CAN WE SEE YOUR RESPONSE BEFORE YOU SPEAK? EXPLORING LINGUISTIC INFORMATION FOUND IN INTER-UTTERANCE PAUSES

Christian Schuler[1]*, Shravan Nayak*[2]*, Debjoy Saha*[3]*, Timo Baumann*[4]

[1]*Universität Hamburg,* [2]*University of Montréal,* [3]*IIT Kharagpur,*
[4]*Ostbayerische Technische Hochschule Regensburg*

christianschuler8989@gmail.com, shravan.nayak@mila.quebec, sahadebjoy10@iitkgp.ac.in, timo.baumann@oth-regensburg.de

Abstract: In this work we assess whether there is information in pauses in-between utterances of the same or different speakers that are predictive of the following speaker's utterance. We present models that connect a person's visual features *before they speak* to their upcoming utterance. In our experiments we find that out-of-the-box pre-trained models can already reach a better-than-chance performance in correlating video embeddings to utterance embeddings. In contrast, models that attempt to predict the first word after the pause do not outperform a unigram model, indicating that our models do not read lips (based e.g. on co-articulation effects) but rather capture more fundamental aspects of the upcoming utterance.

1 Introduction

Dialog systems, virtual agents, and social robots typically pay close attention to their interaction partners while listening to their speech in order to infer the meaning and communicative function of speech and mimicry or gestures. Correspondingly, hesitations and filled pauses during turns may well be considered during the turns when determining the meaning of an utterance. However, this is rarely the case during inter-turn or inter-utterance gaps (which, in contrast, are used to determine turn-taking). This is similarly manifested in current video-audio-text models which model these components in a time-synchronous manner. For example, models may incorporate video when performing speech recognition [1].

We propose to relate pauses that occur in-between speech utterances in video with the next utterance based on the hypothesis that the pause itself holds information about the upcoming utterance. While there is some information for which other sources would exist, e.g. pertaining to the conversation overall (i.e. the setting) and to what has been spoken before (i.e. the preceding utterances), it may also tell us about the addressee's thought process while deliberating their response. This can be useful to process the response and would also be a requirement to more realistically synthesize pausing behaviours.

2 Related Work

Recently, (filled) pauses have received more attention in speech processing research. Advances have been made in detecting filled pauses [2, 3], which is a crucial first step in utilizing their potential to improve speech and dialogue systems alike. Inserting pseudo-filled-pauses has also been shown to improve the naturalness of synthetic speech [4]. Some downstream NLP tasks greatly benefit from pause information, such as entity recognition [5, 6] or multi-speaker text-to-speech using state-of-the-art large language models [7]. An in-depth investigation of turn-transition times (TTT) [8] found that direct answers come with a shorter TTT than responses that are not directly answering a question. It has been hypothesized that certain silent pauses,

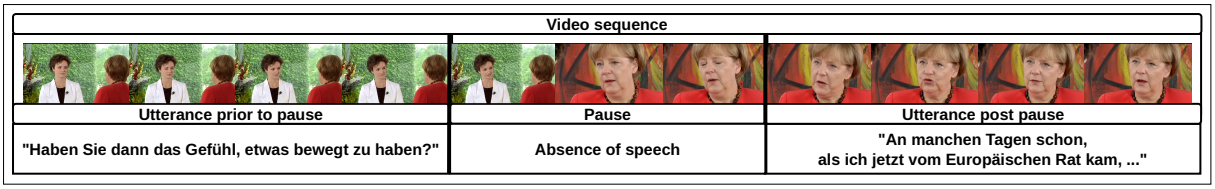

Figure 1 – Data example of a pause sequence and surrounding video sequences

under specific conditions, might also serve an interactional purpose, perceived by the speaker as a prompt to provide clarification [9]. Baumann [10] analyzed the effect of listener audio on speakers using a language modeling approach and found stronger effects turn-initially. The work of Koutsombogera and Vogel [11] provides an investigation into "speech pauses and their patterns in the data, as well as their relationship to the topics of the dialog and the turn-taking mechanism". It presents an argument advocating a focus on analyzing smaller datasets initially, before scaling up to larger data collections. The investigation of pause length effects also provides a broad spectrum of applicability in downstream tasks, such as assessment of conversation partners' cognitive state [12] and various cognitive impairments [13, 14], analysis of which is also relevant for spoken dialogue systems [15].

