The importance of fillers for text representations of speech transcripts

Tanvi Dinkar[1]*; **Pierre Colombo**[1,2]*;
Matthieu Labeau[1], **Chloé Clavel**[1]
[1]LTCI, Telecom Paris, Institut Polytechnique de Paris, [2]IBM GBS France
[1]firstname.lastname@telecom-paris.fr

Abstract

While being an essential component of spoken language, fillers (e.g. "um" or "uh") often remain overlooked in Spoken Language Understanding (SLU) tasks. We explore the possibility of representing them with deep contextualised embeddings, showing improvements on modelling spoken language and two downstream tasks — predicting a speaker's stance and expressed confidence.

1 Introduction

Disfluencies are interruptions in the regular flow of speech, such as pausing silently, repeating words, or interrupting oneself to correct something said previously (Fraundorf et al., 2018). They commonly occur in spoken language, as spoken language is rarely fluent. *Fillers* are a type of disfluency that can be a sound ("um" or "uh") filling a pause in an utterance or conversation.

Recent work has shown that contextualised embeddings pre-trained on large written corpora can be fine-tuned on smaller spoken language corpora to learn structures of spoken language (Tran et al., 2019). However, for NLP tasks, fillers and all disfluencies are typically removed in pre-processing, as NLP models achieve highest accuracy on syntactically correct utterances. This contradicts linguistic studies, which show that fillers are an essential and informative part of spoken language (Clark and Fox Tree, 2002; Yoshida and Lickley, 2010; Brennan and Williams, 1995; Corley et al., 2007; Stolcke and Shriberg, 1996).

So far, the information carried by fillers has only been studied using hand crafted features, for example in Le Grezause (2017); Saini (2017); Dinkar et al. (2020). Besides, Barriere et al. (2017) show that pre-trained word embeddings such as Word2vec (Mikolov et al., 2013), have poor representation of spontaneous speech words such as "uh", as they are trained on written text and do not carry the same meaning as when used in speech. We address the matter of representing fillers with deep contextualised word representations (Devlin et al., 2019), and investigate their usefulness in NLP tasks for spoken language, without handcrafting features.

Hence, the present work is motivated by the following observations: (1) Fillers play an important role in spoken language. For example, a speaker can use fillers to inform the listener about the linguistic structure of their utterance, such as in their (difficulties of) selection of appropriate vocabulary while informing the listener about a pause in their upcoming speech stream (Clark and Fox Tree, 2002). (2) Fillers and prosodic cues have also been linked to a speaker's *Feeling of Knowing (FOK)* or *expressed confidence*, that is, a speaker's certainty or commitment to a statement (Smith and Clark, 1993). Brennan and Williams (1995) observed that fillers and prosodic cues contribute to the listener's perception of the speaker's expressed confidence in their utterance, which they refer to as the *Feeling of Another's Knowing (FOAK)*, also observed by (Wollermann et al., 2013). (3) Recent work has shown that fillers have been successful in *stance* prediction (stance referring to the subjective spoken attitude towards something) (Le Grezause, 2017).

Aim of this work: We want to verify that these observations are still valid when

*Equal contribution

Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pages 7985–7993,
November 16–20, 2020.

we represent fillers in an automatic and efficient way. Hence, our contributions are as follows: (1) Fillers contain useful information that can be leveraged by deep contextualised embeddings to better model spoken language and thus should not be removed. In addition, we study which filler representation strategies are best suited to our task of Spoken Language Modelling (SLM) and investigate the learnt positional distribution of fillers. (2) We show that in a spontaneous speech corpus of spoken monologues, fillers are a discriminative feature in predicting the perception of expressed confidence of the speaker, and perception of a speaker's stance (which we measure by sentiment).

2 Models and Data description

2.1 Model Description

For our work, we consider the two fillers "uh" and "um" (see subsection 2.2). To obtain contextualised word embeddings for fillers, we use bidirectional encoder representations from transformers (BERT) (Devlin et al., 2019), as it has achieved SOTA performance on several NLP benchmarks and are better than Word2Vec for word sense disambiguation by integrating context (Bartunov et al., 2015).

2.1.1 Spoken Language Modelling

For SLM, we use the masked language modelling objective (MLM). It consists of masking some words of the input tokens at random, and then predicting these masked tokens. The MLM objective is classically used to pretrain and then fine-tune BERT. Here, we use this MLM objective to fine-tune a pretrained BERT on a spoken language corpus (see subsection 2.2). Each experiment requires a token representation strategy $\mathcal{T}_i$ and a pre-processing strategy $\mathcal{P}_{Si}$ (additional details are given in the algorithm 1 in Supplementary).

