Prediction of Audience Response from Spoken Sequences, Speech Pauses and Co-speech Gestures in Humorous Discourse by Barack Obama

Costanza Navarretta
University of Copenhagen
Njalsgade 136
2300 Copenhagen S
Denmark
Email:costanza@hum.ku.dk

***Abstract*—In this paper, we aim to predict audience response from simple spoken sequences, speech pauses and co-speech gestures in annotated video- and audio-recorded speeches by Barack Obama at the Annual White House Correspondents' Association Dinner in 2011 and 2016. At these dinners, the American president mocks himself, his collaborators, political adversary and the press corps making the audience react with cheers, laughter and/or applause. The results of the prediction experiment demonstrate that information about spoken sequences, pauses and co-speech gestures by Obama can be used to predict the immediate audience response. This confirms and shows an application of numerous studies that address the importance of speech pauses and gestures in delivering the discourse message in a successful way. The fact that machine learning algorithms can use information about pauses and gestures to build models of audience reaction is also relevant for the construction of intelligent and cognitively based multimodal ICT.**

I. Introduction

In face-to-face communication not only the content of words, but also the way in which they are uttered and the gestures which co-occur with them contribute to the successful deliver of the discourse message. Gestures indicate co-speech body behavior such as head movements, facial expressions and hand gestures. Speech pauses comprise silent pauses, which can be accompanied by audible breath or other sounds, and filled pauses which are pauses and short words such as the English *um*, *ah*, and *uh*.

Speech and gestures are related temporally and semantically [1], [2], and gestures contribute to both the content of the discourse [1] and the management of the interaction [3]. Speech pauses have similar functions and also contribute to the discourse content and structure [4], [5], [6], [7], [8].

The main hypothesis that we want to test in this paper is that speech pauses and gestures are so important means for the successful delivery of discourse message that they can be used to predict audience response in two humorous speeches by Barack Obama at the Annual White House Correspondents' Association Dinner in 2011 and 2016. The speeches have been chosen for different reasons. First, Barack Obama is judged to be an excellent speaker by both the press corps and researchers, such as [9]. Second, the American president, according to the tradition at the Annual White House Correspondents' Association Dinners, makes fun of himself, his wife and near collaborators, his political adversaries and the press corps. This often results in audience laughter and/or applause.

A previous study showed that speech pauses and audience reaction are positively correlated in these speeches [10]. We want to determine whether this correlation can be used to predict audience response. More specifically, we train algorithm on information about speech sequences, pauses and co-speech gestures to predict the success of the delivered message by Obama in terms of immediate audience reaction. Determining to what extent speech pauses are used as means to get audience response is important not only for understanding the way in which humans communicate, but also for implementing spoken and multimodal systems which can interact with humans in a successful and cognitively natural way and for adding this ability to systems adding cognitive capacities to humans with e.g. social impairements [11], [12]

The paper is organized as follows. In section II, we discuss background literature, then in section III we present the data. Section IV describes prediction experiments in which sequences of speech and pauses and co-occurring gestures is used to predict audience response. A discussion follows in section V.

II. Background

Speech pauses have many functions in discourse. For example, they contribute regulating the interaction [5], [13] and can signal that the speakers are planning the discourse [14], [15]. The presence of pauses can indicate that the speakers are searching for a word [16] or are talking about complex concepts [17]. Pauses are also temporally and semantically related to gestures [18], [1], [6], [7], [8] and their presence gives naturalness to conversing software agents [19]. Similarly, co-speech gestures contribute to the content and the structure of discourse [18], [1] and regulate the interaction as feedback and turn management signals [20], [21]. They also show the attitudinal state of the speakers and their interlocutors [22].