Despite these advances, previous work mostly focused on speech alone, neglecting the benefits that visual information might bring to the issues at hand (after all, there is no speech to be analyzed during a pause). To the best of our knowledge, this presents a literature gap that we attempt to fill with this work. Hence, we argue for including video processing and the use of video corpora for this task.

The human intuition that one could "read a response in the face" is quite clear. However, currently available data sets and applied annotation schemes are lacking when investigating the effect of pauses and their potential contribution to automatic speech recognition, language understanding and synthesis of natural speech patterns.

3 Method

We hypothesize that pausing behaviour in dialogue holds information that is helpful to predict subsequent interactional behaviour and that the content of one's upcoming utterance is reflected in the face while pondering and constructing a response.

We perform our analysis using video material from the Merkel Podcast Corpus [16] which contains a large amount of video recordings of former chancellor of Germany, Angela Merkel[1]. The corpus is split roughly equally between (semi-)prepared speeches and interview dialogues. Figure 1 shows an example turn transition in an interview. From watching many an interview, we built the intuition that some aspects of Merkel's response (e.g. relating to conciseness, spontaneity, agreement) can be 'seen' on the face even before that response is uttered.

We test our hypothesis by relating a representation of a video recording of the face during a pause in-between utterances to aspects of the upcoming utterance. We attempt to either predict the first word or assess whether we can relate to the meaning of the utterance via an embedding. If pause information is irrelevant, our models will not show significant results whereas significant results imply that there is at least *some* information in the pausing behaviour.

4 Data

We process recorded speeches and interviews from the Merkel Podcast Corpus [16] up to April 2021. We identified the positions of speech pauses in the video via voice activity detection based

[1] `https://github.com/deeplsd/Merkel-Podcast-Corpus`

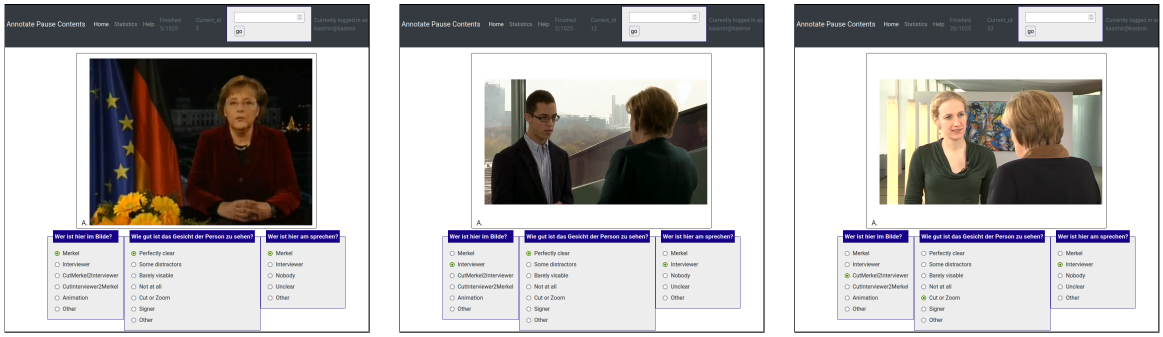

Figure 2 – Examples of pause contents (from left to right): The target speaker, a secondary speaker, a cut to Merkel during the pause.

on WebRTC-VAD[2] and identified all silent segments longer than 500 ms that roughly coincide with sentence boundaries in the annotation for a total of 1025 inter-utterance pauses. The nature of possible pause contents necessitated a more precise approach to our annotation efforts.

In our data, either Merkel, another person (most often an interviewer) or transitioning animations can be visible during the pause. Many inter-turn pauses contain a video cut, especially in interviews.[3] We focus our analysis on Merkel speaking after the cut and being frontally visible during the whole pause (regardless of shot changes). Some examples are shown in Figure 2.