The **token representation strategies** are particularly important for our task, for BERT to learn the distribution of fillers. The three token representation strategies $(\mathcal{T}_1, \mathcal{T}_2, \mathcal{T}_3)$, are described as follows: In $\mathcal{T}_1$, no special treatment is done to the fillers[1], i.e BERT will use its a priori knowledge of the fillers "uh" or "um" to model the language. In $\mathcal{T}_2$, "uh" and "um" are distinguished from other tokens by a special filler tag, and are represented as two different tokens respectively; this strategy aims at forcing BERT to learn a new embedding that focuses both on the position and the context of the fillers. In $\mathcal{T}_3$, both fillers are represented as the same token, suggesting that they have the same pragmatic meaning and are interchangeable. A concrete example is given in Table 1.

Pre-processing strategies, $(\mathcal{P}_{S1}, \mathcal{P}_{S2}, \mathcal{P}_{S3})$, are as follows: In $\mathcal{P}_{S1}$, the sentences have all fillers removed, both during training and inference. In $\mathcal{P}_{S2}$, the sentences have the fillers kept during training, but are removed at inference. In $\mathcal{P}_{S3}$, the fillers are kept both during training and inference. For each pre-processing and token representation strategy, we optionally fine-tune BERT using the same Masked Language Model (MLM) objective as in the original paper (Devlin et al., 2019). Note, if we do not fine-tune, the training dataset ($\mathcal{D}_{train}$) is not used and therefore $\mathcal{P}_{S1}$ and $\mathcal{P}_{S2}$ are equivalent. For language modelling we report the perplexity (ppl) measure to evaluate the quality of the model.

2.1.2 Confidence and Sentiment Prediction

In both our confidence prediction and sentiment analysis task, our goal is to predict a label of confidence/sentiment using our BERT text representations that include fillers. Formally, our confidence/sentiment predictor is obtained by adding a Multi-Layer Perceptron (MLP) on top of a BERT, which has been optionally fine-tuned using the MLM. The MLP is trained by minimising the mean squared error (MSE) loss (additional details are given in algorithm 2 in Supplementary). We keep the same token representation and pre-processing strategies from Section 2.1.1.

2.2 Data Description

We use the Persuasive Opinion Mining (POM) dataset (Park et al., 2014), a dataset of 1000 English monologue videos. Speakers recorded themselves giving a movie review, freely available on ExpoTV.com. The movies were rated from 1 star (most negative) to 5 stars (most

[1]It is interesting to note that BERT provides embedding for "uh" or "um" despite being trained on written text (Wikipedia, BooksCorpus (Zhu et al., 2015), Word Benchmark (Chelba et al., 2014).

Token.	Output Tokenizer
Raw	(umm) Things that (uhh) you usually wouldn't find funny were in this movie.
$\mathcal{T}_1$	['umm', 'things', 'that', 'uh', 'you', 'usually', 'wouldn', "'", 't', 'find', 'funny', 'were', 'in', 'this', 'movie', '.']
$\mathcal{T}_2$	['$FILLER_{UMM}$]', 'things', 'that', '[$FILLER_{UHH}$]', 'you', 'usually', 'wouldn', "'", 't', 'find', 'funny', 'were', 'in', 'this', 'movie', '.']
$\mathcal{T}_3$	['[FILLER]', 'things', 'that', '[FILLER]', 'you', 'usually', 'wouldn', "'", 't', 'find', 'funny', 'were', 'in', 'this', 'movie', '.']

Table 1: Filler representation using different token representation strategies

positive). Annotators were asked to label the video for high-level attributes. For confidence, annotators (3 per video) were asked "How confident was the reviewer?", and had to each give a label respectively; from 1 (not confident) to 7 (very confident), after watching the entire review. Similarly for sentiment, the annotators were asked "How would you rate the sentiment expressed by the reviewer towards this movie?", and were asked to give a label from 1 (strongly negative) to 7 (strongly positive).

