand gestures ($r = 0.66, p = 1.38e^{-08}$) and head movements including facial expressions ($r = 0.38, p = 0.003$) with audience reaction were significant. In the Inaugural Address the correlation of hand gestures and audience response was significant ($r = 0.75, p = 7.8e^{-08}$). In the State of the Union Address the correlation between head movements with facial expressions and audience reaction was significant ($r = 0.38, p = 5.26e^{-05}$). The correlation between gestures and pauses was significant in the State of the Union Address ($r = 0.34, p = 0.0003$).
F1–score (f1), precision (P), and recall (R) for the different models can be seen in tables 4 and 5. F1–score was calculated with equation 1. The values were given as weighted average, which took the total occurrences for each label into account. Therefore, the f1–score could be beyond the range given by precision and recall.

$$f1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \quad (1)$$

In tables 4 and 5 HF stands for head movements and co–occurring facial expressions, HG stands for hand gestures and PA stands for silent pauses.
The various models were produced with the following algorithms: Logistic Regression, k–Nearest Neighbor, Gaussian Naive Bayes, Support Vector Machine and Perceptron. The best results were obtained with the model produced by k–Nearest Neighbor and the improvement with respect to the baseline is significant[3].

[3]Paired corrected t-test and significance level $p < 0.001$.

Feature	Baseline	LR	kNN
HF	f1 = 0.47 P = 0.47 R = 0.48	f1 = 0.61 P = 0.68 R = 0.65	f1 = 0.70 P = 0.73 R = 0.71
HG	f1 = 0.47 P = 0.47 R = 0.48	f1 = 0.61 P = 0.68 R = 0.65	f1 = 0.70 P = 0.73 R = 0.71
HF + HG	f1 = 0.56 P = 0.55 R = 0.56	f1 = 0.64 P = 0.71 R = 0.72	f1 = 0.66 P = 0.66 R = 0.70
PA	f1 = 0.51 P = 0.51 R = 0.50	f1 = 0.72 P = 0.78 R = 0.76	f1 = 0.75 P = 0.75 R = 0.75
PA + HF	f1 = 0.47 P = 0.47 R = 0.48	f1 = 0.64 P = 0.72 R = 0.68	f1 = 0.70 P = 0.73 R = 0.71
PA + HG	f1 = 0.52 P = 0.52 R = 0.53	f1 = 0.65 P = 0.69 R = 0.80	f1 = 0.60 P = 0.62 R = 0.62
PA + HF + HG	f1 = 0.53 P = 0.53 R = 0.53	f1 = 0.64 P = 0.68 R = 0.68	f1 = 0.64 P = 0.64 R = 0.65

Table 7: Results of baseline algorithm, Logistic Regression (LR), and k–Nearest Neighbor (kNN) for different feature combinations.

Feature	NB	SVM	Perceptron
HF	f1 = 0.60 P = 0.72 R = 0.66	f1 = 0.41 P = 0.32 R = 0.57	f1 = 0.26 P = 0.19 R = 0.43
HG	f1 = 0.60 P = 0.72 R = 0.66	f1 = 0.41 P = 0.32 R = 0.57	f1 = 0.26 P = 0.19 R = 0.43
HF + HG	f1 = 0.58 P = 0.56 R = 0.69	f1 = 0.57 P = 0.49 R = 0.70	f1 = 0.63 P = 0.74 R = 0.72
PA	f1 = 0.64 P = 0.64 R = 0.64	f1 = 0.71 P = 0.77 R = 0.75	f1 = 0.17 P = 0.12 R = 0.34
PA + HF	f1 = 0.60 P = 0.72 R = 0.66	f1 = 0.41 P = 0.32 R = 0.57	f1 = 0.26 P = 0.19 R = 0.43
PA + HG	f1 = 0.53 P = 0.72 R = 0.62	f1 = 0.41 P = 0.32 R = 0.57	f1 = 0.26 P = 0.19 R = 0.43
PA + HF + HG	f1 = 0.48 P = 0.57 R = 0.62	f1 = 0.47 P = 0.38 R = 0.62	f1 = 0.62 P = 0.68 R = 0.67

Table 8: Results of Gaussian Naive Bayes (NB), Support Vector Machine (SVM), and Perceptron for different feature combinations.