Quaglio [23] notices that ungrammatical silent pauses and rate of articulation in the sitcom *Friends* provide spoken features to the written acted manuscripts. Studies of comedy suggest that jokes are presented changing the speech rate

978-1-5386-1264-4/17/$31.00 ©2017 IEEE

and using pauses before punch lines. However, recent corpus based studies of humorous and non humorous discourse do not confirm this, but find that speakers smile and laugh more when they present humorous discourse than when they talk seriously [24]. In comedy, speech pauses have been found to structure and emphasize the discourse, and they give the audience time to reflect on the conveyed message [25], [26]. Speech pauses have also been studied in political speeches. Duez [27] compares different types of French televised interviews and political speeches and concludes that silent pauses are 50% longer in political speeches than in interviews, and that the longer pauses have a stylistic function. Analyzing Italian speeches by Silvio Berlusconi, Salvati and Pettorini [28] find an higher presence of emphatic pauses in political speeches than in other types of discourse. Guerini et al. [29] collect a corpus of transcriptions of American political speeches and add to the transcriptions of the speeches occurrences of audience reaction in the form of laughter and applause in order to find prominent discourse segments in them. Differing from Guerini et al. [29]'s data, we also account for Obama's speech pauses and co-speech gestures and investigate their relation to the audience's reaction. Audience response in these speeches is always positive and consists of applause, laughter and/or cheers.

Navarretta [10] analyzes and compares speech pauses and co-speech gestures (head movements, facial expressions, hand gestures) in two Obama's speeches at the Annual White House Correspondents' Association Dinner in 2011 and 2016. In the following, we call the two speeches *speech2011* and *speech2016*, respectively. The analysis of the speeches shows that Obama's speech rate and gestural rate do not change in the two speeches with one exception. Obama produces significantly more hand gestures in *speech2016* than in *speech2011*. An analysis of Obama's hand gestures in political speeches from the same years confirms this change in hand gesturing. Only few filled pauses were found in the data, and their main use in these data is to emphasize the preceding words. Silent pauses in the two speeches delimit grammatical phrases or topic shifts. When they precede single words, they emphasize the following speech segment. Obama also uses pauses to let the audience get the point, and in numerous cases after these pauses the audience react by laughing and/or applauding the president. A high degree of positive correlation between speech pauses and audience response was found in the two speeches. More specifically, the Pearson 2-tailed correlation r is equal to 0.465 and the correlation level is highly significant ($r(1541) < 0.0001$).

In the present work, we build on the work by [10] investigating further the relation between pauses and audience reaction in Obama's humorous speeches. Using a number of features extracted from the annotated videos, we want to determine to what extent information about sequences of speech pauses, speech and co-speech gestures can be used to predict audience response. Our expectation is that simple information about sequences of speech and speech pauses is useful for predicting audience response in Obama's humorous speeches, since both statistical analysis and qualitative analysis indicates that Obama uses pauses especially to emphasize his jokes and let the audience get the point. We also expect co-speech gestures, or the lack of them, to have some influence on audience reaction.

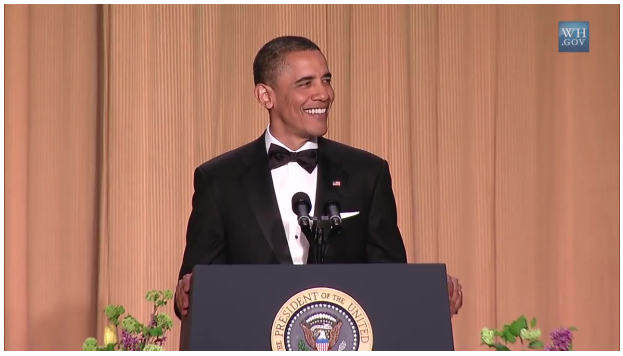

Fig. 1. Barack Obama in a snapshot from *speech2011*

Fig. 2. Barack Obama in a snapshot from *speech2016*

III. The data

In the present work, we work with multimodal annotations of *speech11* and *speech16*. The two annotated video- and audio-recorded Obama's speeches were downloaded from http:\\www.WH.gov. In these videos Obama is recorded frontally, and therefore his hand gestures, head movements and facial expressions are clearly visible. Two snapshots from *speech11* and *speech16* are in Figure 1 and Figure 2, respectively.