We choose this dataset for the following reasons: (1) The corpus has been manually transcribed with fillers "uh" and "um", where $\approx 4\%$ of the speech consists of fillers (for comparison, the Switchboard (Godfrey et al., 1992) dataset of human-human dialogues, consists of $\approx$ 1.6% of fillers (Shriberg, 2001)). Sentence markers have been manually transcribed, with the practice of the filler being annotated sentence-initially, if the filler occurs between sentences. (2) The dataset consists of monologues, where the speaker is conscious of an *unseen* listener, but dialogue-related disfluencies (such as backchannels) are not present, allowing us to concentrate on fillers of the narratives of the speaker (Swerts, 1998). (3) Only reviews with a 1-2 star or a 5 star rating were chosen for annotation, to clearly demarcate sentiment/stance polarity. (4) *FOAK*, which we measure by the given label of confidence, has been annotated with high inter-annotator agreement (Krippendorff's alpha = 0.73).

Details can be found in the supplementary material and in Park et al. (2014). Confidence labels are obtained by taking the root mean square (RMS) value of the labels given by the 3 annotators[2]. Sentiment labels are calculated by taking the mean of the 3 labels, which were obtained from Zadeh (2018a)[3].

3 Experiments and Analysis

3.1 Information contained by fillers can be leveraged to model spoken language.

Language Modelling with fillers. We compare the perplexity of the LM with different pre-processing strategies with a fixed token representation $\mathcal{T}_1$. Results are reported in Table 2(a). We compare $\mathcal{P}_{S1}$,$\mathcal{P}_{S2}$ $\mathcal{P}_{S3}$ with or without fine-tuning and observe that adding fillers, both during training and inference, leads to a model with lower perplexity and a perplexity reduction of at least 10%. Hence, fillers contain information that can be leveraged by BERT.

As shown, the fine-tuning procedure reduces the perplexity of the language model. Even without fine-tuning, we observe that $\mathcal{P}_{S3}$ outperforms $\mathcal{P}_{S1}/\mathcal{P}_{S2}$, as the perplexity reduces when adding fillers. This suggests that BERT has a priori knowledge of spoken language, in terms of fillers.

Hence, fillers can be leveraged to reduce uncertainty of BERT for SLM. This is not an expected result, as intuitively, one might think that the perplexity would reduce when fillers are excluded from both training and inference, due to the fact that the utterance is shorter and "simplified". The fact that $\mathcal{P}_{S3}$ outperforms the other pre-processing methods also suggests that the MLM procedure is an effective way to learn this information.

Best Token representation: We observe that $\mathcal{T}_1$ outperforms the other representations in a fine-tuning setting, as shown in Table 2(b). Given the restricted size of our data and the dimension of the BERT embeddings (768), it is better to keep the existing representations (with $\mathcal{T}_1$), than adding and learning new representations from scratch.

[2]Though the inter-annotator agreement for confidence is high, we choose RMS as a way to handle disagreement between annotators. For example, annotation labels $\{3, 5, 7\}$ would result in mean value of 5, not highlighting that one annotator found the reviewer particularly confident. The RMS value however (≈ 5.3), slightly enhances the high confidence label.

[3]A toolkit for multimodal analysis. Please refer to the **Usage** and the **Supported Datasets** sections, which include instructions to download the data.

Fine.	Setting	Token.	Ppl
w/o	$\mathcal{P}_{S1}$	$\mathcal{T}_1$	22
	$\mathcal{P}_{S2}$	$\mathcal{T}_1$	22
	$\mathcal{P}_{S3}$	$\mathcal{T}_1$	**20**
w	$\mathcal{P}_{S1}$	$\mathcal{T}_1$	5.5
	$\mathcal{P}_{S2}$	$\mathcal{T}_1$	5.6
	$\mathcal{P}_{S3}$	$\mathcal{T}_1$	**4.6**

(a)

Setting	Token.	Ppl
$\mathcal{P}_{S3}$	$\mathcal{T}_1$	**4.6**
$\mathcal{P}_{S3}$	$\mathcal{T}_2$	4.7
$\mathcal{P}_{S3}$	$\mathcal{T}_3$	4.7

(b)

Fine.	Model	FOAK	Sent
w/o	$\mathcal{P}_{S1}$	1.47	1.98
	$\mathcal{P}_{S2}$	1.45	1.75
	$\mathcal{P}_{S3}$	**1.30**	**1.44**
w	$\mathcal{P}_{S1}$	1.32	1.39
	$\mathcal{P}_{S2}$	1.31	1.40
	$\mathcal{P}_{S3}$	**1.24**	**1.22**

(c)

Table 2: From left to right, the (a) LM Task, (b) Best token representation, (c) MSE of Confidence (FOAK) and the Sentiment (Sent) prediction task. Wilcoxon test (10 runs with different seeds) has been performed. Highlighted results exhibit significant differences (p-value < 0.005). Data split is fixed according to Zadeh (2018b) and results are given on the test set (see supplementary materials for for additional details).