We manually annotated all pauses with the speaker(s) visible (Merkel and/or non-target speakers) using Potato: the POrtable Text Annotation TOol[4] (see Figure 2). We extracted short sequences of 1 second duration in 2 second long intervals prior and following each pause snippet for annotation) that were then reviewed by a human annotator followed by automatic filtering based on the provided labels to quickly adjust the desired degree of data purity.

During annotation, we noticed many small errors in the provided transcriptions which we corrected. Many of those we attribute to the creators of subtitles trying to cover for the natural messiness of spontaneous speech such as unnecessary repetitions of words, fill-words, small mispronunciations, etc. The corrected files are part of our project repository and can be used as a partial update to the Merkel Podcast Corpus.

Filtering to only include items in which Merkel is clearly and frontally visible during the pause and also the person speaking after the pause has ended yields 684 items. In these we observed 194 unique first words of a very imbalanced distribution (137 words seen only once). We find the mean entropy for encoding each of the 194 first words of each utterance to be 6.25. The mean/median duration of pauses is 750/680 ms, 95 % of pauses are shorter than 1230 ms (note that we exclude pauses shorter than 500 ms).

5 Experiment

We assess our hypothesis with two kinds of experiments:

1. We train a classifier to predict the first word of the utterance based on a representation of the pause video. If this model outperforms the unigram perplexity, there must be information in the pause.
2. We design a discriminator that estimates whether a pause video and utterance form a pair or not; we train this discriminator using contrastive learning. If this model is able to predict pairings with better-than-random performance, there must be information in the pause.

We try three kinds of video representations: We utilized **OpenFace** [17][5] to harvest facial

[2] `https://github.com/wiseman/py-webrtcvad`

[3] Not all cuts change the scene, as they can transition e.g. from a long shot to a close-up.

[4] `https://github.com/davidjurgens/potato`

[5] `https://github.com/TadasBaltrusaitis/OpenFace`

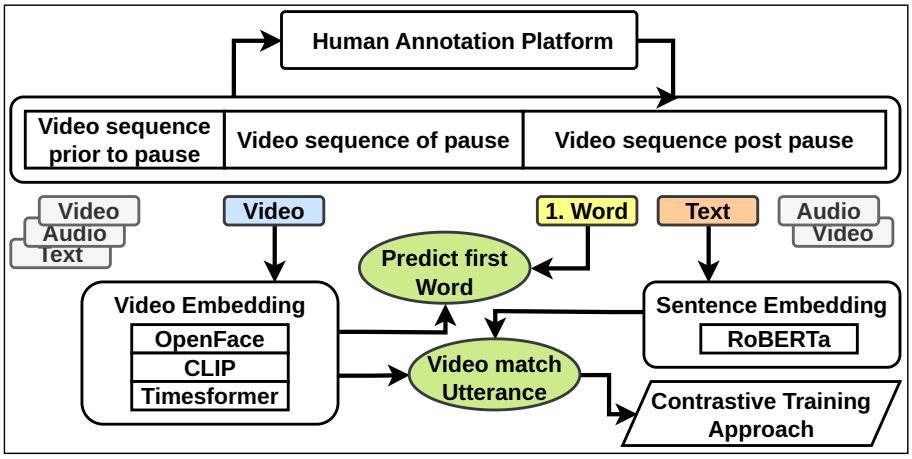

Figure 3 – General overview of our experiment setup for first word prediction and discriminator training.

features, including facial landmarks, eye gaze, head pose, and action units. To derive a single embedding for each video, we calculated the mean and standard deviation of these features across various frames. These statistical measures were then concatenated to form the final feature vector, providing a comprehensive representation of facial dynamics throughout the video. Secondly, we use the **CLIP** [18] image encoder to process the video frames. CLIP has been pre-trained on millions of image-text pairs and has demonstrated remarkable capabilities in zero-shot learning scenarios. By averaging the CLIP features extracted from all frames, we obtained a cohesive video vector that we use to train our models further. Lastly, we explored feature extraction through **Timesformer** [19], a transformer-based [20] model specifically designed for video classification. Timesformer approaches video analysis by dissecting the video into a series of patches derived from individual frames and applying a divided space-time attention mechanism to effectively capture the temporal and spatial essence of the video. The features obtained from the last layer of Timesformer were averaged to produce the final video embedding. This method allows for a nuanced understanding of the video content, leveraging the model's ability to interpret complex visual sequences.