The duration of the annotated *speech2011* video segments is 13 minutes and 22 seconds while the duration of the annotated *speech2016* video segments is 30 minutes. Obama's speeches and pauses of at least 0.02 seconds duration were semi-automatically transcribed in PRAAT and then imported in the ANVIL tool in which the gestures were annotated according to the MUMIN annotation scheme [21].

The data consist of different annotation tracks which are temporally aligned. One track consists of the transcriptions of Obama's speeches, comprising silent and filled pauses, while three gestural tracks contain shape and functional annotations of head movements, facial expressions and hand gestures, respectively. Finally, audience response and presence of external data, such as the videos and pictures which Obama's shows to the audience, are annotated in the last two tracks.

We extracted all the information from the tracks as comma separated files and we transformed the transcriptions of Obama's speeches to simple sequences of spoken tokens, pauses tokens, audience response tokens, and external data

TABLE I. SHAPE FEATURES

Attribute	Value
HeadMovement	Nod, UpNod, HeadForward, HeadBackward, Shake, Waggle, HeadOther, Tilt, SideTurn
General face	Smile, Laugh, Scowl, FaceOther
Handedness	BothHandsSym, BothHandsAsym, RightSingleHand, LeftSingleHand

TABLE II. THE CO-OCCURRING MULTIMODAL INFORMATION

Duration	Audio	Face	Head	Hand
0.78	speech	none	nod-single	none
0.56	pause	none	none	none
1.99	speech	none	none	both-hands-single
3.64	response	none	forward-single	none
0.85	speech	smile	none	none
0.1	pause	smile	none	none

tokens. The total number of these transcribed tokens is 1563. The transcribed tokens are called auditory tokens in what follows. Over half of the auditory tokens, 811, are spoken sequence tokens, 478 are speech pauses, 261 are audience response tokens and 13 are external data tokens. Successively, via a perl script, we found the descriptions of the gestures which co-occurred with the auditory tokens and attached them to the tokens. The gesture's shape features which we used are in Table I. The shape features for head movements describe the type of movement, facial expressions are described via a simple general face attribute, and hand gestures are illustrated by the hands involved in the movement. Table II shows examples of transcription tokens and co-occurring gestures. The value *none* in the table indicates that no gesture of that type co-occurred with the auditory token.

A row in Table II is a multimodal unigram since it contains information about both the auditory modality (speech) and the visual modality (gesture). From the 1563 unigrams, we also obtained 1562 multimodal bigrams and 1560 multimodal trigrams, which were automatically obtained from the unigrams independently from the duration of the unigrams. We used bigrams and trigrams in order to determine whether information about a larger context improves the prediction of audience response. The annotations distinguish six types of audience response: *cheers*, *laughter*, *applause*, *cheers/applause*, *laughter/applause*, *cheers/laughter*. All these types of response are positive, and in our machine learning experiments, we collated them in a single category which is positive audience response. In Tables III, IV and V we show the unigrams , bigrams and trigrams data obtained from the unigrams in Table II. The last column in the three tables is the feature indicating the presence (YES) or absence (NO) of positive audience response. This is the information that the machine learning algorithms have to predict.

IV. THE PREDICTION EXPERIMENTS

The Weka machine learning platform [30] was used in the prediction experiments. The data were trained and tested in two ways. First the data was divided in two subparts. The first subpart consisting of 2/3 of the data was used for training, while the remaining subpart was used for testing. The second test and evaluation method consisted of 10-fold

TABLE III. AN EXAMPLE OF THE MULTIMODAL UNIGRAMS

multimodal1	RESULT
speech+gestures	NO
pause+gestures	NO
speech+gestures	YES
response+gestures	NO
speech+gestures	NO
pause+gestures	NO

TABLE IV. AN EXAMPLE OF THE MULTIMODAL BIGRAMS

multimodal1	multimodal2	RESULT
speech+gestures	pause+gestures	NO
pause+gestures	speech+gestures	YES
speech+gestures	response+gestures	NO
response+gestures	speech+gestures	NO
speech+gestures	pause+gestures	NO