Interestingly, $\mathcal{T}_2$ and $\mathcal{T}_3$ perform the same. This can be explained by "um" and "uh" being only distinguished in duration (Clark and Fox Tree, 2002), the hypothesis being that "uh" is used for a shorter pause in speech; which cannot not be reflected in text. Given these results, we fix $\mathcal{T}_1$ as the token representation strategy for the rest of the experiments.

Learnt Positional distribution of fillers: We additionally test whether our model has learnt information about the placement of fillers. We use fine-tuned BERT on $\mathcal{D}_{train}$ with fillers to see where the model estimates the most probable position of the fillers (which we call $\mathcal{LM}_{fillers}$) to be. Given a sentence S of length L, we insert after word j the mask token ('[MASK]') to obtain the corrupted sentence $\widetilde{S}$[4]. We compute the probability of the appearance of a filler in position $j+1$ according to the LM, which corresponds to $P([MASK] = filler|\widetilde{S})$, as illustrated by Figure 1. Formally, we plot the average of the probability of the masked word to be a filler given its position in the sentence, as shown in Figure 2. We observe that the fine-tuned BERT on $\mathcal{D}_{train}$ with fillers ($\mathcal{LM}_{fillers}$) predicts with high probability fillers occurring at the first position in the sentence (please refer to Table 5 supplementary for example sentences). This is consistent with the actual distribution of fillers in the dataset, as can be seen in Figure 2. The fine-tuned BERT on $\mathcal{D}_{train}$ without fillers ($\mathcal{LM}_{nofillers}$) predicts a constant low probability. Given the available segmentation of sentence boundaries (fine-grained discourse annotations are not available), it is interesting to note that our model was able to capture similar positional distribution of fillers that are reported in Swerts (1998); Shriberg (2001); Swerts and Geluykens (1994); Yoshida and Lickley (2010).

In this section we show that although BERT uses contextualised word embeddings, the information contained in fillers can be leveraged to achieve a better modelling of spoken language.

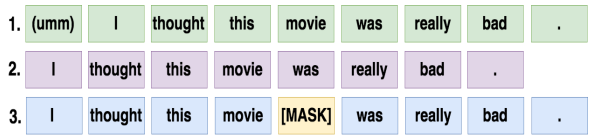

Figure 1: Predicting the probability of a filler, where 1. Raw input, 2. Pre-processed text with the filler removed, and 3. Illustrates the [MASK] procedure for predicting the probability of a filler at position 5.

3.2 Fillers are a discriminative feature for *FOAK* and *stance* prediction.

We observe the impact that fillers have on two downstream tasks, a novel FOAK prediction task, and a ubiquitous sentiment analysis task. Psycholinguistic studies have observed the link between fillers and expressed confidence (Smith and Clark, 1993; Brennan and Williams, 1995; Wollermann et al., 2013). Previous research on the link between fillers and their relation to a speaker's expressed confidence has been confined to a narrow range of QA tasks (Schrank and Schuppler, 2015). Fillers have also been linked to stance prediction (Le Grezause, 2017),

[4]For clarity we abuse the notation and remove dependence in j.

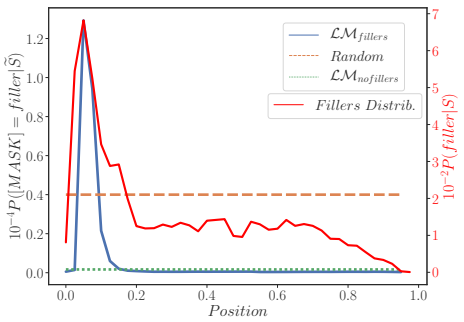

Figure 2: Predicting the position of fillers. *Fillers Distrib.* stands for the actual filler distribution in the dataset. *Random* stands for the random predictor which predicts $P([MASK] = filler|\widetilde{S}) = \frac{2}{|\mathcal{V}|}$ where $|\mathcal{V}|$ is the size of the vocabulary, and 2 represents both fillers.

which we measure using sentiment. We show that in a spontaneous speech corpus of spoken monologues, fillers can play a role in predicting both the perception of the speaker's expressed confidence and speaker's stance.

