How Confident Are You? Exploring The Role Of Fillers In The Automatic Prediction Of A Speaker's Confidence

Tanvi Dinkar, Ioana Vasilescu, Catherine Pelachaud, Chloé Clavel

HOW CONFIDENT ARE YOU? EXPLORING THE ROLE OF FILLERS IN THE AUTOMATIC PREDICTION OF A SPEAKER'S CONFIDENCE

Tanvi Dinkar[1] *, Ioana Vasilescu*[2]*, Catherine Pelachaud*[3]*, Chloé Clavel*[1]

1. Institut Mines-Telecom, Telecom Paris, CNRS-LTCI, Paris, France
2. LIMSI, CNRS, Université Paris-Saclay, Orsay, France
3. CNRS-ISIR, UPMC, Paris, France

(tanvi.dinkar,chloe.clavel)@telecom-paristech.fr, Ioana.Vasilescu@limsi.fr, catherine.pelachaud@upmc.fr

ABSTRACT

"Fillers", example "um" in English, have been linked to the "Feeling of Another's Knowing (FOAK)" or the listener's perception of a speaker's expressed confidence. Yet, in Spoken Language Processing (SLP) they remain unexplored, or overlooked as noise. We introduce a new and challenging task, that is the prediction of FOAK, which we think has widespread applicability, given the increasing popularity of automatic processing of educational and job interviews, reviews and speeches. We design a set of filler features based on linguistic literature, and investigate their potential in FOAK prediction. We show that the integration of information related to implicature meanings allows an improvement in the FOAK model and that the different functions of fillers are differently correlated with confidence.

Index Terms— Spoken Language Processing, Disfluency, Fillers, Confidence, Educational Applications

1. INTRODUCTION

This paper presents a study which is a part of ANIMATAS (Advancing intuitive human-machine interaction with human-like social capabilities for education in schools), an H2020 Marie Sklodowska Curie European Training Network that aims to enhance the pedagogical environment through human-agent interaction. In this project, we investigate the role of spoken language in this educational context [1].

An essential part of spoken language are disfluencies; which can be defined as breaks, irregularities or non-lexical vocables that occur within the flow of otherwise fluent speech. They are frequent in spoken language, as spoken language is rarely fluent. In the past few years, there has been a widespread interest in SLP. However, methodologies in SLP to study the disfluencies of the speakers and the information they can provide, are often overlooked. *Fillers*, are a type of disfluency that can be a sound (*um* or *uh* in English) or word/phrase (*well*, *you know*) filling a pause in an utterance or conversation. Research on fillers conclude that they are informative in understanding spoken language [2, 3, 4, 5]. There are several roles that fillers play within the verbal message. The speaker can use a filler to indicate a pause in speech [2] or hesitation [6]. A speaker can use fillers to inform about the linguistic structure of their utterance, such as in their (difficulties of) selection of appropriate vocabulary while maintaining their turn (in dialogue). Importantly, fillers are linked to the metacognitive state of the speaker. It was observed that fillers and prosodic cues are linked to a speaker's *Feeling of Knowing (FOK)* or *expressed confidence*, that is, a speaker's certainty or commitment to a statement [7].

However, the meanings of fillers are contextual, and dependent on the perception of the listener [8, 2]. The speaker encodes the meaning into their speech, while the listener decodes (or perceives) this meaning depending on the context. Hence, studies have also looked at the *comprehension* of fillers, that is, by taking into account the listener's understanding of speech uttered by the speaker [9]. [4] found that listeners can *perceive* a speaker's metacognitive state, by using fillers and prosody as cues to study the *Feeling of Another's Knowing (FOAK)*; or the listener's perception of a speaker's expressed confidence/ commitment to their speech.

Despite the rich literature regarding fillers, from an SLP perspective, fillers remain mostly unexplored, or overlooked as noise. The aim of this paper is to do a feature study of fillers, to see whether the functions of fillers, can contribute to the prediction of perception of expressed confidence, or FOAK. We think that this task of FOAK prediction, is important and has widespread applicability, given the increasing popularity of automatic processing of educational and job interviews, reviews and speeches. We approach this task from the listener's perspective, as ultimately, the listener's perspective in these contexts determines the outcome of the speech. The research questions are as follows: **RQ1.** Are the different functions of fillers correlated differently to the listener's FOAK? **RQ2.** Can fillers be informative in the prediction of the listener's FOAK?