TABLE V. AN EXAMPLE OF THE MULTIMODAL TRIGRAMS

multimodal1	multimodal2	multimodal3	RESPONSE
speech+gestures	pause+gestures	speech+gestures	YES
pause+gestures	speech+gestures	response+gestures	NO
speech+gestures	response+gestures	speech+gestures	NO
response+gestures	speech+gestures	pause+gestures	NO

cross-validation, in which the data are randomly divided in ten subparts of approximately same size. The algorithms are then trained on the first nine subparts and tested on the remaining part (a fold). The process is repeated 10 times using each time a different fold for testing and the remaining nine subparts for training. Finally, the results from all ten folds are averaged in order to get a single estimation. The best results were obtained with 10-fold cross-validation and, therefore, we report the results obtained with this method.

We trained and tested various prediction algorithms on the data: Naive Bayes, an implementation of neural networks (DI4JMIp), simple logistic, an implementation of support vector machine (SMO) and an implementation of a Multilayer Perceptron using back-propagation to classify instances. The prediction algorithms were run on the unigrams, bigrams and trigrams obtained from the annotations of *speech2011* and *speech2016* as described in section III.

In each experiment, we first trained the prediction algorithms on unimodal information, that is the transcriptions of audio data and, successively, on multimodal information consisting of the transcriptions of speech and information about co-speech gestures as shown in Tables II, III, IV and V. The best results were obtained with the Naive Bayes and the second best were achieved by the Multilayer Perceptron. In what follows, we only report the results of the former algorithm. The results of the Naive Bayes on each dataset are shown in Table VI. We use as baseline the results obtained by the algorithm on the unimodal unigram information.

The first column of Table VI shows the dataset used in prediction while the following three columns show the

TABLE VI. RESULTS OF THE PREDICTION EXPERIMENTS

data	P	R	F-score
baseline	0.694	0.833	0.757
unigram:transcr	0.820	0.834	0.761
unigram:transcr+duration	0.745	0.830	0.759
unigram:transcr+multimodal	0.786	0.834	0.776
bigram:transcr	0.822	0.830	**0.825**
bigram:transcr+duration	0.778	0.832	0.774
bigram:transcr+multimodal	0.807	0.827	0.814
trigram:transcr	0.819	0.834	**0.825**
trigram:transcr+duration	0.751	0.826	0.766
trigram:transcr+multimodal	0.816	0.829	0.822

TABLE VII. RESULTS OF THE PREDICTION WITH ONLY PAUSE DURATION

data	P	R	F-score
baseline	0.745	0.830	0.759
unigram:transcr+dur	0.820	0.834	0.761
baseline	0.795	0.820	0.804
bigram:transcr+dur	0.818	0.823	0.82
baseline	0.751	0.826	0.766
trigram:transcr+duration	0.806	0.821	0.813

Precision (P), Recall (R) and F-score for the prediction. The F-score is calculated as follows:

$$F - score = \frac{2 \times Precision \times Recall}{Precision + Recall}. \quad (1)$$

Table VI shows that the Naive Bayes algorithm trained on the unigrams of the speech transcriptions achieves a 0.758 F-score. The confusion matrix from this experiment reveals that the algorithm obtains the same results of a majority classifier on this dataset, in fact it always predicts that there is no audience response. The table also shows that adding information about gestures improves prediction in unigrams, but it does not when a larger context is used, that is in the case of bigrams and trigrams. Prediction using unimodal bigrams and trigrams gives better results than training the algorithms on unigrams, and the best results are achieved on the unimodal auditory bigrams and trigrams. All the improvements in the table are statistically significant.[1]. Table VI also shows that adding information about the duration of the auditory tokens does not improve the prediction. This is due to the fact that the duration of spoken segments and audience response varies and is not connected, while the duration of pauses probably is. In order to test this, we changed the duration of not pause tokens to 0 and tested prediction on the data in which the auditory tokens were supplied with duration information. The results of these experiments are in Table VII. We use as baseline in each experiments (unigrams, bigrams and trigrams) the results obtained on the unimodal tokens in the preceding experiments.