In Table 2(c) we observe that both with and without fine-tuning the $\mathcal{P}_{S3}$ decreases the MSE compared to $\mathcal{P}_{S1}$ and $\mathcal{P}_{S2}$. $\mathcal{P}_{S1}$ and $\mathcal{P}_{S2}$ have similar MSE because fillers are not added during the inference phase. We observe that $\mathcal{P}_{S2}$ leads to higher MSE, possibly because of the discrepancy created between $\mathcal{D}^{labelled}_{train}$ and $\mathcal{D}^{labelled}_{test}$. This shows that fillers can be a discriminative feature in both FOAK and stance (Le Grezause, 2017) prediction, apart from overt lexical cues [5].

Does the addition of fillers always improve the results for downstream spoken language tasks? In the subsection 3.1, we show that by including fillers , the MLM achieves a lower perplexity. An assumption one could make based on the work by Radford et al. (2019), is that with this model, the results for any further downstream task would be improved by the presence of fillers. However, we observe that to predict the persuasiveness of the speaker (using the high level attribute of persuasiveness annotated in the dataset (Park et al., 2014)), following the same procedure as outlined in subsubsection 2.1.2, that fillers, in fact, are not a discriminative feature.

[5] by overt lexical cues, we mean words that explicitly express uncertainty/confidence, such as *maybe, I'm unsure* or sentiment, *amazing, disgusting*)

4 Conclusion

When working with deep contextualised representations of transcribed spoken language, we showed that retaining fillers can improve results, both when modelling language and on a downstream task (FOAK and stance prediction). Besides, we propose and compare various token representation and pre-processing strategies in order to integrate fillers. We plan to extend these results by studying the mixing of such textual filler-oriented representations with acoustic representations, and further investigate the representation of fillers learnt during pre-training.

Acknowledgements

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 765955 and the French National Research Agency's grant ANR-17-MAOI.

References

Valentin Barriere, Chloé Clavel, and Slim Essid. 2017. Opinion Dynamics Modeling for Movie Review Transcripts Classification with Hidden Conditional Random Fields. In *Proceedings of Interspeech 2017*, Stockholm, Sweden.

Sergey Bartunov, Dmitry Kondrashkin, Anton Osokin, and Dmitry P. Vetrov. 2015. Breaking Sticks and Ambiguities with Adaptive Skip-Gram. *CoRR*, abs/1502.07257.

Susan E. Brennan and Maurice Williams. 1995. The Feeling of Another's Knowing: Prosody and Filled Pauses as Cues to Listeners about the Metacognitive States of Speakers. *Journal of Memory and Language*, 34(3):383 – 398.

Ciprian Chelba, Tomas Mikolov, M. Schuster, Qi Ge, T. Brants, Phillipp Koehn, and T. Robinson. 2014. One Billion Word Benchmark for Measuring Progress in Statistical Language Modeling. In *Proceedings of Interspeech 2014*.

Herbert H. Clark and Jean E. Fox Tree. 2002. Using uh and um in Spontaneous Speaking. *Cognition*, 84(1):73 – 111.

Pierre Colombo, Emile Chapuis, Matteo Manica, Emmanuel Vignon, Giovanna Varni, and Chloe Clavel. 2020. Guiding Attention in Sequence-to-Sequence Models for Dialogue Act Prediction. In *AAAI-20*, pages 7594–7601.

Pierre Colombo, Wojciech Witon, Ashutosh Modi, James Kennedy, and Mubbasir Kapadia. 2019. Affect-Driven Dialog Generation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3734–3743, Minneapolis, Minnesota. Association for Computational Linguistics.

Martin Corley, Lucy J. MacGregor, and David I. Donaldson. 2007. It's the Way that You, er, Say it: Hesitations in Speech Affect Language Comprehension. *Cognition*, 105(3):658 – 668.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Tanvi Dinkar, Ioana Vasilescu, Catherine Pelachaud, and Chloé Clavel. 2020. How Confident are You? Exploring the Role of Fillers in the Automatic Prediction of a Speaker's Confidence. In *Proceedings of the 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8104–8108. IEEE.

Scott H. Fraundorf, Jennifer Arnold, and Valerie J. Langlois. 2018. Disfluency.