For the contrastive learning approach, we encode the text of the utterance using the Sentence Transformers Python framework [21] and the Cross English & German RoBERTa for Sentence Embeddings model[6], which works well for computing sentence embeddings for English and German text.

Our classification approach uses the video representation as input and contains simple softmax layer that is initialized with the unigram probabilities of the corpus.

Our discrimination approach concatenates the video and utterance representations that feeds to a single logistic regression to determine whether video and utterance belong together. We implement contrastive learning by selecting, in addition to each pair in our training set, a random combination of video and utterance and negative samples. The strength of this approach is the ability to generate a diverse set of distractors which is important given our relatively small dataset.

Given the small size of our networks and the little available data, we perform 100-fold cross-validation and report results for the full data set. We employ the Adam optimizer with a learning rate of 0.001 for model training over 5 epochs of the data.

[6] `https://huggingface.co/T-Systems-onsite/cross-en-de-roberta-sentence-transformer`

6 Results

For classification (i.e., language modelling the first word of the sentence), our attempts fail. Our best results yield entropies of around 7.56 bit, roughly 1.3 bit more than the unigram baseline of 6.25 bit. Results vary little between the three types of video representations used.

For discrimination trained with contrastive learning, we report the proportion of correctly classified examples as well as mean/stddev/median of the discriminator's output value for all three types of video representations in Table 1. We also report the p-values of a binomial test on the proportion of correct discriminations as well as for a one-sided t-test that assesses whether the discriminator's mean output significantly differs from 0.5 (i.e., chance).

We find that CLIP and Timesformer yield discriminations highly significantly above chance level with CLIP outperforming the Timesformer. OpenFace also produces more correct than incorrect results but this may be attributed to chance.

					p-values	
	Correct	**Mean**	**Stddev**	**Median**	**binomial**	**t-test**
OpenFace	0.51	0.52	0.47	0.57	.28	.12
CLIP	0.56	0.51	0.07	0.51	$<.002$	$<.0001$
Timesformer	0.53	0.51	0.08	0.51	$<.07$	$<.0001$

Table 1 – Comparing the experiment result statistics of OpenFace, CLIP, and Timesformer.

7 Discussion

We find that video during pauses is not predictive of the first word spoken after the pause, at least not in our straightforward classification approach. A particular issue here is the sparse vocabulary (most words occuring only once, i.e., either in training or test sets even with leave-one-out-training). We tried to counter this with including the correct unigram probabilities into the network. We believe that the network nevertheless 'unlearned' these probabilities during training and focused on noise in the data instead.

However, we find that our discrimination-based models are able to find pairs of pause videos and utterance texts at above-chance levels. This means that there is information in the video pause that correlates with the following text. We furthermore find that more elaborate DL-based image and video encoders are superior to feature extraction from OpenFace (although they are not trained specifically for facial features).

Our models' prediction performance is only very modestly above chance levels. However, this cannot be surprising as spoken interaction obviously transports most information via speech (or text) rather than via looking at someone before they speak.

Our approach also excludes the eventuality that the model 'sees' paralinguistic aspects of the speech (such as tempo or the like) during the pause, as we use only textual embeddings of the utterance, rather than speech audio.

We believe that our models indeed 'read the face' rather than merely the lips – if the latter were true, the first-word classification should likely have yielded better results. This indicates that we have not been training to read the first word from the lips of the speaker prior to talking, accidentally drawing from co-articulation caused by pre-vocal movements.

8 Conclusions

Our contributions are threefold. First, we expand upon the available annotation of speech in the Merkel Podcast Corpus [16]. Second, we propose an easy to replicate workflow for annotating data utilizing the potato annotation tool [22] that is aimed at investigating video content of speaker pauses[7]. Third, we present results from initial experiments to assess the information contained in video during inter-utterance and inter-turn pauses.