The results in Table VII show that using pause duration improves the results with respect to the dataset in which all auditory tokens have duration information, but classification results are still not better than those achieved when the algorithms are trained on auditory data without duration information.

[1]Significance is measured with Paired Corrected t-test and the significance level is $p < 0.05$

V. DISCUSSION

We run a number of machine learning experiments on unigrams, bigrams and trigrams consisting of speech sequences, pauses and other auditory data as well as of co-speech gestures by Obama. The unigrams, bigrams, and trigrams were produced using the annotations of two Obama's humorous speeches at the Annual White House Correspondents' Association Dinner and the experiments were aimed to predict audience response. Various machine learning algorithms were tested and the best results were obtained by Naive Bayes trained and tested via 10-fold cross-validation.

The results of the experiments confirm that information comprising speech pauses of the duration of over 0.06 seconds is useful for predicting audience response. Co-occurring gestures also contribute to the prediction, but to a lesser extent than auditory data when bigrams and trigrams data are used. The best results are achieved with unimodal bigrams and trigrams. The results confirm the important role of speech pauses in the presentation of discourse in general [4], [5] and of humorous discourse in particular [25], [26]. The outcome of the experiments also demonstrates that both auditory and visual information play an important role in the successful delivery of discourse, and that they can be uses to train machine learning models which can and should be included in ICT systems. Contrary to our expectations, the duration of speech pauses does not contribute to prediction, but it must be noted that speech pauses shorter than 0.02 were in advance not annotated, since this was the threshold given to the PRAAT script which automatically extracted longer pauses from Obama's speeches.

We did not use speech content and information about intonation in our experiments. Both are essential in spoken discourse, and they should therefore be accounted for in the future, and combined with the gestural features. Other aspects that should be considered in models of communication are the communicative situation, the audience, the relation between the speaker and the audience.

REFERENCES

[1] A. Kendon, *Gesture - Visible Action as Utterance* . Cambridge University Press, 2004.

[2] D. McNeill, *Gesture and thought*. University of Chicago Press, 2005.

[3] J. Allwood, J. Nivre, and E. Ahls'en, "On the semantics and pragmatics of linguistic feedback," *Journal of Semantics*, vol. 9, pp. 1–26, 1992.

[4] F. Goldman-Eisler, *Psycholinguistics: Experiments in spontaneous speech*. London: Academic Press, 1968.

[5] S. Duncan and D. Fiske, *Face-to-face interaction*. Hillsdale, NJ: Erlbaum, 1977.

[6] A. Esposito, K. E. McCullough, and F. Quek, "Disfluencies in gesture: gestural correlates to filled and unfilled speech pauses," in *Proceedings of IEEE International Workshop on Cues in Communication*, Hawai, 2001.

[7] A. Esposito and A. M. Esposito, "On Speech and Gesture Synchrony," in *Communication and Enactment - The Processing Issues*, ser. LNCS, A. Esposito, A. Vinciarelli, K. Vicsi, C. Pelachaud, and A. Nijholt, Eds. Springer-Verlag, 2011, vol. 6800, pp. 252–272.

[8] C. Navarretta, "The functions of fillers, filled pauses and co-occurring gestures in danish dyadic conversations," in *Postproceedings of the 3rd European Symposium on Multimodal Communication*, L. U. E. Press, Ed., vol. 105, 2016, pp. 55–61.

[9] M. M. Cooper, "Rhetorical agency as emergent and enacted," *College Composition and Communication*, vol. 62, no. 3, pp. 420–449, 2011.

[10] C. Navarretta, "Speech pauses, gestures and audience laughter in english humorous speech. , copenhagen, september 29-30, 2016.fillers, filled pauses and gestures in danish first encounters," in *Extended Abstracts of 4th European and 7th Nordic Symposium on Multimodal Communication on Multimodal Communication*. Copenhagen: University of Copenhagen, September 2016, pp. 1–3.

[11] P. Baranyi and A. Csapó, "Definition and synergies of cognitive infocommunications," *Acta Polytechnica Hungarica*, vol. 9, no. 1, pp. 67–83, 2012.