John J Godfrey, Edward C Holliman, and Jane McDaniel. 1992. SWITCHBOARD: Telephone Speech Corpus for Research and Development. In *Proceedings of the 1992 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 1, pages 517–520. IEEE.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. *CoRR*, abs/1412.6980.

Taku Kudo and John Richardson. 2018. SentencePiece: A Simple and Language Independent Subword Tokenizer and Detokenizer for Neural Text Processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations (EMNLP)*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Esther Le Grezause. 2017. *Um and uh, and the Expression of Stance in Conversational Speech*. Theses, Université Sorbonne Paris Cité ; University of Washington.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep Learning. *Nature*, 521(7553):436.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. *arXiv preprint arXiv:1301.3781*.

Sunghyun Park, Han Suk Shim, Moitreya Chatterjee, Kenji Sagae, and Louis-Philippe Morency. 2014. Computational Analysis of Persuasiveness in Social Multimedia: A Novel Dataset and Multimodal Prediction Approach. In *Proceedings of the 16th International Conference on Multimodal Interaction*, ICMI 2014, page 50–57, New York, NY, USA. Association for Computing Machinery.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners. *OpenAI Blog*, 1(8).

Divya Saini. 2017. The Effect of Speech Disfluencies on Turn-Taking.

Tobias Schrank and Barbara Schuppler. 2015. Automatic Detection of Uncertainty in Spontaneous German Dialogue. In *Sixteenth Annual Conference of the International Speech Communication Association*.

Elizabeth Shriberg. 2001. To 'errrr' is Human: Ecology and Acoustics of Speech Disfluencies. *Journal of the International Phonetic Association*, 31(1):153–169.

Vicki L. Smith and Herbert H. Clark. 1993. On the Course of Answering Questions. *Journal of Memory and Language*, 32(1):25 – 38.

Andreas Stolcke and Elizabeth Shriberg. 1996. Statistical Language Modeling for Speech Disfluencies. In *Proceedings of the 1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings (ICASSP)*, volume 1, pages 405–408. IEEE.

Marc Swerts. 1998. Filled Pauses as Markers of Discourse Structure. *Journal of Pragmatics*, 30(4):485 – 496.

Marc Swerts and Ronald Geluykens. 1994. Prosody as a Marker of Information Flow in Spoken Discourse. *Language and Speech*, 37(1):21–43.

Trang Tran, Jiahong Yuan, Yang Liu, and Mari Ostendorf. 2019. On the Role of Style in Parsing Speech with Neural Models. In *Proceedings of Interspeech 2019*, pages 4190–4194.

Wojciech Witon, Pierre Colombo, Ashutosh Modi, and Mubbasir Kapadia. 2018. Disney at IEST 2018: Predicting Emotions Using an Ensemble. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 248–253, Brussels, Belgium. Association for Computational Linguistics.

Charlotte Wollermann, Eva Lasarcyk, Ulrich Schade, and Bernhard Schröder. 2013. Disfluencies and Uncertainty Perception-Evidence from a Human-Machine Scenario. In *Sixth Workshop on Disfluency in Spontaneous Speech (DISS)*.

Etsuko Yoshida and Robin J Lickley. 2010. Disfluency Patterns in Dialogue Processing. In *DiSS-LPSS Joint Workshop 2010*.

Amir Zadeh. 2018a. CMU-Multimodal SDK. https://github.com/A2Zadeh/CMU-MultimodalSDK.

Amir Zadeh. 2018b. CMU-Multimodal SDK, Standard data folds. https://github.com/A2Zadeh/CMU-MultimodalSDK/blob/master/mmsdk/mmdatasdk/dataset/standard_datasets/POM/pom_std_folds.py.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning Books and Movies: Towards Story-Like Visual Explanations by Watching Movies and Reading Books. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 19–27.

Supplementary

Model

Detailed algorithms: In algorithm 1 and algorithm 2, we provide additional details of the procedure used for the language modelling task and confidence prediction task. For stance prediction, the procedure is the same as for confidence.