6. Discussion

A corpus of annotated speeches by Donald Trump is created to predict audience response from gestures. Several

machine learning models are built to predict audience response. Furthermore, different combinations of gestures and silent pauses during the speeches included in the corpus are used as input. Gesture types are chosen based on their correlation with audience reaction.

The best results are obtained for k–Nearest Neighbor (kNN) using only pauses as input ($f1 = 0.75$). kNN is also the best overall model, except for input using a combination of pauses and hand gestures and a combination of pauses, hand gestures, and head movements with facial expressions. For both Logistic Regression produces better or the same results. Logistic Regression achieves the second best results using the other combinations.

The best feature is only pauses. This implies that pauses are better indicators of audience response than gestures. The results could, however, be influenced by the State of the Union Address, which accounts for 2/3 of the data and during which Donald Trump more frequently uses pauses than gestures (see table 3). The second best feature combinations are the combination of hand gestures and head movements with facial expressions, and the combination of pauses and head movements with facial expressions. Since the common feature of the two combinations are head movements, this could mean that head movements are more informative than other kinds of gestures with respect to the onset of audience reaction. A short analysis shows that most head gestures coinciding with applause are of the type *IndexNon–deictic* and often are repeated. The most common types of head movements and facial expressions coinciding with applause are tilts, moving the head forward, and nodding with over a 200 occurrences for tilts and over 100 instances for forward movements and nodding.

Navarretta (2017b) predicts audience reaction for humorous speeches of Barack Obama based on multimodal n–grams consisting of sequences of speech, pauses and co-speech gestures, as well as audience response. The main difference to the models in the present study is therefore that Navarretta (2017b) uses more speech features, such as speech duration and sequences (bi– and tri–grams) of multimodal events, as data. Actual speech contains more information, which probably is the reason for better predictions. Navarretta (2017b)'s best results are a f1–score of 0.825 for both bi– and trigrams of events. The best results presented here are for pauses only ($f1 = 0.75$). This is similar as pauses are related to speech and it supports Kurzon (1996), indicating that speech and linguistic features are the most important factor for predicting applause. The results also show that speech pauses are a means to gain applause, endorsing Atkinson (1984), and can be used to predict it. This is particularly promising since speech pauses can be automatically extracted in tools such as PRAAT.

However, the results imply that gestures can be employed to predict audience response to some extent as well. Particularly head movements seem to be good indicators. This supports Navarretta (2017b) and demonstrates that her findings are valid for other types of speeches and different politicians.

The results also confirm that machine learning can be used to predict audience reaction. Furthermore, they indicate that these techniques can be employed in other areas, such as communication technology, in order to improve HCI, for example, by learning the amount of gestures that have to be employed to successfully get a response.

There are some limitations to this study. First, only one annotator generated the corpus. Second, the distribution of and frequency of various kinds of gestures is quite different between the three speeches included in the corpus. This could have influenced the results. Third, no distinction is made between gestures categorised as *other*, and, finally, facial expressions and head movements are annotated in the same track since facial expressions were rare in the data.

In the future, we should investigate the role of pauses and gestures on audience response in more types of data and apply Navarretta (2017b)'s strategy of predicting audience response using trigrams of sequences of speech, pauses and audience response on this corpus.

7. Bibliographical References

Allwood, J., Ahlsén, E., Lund, J., and Sundqvist, J. (2005a). Multimodality in own communication management. In *Proceedings from the Second Nordic Conference on Multimodal Communication*.

Allwood, J., Cerrato, L., Dybkjaer, L., Jokinen, K., Navarretta, C., and Paggio, P. (2005b). The mumin multimodal coding scheme. *NorFA yearbook*, 2005:129–157.