[12] P. Baranyi, A. Csapo, and G. Sallai, *Cognitive Infocommunications (CogInfoCom)*. Springer, 2015.

[13] H. H. Clark and J. E. Fox-Tree, "Using uh and um in spontaneous speaking," *Cognition*, vol. 84, pp. 73–11, 2002.

[14] W. Chafe, "Cognitive Constraint on Information Flow," in *Coherence and Grounding in Discourse*, R. R. Tomlin, Ed. Amsterdam: John Benjamins, 1987, pp. 20–51.

[15] J. Hirschberg and C. Nakatani, "Acoustic Indicators of Topic Segmentation," in *Proceedings of ICSLP-98*, Sidney, 1998.

[16] R. Krauss, Y. Chen, and R. F. Gottesman, "Lexical gestures and lexical access: a process model," in *Language and gesture*, D. McNeill, Ed. Cambridge University Press, 2000, pp. 261–283.

[17] A. Reynolds and A. Paivio, "Cognitive and emotional determinants of speech," *Canadian Journal of Psychology*, vol. 22, pp. 164–175, 1968.

[18] D. McNeill, *Hand and mind: What gestures reveal about thought*. Chicago: University of Chicago Press, 1992.

[19] J. Cassell, "Embodied conversational interface agents," *Communications of the ACM*, vol. 43, no. 4, pp. 70–78, 2000.

[20] H. Sacks, E. Schegloff, and G. Jefferson, "A simplest systematics for the organization of turn-taking for conversation," *Language*, vol. 50, no. 4, pp. 696–735, 1974.

[21] J. Allwood, L. Cerrato, K. Jokinen, C. Navarretta, and P. Paggio, "The mumin coding scheme for the annotation of feedback, turn management and sequencing," *Multimodal Corpora for Modelling Human Multimodal Behaviour. Special Issue of the International Journal of Language Resources and Evaluation*, vol. 41, no. 3–4, pp. 273–287, 2007.

[22] P.Ekman, "An argument for basic emotions," *Cognition and Emotion*, vol. 6, no. 3/4, pp. 169–200, 1992.

[23] P. Quaglio, *Television Dialogue. The sitcom Friends vs. natural conversation*. Amsterdam/Philadelphia: John Benjamins Publishing Company, 2009.

[24] S. Attardo, L. Pickering, and A. Baker, "Prosodic and Multimodal Markers of Humor in Conversation," *Pragmatics and Cognition*, vol. 19, no. 2, pp. 224–247, 2011.

[25] J. Sankey, *Zen and the Art of Stand-Up Comedy*. New York: Routledge, 1998.

[26] B. Oliver, *The Tao of Comedy: Embrace the Pause*. Oliver, 2013.

[27] D. Duez, "Silent and non-silent pauses in three speech styles," *Language and Speech*, vol. 25, no. 1, pp. 11–28, 1982.

[28] L. Salvati and M. Pettorino, "A Diachronic Analysis of Face-to-Face Discussions: Berlusconi, Fifteen Years Later," in *Multimodal Communication in Political Speech. Shaping Minds and Social Action: International Workshop, Political Speech 2010, Rome, Italy, November 10-12, 2010, Revised Selected Papers*, I. Poggi, F. D'Errico, L. Vincze, and A. Vinciarelli, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 65–74.

[29] M. Guerini, D. Giampiccolo, Giovanni, Moretti, , R. Sprugnoli, and C. Strapparava, "The new release of corps: A corpus of political speeches annotated with audience reactions," in *Multimodal Communication in Political Speech. Shaping Minds and Social Action: International Workshop, Political Speech 2010, Rome, Italy, November 10-12, 2010, Revised Selected Papers*, I. Poggi, F. D'Errico, L. Vincze, and A. Vinciarelli, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 86–98.

[30] I. H. Witten and E. Frank, *Data Mining: Practical machine learning tools and techniques*, 2nd ed. San Francisco: Morgan Kaufmann, 2005.