Algorithm 1: Spoken Language Modelling

Input : $\mathcal{P}_{Si}$, $\mathcal{T}_i$, Pret. BERT $\mathcal{LM}$
Output : $(\mathcal{LM}, Perplexity)$

$(\mathcal{D}_{train}, \mathcal{D}_{dev}, \mathcal{D}_{test}) \leftarrow$ (train, dev, test set) according to $(\mathcal{P}_{Si}, \mathcal{T}_i)$
if *Do Finetuning* **then**
  $\mathcal{LM} \leftarrow \mathcal{LM}(\mathcal{D}_{train})$ using (MLM).
end
Evaluate: $Perplexity \leftarrow \mathcal{LM}$ on $\mathcal{D}_{test}$

Algorithm 2: Confidence prediction

Input : $\mathcal{P}_{Si}$, $\mathcal{T}_i$, $\mathcal{LM}$ from algorithm 1
Output : $(CONF_p, MSE)$

$(\mathcal{D}^{labelled}_{train}, \mathcal{D}^{labelled}_{dev}, \mathcal{D}^{labelled}_{test}) \leftarrow$ (train, dev, test set) according to $(\mathcal{P}_{Si}, \mathcal{T}_i)$
$CONF_p \leftarrow \mathcal{LM} + MLP$
$CONF_p \leftarrow CONF_p(\mathcal{D}^{labelled}_{train})$ using (MSE).
Evaluate: $MSE \leftarrow CONF_p$ on $\mathcal{D}_{test}$

Example of token representation strategies: Our token representation strategies are built on the tokenizer introduced by Devlin et al. (2019) and used the Sentence Piece algorithm (Kudo and Richardson, 2018). An example is given in Table 3.

Dataset: Additional details

We highlight relevant information about the dataset in Table 4. The count of each "uh" and "um" filler is roughly the same. After discarding some videos due to missing labels, only 100 of them do not contain fillers. We use the original standard training, testing and validation folds provided in the CMU-Multimodal SDK (Zadeh, 2018b).

The process of transcription of fillers is described in (Park et al., 2014). The transcriptions were carried out via Amazon Mechanical Turk, using 18 native English speaking workers based in the United States. These workers were from the same pool of workers used to annotate the videos for high level attributes. Each transcription was then reviewed and edited by in-house experienced transcribers for accuracy.

In Table 5 we give example sentences extracted from the POM dataset. In these examples, we can observe that the fillers are commonly located sentence-initially. Note, the corpus annotates "uh" and "um" as "uhh" and "umm" respectively, reflected in our examples taken from the dataset.

Hyper-parameters for our experiments

All the hyper-parameters have been optimised on the validation set based on the minimum of the training loss (MSE for confidence/sentiment prediction and perplexity for LM) accuracy computed on the last tag of the sequence. We used Adam optimizer (Kingma and Ba, 2015) with a learning rate of $10^{-}5$, which is updated using a polynomial decay. The gradient norm is clipped to 5.0, weight decay is set to 10^{-6}, and dropout (LeCun et al., 2015) is set to 0.2. Models have been implemented in PyTorch and trained on a v100 using the same procedure as in (Colombo et al., 2019, 2020; Witon et al., 2018).

Token.	Output Tokenizer
Raw	(umm) It's an interesting movie to say the least.
T1	['umm', 'it', "'", 's', 'an', 'interesting', 'movie', 'to', 'say', 'the', 'least', '.']
T2	['$[FILLER_{UMM}]$', 'it', "'", 's', 'an', 'interesting', 'movie', 'to', 'say', 'the', 'least', '.']
T3	['[FILLER]', 'it', "'", 's', 'an', 'interesting', 'movie', 'to', 'say', 'the', 'least', '.']

Table 3: Additional example of the different token representation strategies

Description	**Value**
Videos that contain fillers	**792**
Total *um* fillers in the corpus	4969
Total *uh* fillers in the corpus	4967
Total fillers in the corpus	**9936**
Number of tokens in the corpus	230462
% of tokens that are fillers	**4.31**
Average length (in tokens) of a video	255.9

Table 4: Details about the POM dataset

Samples
(umm) the title actually translates to The Brotherhood of War.
(umm) The movie itself is a lot like Saving Private Ryan and Band of Brothers.
(uhh) Morgan Freeman is great in this movie, and **(uhh)** so is Tim Robbins.
(umm) You'll only like it if you're into kid of strange, bizarre humor.
It's just **(uhh)** pretty obvious stuff you know.
But **(umm)** a lot of the movie didn't really make sense.
(umm) It's really funny, there there's (stutter) some really funny parts in it.
(umm) But, I recommend watching this movie it's really good.
(umm) The acting is only so-so.
And so **(umm)** I wouldn't really recommend it.
(umm) Yeah, but that's it.

Table 5: Some samples from the dataset. As can be seen, many of the fillers occur sentence-initially.