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 765955

The research on the FOK has focused on the link to prosody [10, 7, 4], fillers [7, 4], facial expressions and gestures [11], and overt lexical cues [12]. By overt lexical cues, we mean words that explicitly mark uncertainty/certainty, such as (*I'm unsure, definitely...*). Similar to our task, are studies that automatically predict points of uncertainty in speech [13, 14].

In affective computing, fillers (such as filler count) are commonly used as an attribute to study persuasiveness [15] and big 5 personality traits [16]. The Computational Paralinguistics Challenge 2013, focused on sensing fillers, which they considered to be a *social signal* [17]. Research in personality computing has the most consistent correlation, with observations made from speech; including paralanguage, such as fillers [18, 19]. To our knowledge, only a few studies use fillers as the *focal* feature in machine learning tasks (unless in disfluency detection); fillers are found to be successful in *stance* prediction (stance referring to the subjective spoken attitudes towards something [20]) [21], *turn-taking* prediction [22], and to enhance the *naturalness* of the text-to-speech systems [23].

FOK studies are usually statistical; there has not been a vast interest in the automatic prediction of confidence. These studies are limited to a narrow range of question-answering (QA) tasks [13]. The speaker is typically asked to give an answer (usually the length of a sentence) to a question with a filler inserted, based on a script. The listener is then asked to form a perception based on this answer, and may or may not notice the use of filler due to the short answer length. However, in natural conversation, listener's are not aware of the use of fillers, unless overused or used in the wrong context [24]. We propose to study free speech in a monologue, which is significantly longer, and has naturally occurring fillers. The annotations that the listener gives are impressions created after hearing the entire monologue (details in Section 3).

This highlights why our task differs from detecting points of uncertainty in speech [13, 14]; as uncertainty detection typically obtains fine grained uncertainty labels based on overt lexical cues, and not the listener's *overall* impression. Points of uncertainty in speech, could still lead to a general impression of confidence of the speaker.

The applications of filler features in affective computing are common, but there aren't studies that do a focused analysis of them. Certain studies also predetermine based on the context (such as in job interviews, where fluency is desirable) whether the impact of fillers is positive or negative [25], which may not always be the case. When fillers are used as features, typically their count is taken [25], or n-gram features [21], but richer representations of their roles are required. Based on state of the art linguistic literature, we propose a novel way to *represent* our filler features, as outlined in Section 2.

The rest of the paper is organised as follows: In Section 2, we describe how we design our filler features based on linguistic literature. In Section 3, we discuss the dataset used for our experiments. Section 4 outlines the methodology, the experiments and the results of the research questions. Section 5 discusses the conclusion of the experiments.

2. DESIGNING FILLER FEATURES

In the ***filler-as-word*** **hypothesis**, a filler functions as an interjection, therefore its meaning is highly dependent on the context. Fillers are hence distinguished by their *basic meaning* and their implied meaning, or *implicature* [2]. We use this hypothesis of *basic and implicature* interjections as a basis for our feature representation. The **basic meaning** of a filler, is to announce the initiation, at $t(filler)$ by the speaker, of what is expected to be a delay in speech [2]. Although fillers have one basic meaning; they have several **implicatures**. A filler can have an *unable to proceed* implicature, indicating hesitation, nervousness, or uncertainty of the speaker [6]. A filler can be used by the speaker in a *syntactic marking* implicature, to indicate pausing to (re)formulate thoughts at discourse boundaries. Fillers can have an implicature linked to a speaker's *stance*, and *stance polarity* [21]. Fillers are associated with the cognitive load of the speaker [26]; disfluency rates increase for longer sentences, suggesting an increase in cognitive load.