We analyze the information that can be derived from looking at the pauses before a person is speaking. We find that pausing behaviours can be predictive of the person's upcoming speech. Such information can be directly used to improve incremental understanding of user utterances. Furthermore, the correlation of pausing behaviour and the following speech also means that an intelligent virtual agent should act out their pausing behaviours so as to make them compatible to their own utterances.

In the future, we intend to broaden our experiments to multi-speaker scenarios and to include speech audio into the analysis. Baumann [10] has previously shown that audio interpretation of pausing behaviour reflects onto the speaker. We believe that the video channel will typically contain much more information in face-to-face (or video-conferencing) interaction and that analyzing the visual backchannel will be fruitful for systems developers as it becomes computationally feasible.

Acknowledgment: This work was supported in part by a Google Cloud research award.

References

[1] SHILLINGFORD, B., Y. ASSAEL, M. W. HOFFMAN, T. PAINE, C. HUGHES, U. PRABHU, H. LIAO, H. SAK, K. RAO, L. BENNETT, M. MULVILLE, B. COPPIN, B. LAURIE, A. SENIOR, and N. DE FREITAS: *Large-scale visual speech recognition*. 2018. 1807.05162.

[2] CHATZIAGAPI, A., D. SGOUROPOULOS, C. KAROUZOS, T. MELISTAS, T. GIANNAKOPOULOS, A. KATSAMANIS, and S. NARAYANAN: *Audio and ASR-based Filled Pause Detection*. In *2022 10th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pp. 1–7. IEEE, Nara, Japan, 2022. doi:10.1109/ACII55700.2022.9953889.

[3] KALIYEV, A., S. V. RYBIN, and Y. MATVEEV: *The Pausing Method Based on Brown Clustering and Word Embedding*. In A. KARPOV, R. POTAPOVA, and I. MPORAS (eds.), *Speech and Computer*, Lecture Notes in Computer Science, pp. 741–747. Springer International Publishing, Cham, 2017. doi:10.1007/978-3-319-66429-3_74.

[4] MATSUNAGA, Y., T. SAEKI, S. TAKAMICHI, and H. SARUWATARI: *Improving robustness of spontaneous speech synthesis with linguistic speech regularization and pseudo-filled-pause insertion*. In *12th Speech Synthesis Workshop (SSW) 2023*. 2023.

[5] DENDUKURI, S., P. CHITKARA, J. R. A. MONIZ, X. YANG, M. TSAGKIAS, and S. PULMAN: *Using Pause Information for More Accurate Entity Recognition*. In *Proceedings of the 3rd Workshop on Natural Language Processing for Conversational AI*, pp. 243–250. Association for Computational Linguistics, Online, 2021. doi:10.18653/v1/2021.nlp4convai-1.22.

[6] SCHLANGEN, D., T. BAUMANN, and M. ATTERER: *Incremental Reference Resolution: The Task, Metrics for Evaluation, and a Bayesian Filtering Model that is Sensitive to*

[7] https://github.com/christianschuler8989/PauseProcessing

Disfluencies. *Proceedings of SIGDIAL 2009:*, 10th(0th Annual Meeting of the Special Interest Group in Discourse and Dialogue), pp. 30–37, 2009.

[7] YANG, D., T. KORIYAMA, Y. SAITO, T. SAEKI, D. XIN, and H. SARUWATARI: *Duration-aware pause insertion using pre-trained language model for multi-speaker text-to-speech*. 2023. doi:10.48550/arXiv.2302.13652. 2302.13652.

[8] HOOGLAND, D., L. WHITE, and S. KNIGHT: *Speech Rate and Turn-Transition Pause Duration in Dutch and English Spontaneous Question-Answer Sequences*. *Languages*, 8(2), p. 115, 2023. doi:10.3390/languages8020115.

[9] SCHETTINO, L., M. D. MARO, and F. CUTUGNO: *Silent pauses as clarification trigger*. *Laughter and Other Non-Verbal Vocalisations Workshop: Proceedings (2020)*, 2020. doi:10.4119/lw2020-927.