Allwood, J., Cerrato, L., Jokinen, K., Navarretta, C., and Paggio, P. (2007). The mumin coding scheme for the annotation of feedback, turn management and sequencing phenomena. *Language Resources and Evaluation*, 41(3-4):273–287.

Allwood, J. (2002). Bodily communication dimensions of expression and content. In *Multimodality in language and speech systems*, pages 7–26. Springer.

Argyle, M. (2010). *Bodily Communication*. Routledge, Florence.

Atkinson, M. (1984). *Our masters' voices, the language and body language of politics*. Routledge, London.

Boersma, P. and Weenink, D., (2009). *Praat: doing phonetics by computer*. Retrieved May 1, 2009, from http://www.praat.org/.

Brownlee, J. (2016). Your first machine learning project in python step–by–step. Technical report, Machine Learning Mastering.

Burkov, A. (2019). *The Hundred–Page Machine Learning Book*. Andriy Burkov.

Charteris-Black, J. (2018). *Analysing political speeches*. Macmillan International Higher Education.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Esposito, A. and Esposito, A. M. (2011). On speech and gestures synchrony. In *Analysis of Verbal and Nonverbal Communication and Enactment. The Processing Issues*, pages 252–272. Springer.

Hirschberg, J. and Nakatani, C. H. (1998). Acoustic indicators of topic segmentation. In *Fifth International Conference on Spoken Language Processing*.

Itauma, I. I., Kivrak, H., and Kose, H. (2012). Gesture imitation using machine learning techniques. In *2012*

20th Signal Processing and Communications Applications Conference (SIU), pages 1–4. IEEE.

Kendon, A. (2004). *Gesture: Visible action as utterance*. Cambridge University Press.

Kimbara, I. (2014). The interactive design of gestures. In *Body – Language – Communication: An International Handbook on Multimodality in Human Interaction*, pages 1368–1374. De Gruyter Mouton.

Kipp, M. (2001). Anvil-a generic annotation tool for multimodal dialogue. In *Seventh European Conference on Speech Communication and Technology*.

Kipp, M. (2005). *Gesture generation by imitation: From human behavior to computer character animation*. Universal-Publishers.

Kranish, M. and Fisher, M. (2017). *Trump revealed: The definitive biography of the 45th president*. Simon and Schuster.

Kurzon, D. (1996). The white house speeches: Semantic and paralinguistic strategies for eliciting applause. *Text-Interdisciplinary Journal for the Study of Discourse*, 16(2):199–224.

Loehr, D. P. (2007). Aspects of rhythm in gesture and speech. *Gesture*, 7(2):179–214.

Longobardi, F. (2010). Linguistic factors in political speech. In *International Workshop on Political Speech*, pages 233–244. Springer.

Mann, R. P., Faria, J., Sumpter, D. J., and Krause, J. (2013). The dynamics of audience applause. *Journal of The Royal Society Interface*, 10(85):20130466.

McNeill, D. (2015). *Why we gesture: The surprising role of hand movements in communication*. Cambridge University Press.

Mori, A., Uchida, S., Kurazume, R., Taniguchi, R.-i., Hasegawa, T., and Sakoe, H. (2006). Early recognition and prediction of gestures. In *18th International Conference on Pattern Recognition (ICPR'06)*, volume 3, pages 560–563. IEEE.

Navarretta, C. (2017a). Barack obama's pauses and gestures in humorous speeches. In *European and 7th Nordic Symposium on Multimodal Communication (MMSYM 2016)*, pages 28–36.

Navarretta, C. (2017b). Prediction of audience response from spoken sequences, speech pauses and co-speech gestures in humorous discourse by barack obama. In *Cognitive Infocommunications (CogInfoCom), 2017 8th IEEE International Conference on*, pages 000327–000332. IEEE.

Paggio, P. and Navarretta, C. (2008). Mumin coding scheme. Technical report, University of Copenhagen.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830.

Strapparava, C., Guerini, M., and Stock, O. (2010). Predicting persuasiveness in political discourses. In *LREC*, pages 1342–1345.

Theodoridis, S. (2015). *Machine learning: a Bayesian and optimization perspective*. Academic Press.