[10] BAUMANN, T.: *How a Listener Influences the Speaker*. In *Proc. Speech Prosody 2020*, pp. 970–974. 2020. doi:10.21437/SpeechProsody.2020-198.

[11] KOUTSOMBOGERA, M. and C. VOGEL: *Speech Pause Patterns in Collaborative Dialogs*. In A. ESPOSITO, A. M. ESPOSITO, and L. C. JAIN (eds.), *Innovations in Big Data Mining and Embedded Knowledge*, Intelligent Systems Reference Library, pp. 99–115. Springer International Publishing, Cham, 2019. doi:10.1007/978-3-030-15939-9_6.

[12] MATZINGER, T., M. PLEYER, and P. ŻYWICZYŃSKI: *Pause Length and Differences in Cognitive State Attribution in Native and Non-Native Speakers*. *Languages*, 8(1), p. 26, 2023. doi:10.3390/languages8010026.

[13] LIU, J., F. FU, L. LI, J. YU, D. ZHONG, S. ZHU, Y. ZHOU, B. LIU, and J. LI: *Efficient Pause Extraction and Encode Strategy for Alzheimer's Disease Detection Using Only Acoustic Features from Spontaneous Speech*. *Brain Sciences*, 13(3), p. 477, 2023. doi:10.3390/brainsci13030477.

[14] GREDEN, J. F., A. A. ALBALA, I. A. SMOKLER, R. GARDNER, and B. J. CARROLL: *Speech pause time: A marker of psychomotor retardation among endogenous depressives*. *Biological Psychiatry*, 16(9), pp. 851–859, 1981.

[15] ADDLESEE, A., A. ESHGHI, and I. KONSTAS: *Current Challenges in Spoken Dialogue Systems and Why They Are Critical for Those Living with Dementia*. 2019. doi:10.48550/arXiv.1909.06644. 1909.06644.

[16] SAHA, D., S. NAYAK, and T. BAUMANN: *Merkel Podcast Corpus: A Multimodal Dataset Compiled from 16 Years of Angela Merkel's Weekly Video Podcasts*. In N. CALZOLARI, F. BÉCHET, P. BLACHE, K. CHOUKRI, C. CIERI, T. DECLERCK, S. GOGGI, H. ISAHARA, B. MAEGAARD, J. MARIANI, H. MAZO, J. ODIJK, and S. PIPERIDIS (eds.), *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pp. 2536–2540. European Language Resources Association, Marseille, France, 2022.

[17] BALTRUSAITIS, T., A. ZADEH, Y. C. LIM, and L.-P. MORENCY: *OpenFace 2.0: Facial Behavior Analysis Toolkit*. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pp. 59–66. 2018. doi:10.1109/FG.2018.00019.

[18] RADFORD, A., J. W. KIM, C. HALLACY, A. RAMESH, G. GOH, S. AGARWAL, G. SASTRY, A. ASKELL, P. MISHKIN, J. CLARK, G. KRUEGER, and I. SUTSKEVER: *Learning transferable visual models from natural language supervision*. 2021. 2103.00020.

[19] BERTASIUS, G., H. WANG, and L. TORRESANI: *Is space-time attention all you need for video understanding?* 2021. 2102.05095.

[20] VASWANI, A., N. M. SHAZEER, N. PARMAR, J. USZKOREIT, L. JONES, A. N. GOMEZ, L. KAISER, and I. POLOSUKHIN: *Attention is all you need*. In *Neural Information Processing Systems*. 2017. URL https://api.semanticscholar.org/CorpusID:13756489.

[21] REIMERS, N. and I. GUREVYCH: *Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation*. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 4512–4525. Association for Computational Linguistics, Online, 2020. doi:10.18653/v1/2020.emnlp-main.365.

[22] PEI, J., A. ANANTHASUBRAMANIAM, X. WANG, N. ZHOU, A. DEDELOUDIS, J. SARGENT, and D. JURGENS: *Potato: The portable text annotation tool*. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. 2022.