We thus design two sets of filler features; a *basic set* and an *implicature set*. An example of the extracted features is given in Figure 1. The basic set includes three features: *f_num, f_uh, f_um*, corresponding to the total number of fillers, the total number of filler *uh* and the total number of filler *um*, respectively. The implicature set includes the following features:

- *f_uncertain*: the number of fillers present in sentences that contain stutter markings (fillers to denote an *unable to proceed implicature*).
- *f_stance*: the number of fillers occurring in sentences that contain tokens (*token* refers to both fillers and words) that are very positive, positive, negative and very negative (to denote fillers associated with *stance*). Token-level stance and polarity labels for the dataset are taken from [27].
- *f_start, f_mid*: the number of fillers that occur at the start of the sentence and within the sentence respectively (fillers to denote *syntactic marking*).
- *f_cog, sen_len*: the length of the sentence in tokens (sen_len) and the position of the *f_mid* in the sentence (as an indication of *Cognitive load*).
- *rate*: the rate of speaking that is computed by dividing the number of tokens by the duration of the video (as a way to measure deletions vs. repetitions : acoustic analysis [28] shows that speakers that use deletions "it was very *uh* she liked it", have a faster speaking rate than speakers that use repetitions "it's *stutter um* it gets hilarious").

3. DATASET

We use the Persuasive Opinion Mining (POM) dataset [15], an English multimedia corpus of 1000 movie review videos.

Video$_1$

Today I'll be reviewing Saving Private Ryan DVD. (umm) Saving Private Ryan is a very good movie starring Tom Hanks. f_start
It's about four brothers all with the last name Ryan who are in World War Two, and three of them die and the last
one has to be saved because of (uhh) U.S. law. f_cog
...
And (umm) The very good movie teaches you a lot about history and about the soldiers and what they went through during f_stance
World War Two. It's (uhh) one of my favorite movies and (umm) it's won five Academy Awards... f_mid

Video$_2$

I think there's been great reviews (stutter) about this movie and it's very (umm) fun to watch with friends and I f_uncertain
have always had a fun time watching it.

Fig. 1. Example annotation scheme for the fillers. Video transcripts are taken from the corpus.

Description	Value
Videos that contain fillers	**792**
Total number of videos used	892
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 1. Details about the POM dataset.

Speakers record videos of themselves giving a movie review. After watching the review, annotators are asked to label the video for high-level attributes, such as confidence. Annotators (3 per video) were asked "How confident was the reviewer?", and had to give a label based on a Likert scale; from 1(Not Confident) to 7(Very Confident), at the granularity of the video level. The inter-annotator score for confidence is high, (Krippendorff's alpha : 0.73) [15]. Further details can be found in [15].

The corpus has been manually transcribed to include the two fillers *uh* and *um*. *stutter* markings have been annotated, although the annotation context varies, and is dependent on the perception of the transcriber. We highlight relevant information about the dataset in Table 1. The occurrences of the fillers are not sparse, as seen in Table 1, where $\approx 4\%$ of the speech consists of fillers. This is relatively high; the Switchboard [29] dataset of human-human free speech for example, has $\approx 1.6\%$ of tokens that are fillers [26]. The count of each *uh* and *um* filler is roughly the same. Out of the videos used, only 100 of them do not contain fillers.

We take the Root Mean Squared (RMS) value for the 3 annotations per video as the final label, to reflect higher annotation scores. We remove the labels in the 1-2 range, due to sparsity of these labels.

4. EXPERIMENTS AND RESULTS

The contextual information provided by fillers, may contain information that is essential for predicting the listener's FOAK of the speaker. We incorporate such information into our

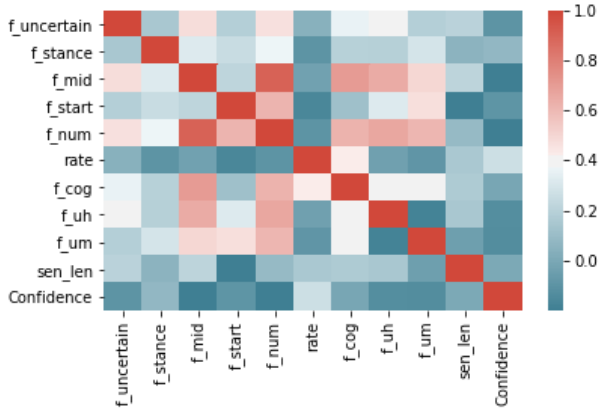

Fig. 2. Feature correlation of all features and Confidence using a Pearson's correlation coefficient. Confidence is given in the last column.

experiments through feature representation that is based on state of the art linguistic literature [2, 21, 6, 26, 7, 4], statistical analysis and linear machine learning models.

We use the transcripts of the dataset, and duration information as our input. We decided not to utilise certain audio features already provided by the CMU-Multimodal SDK, due to the poor results of the forced alignment [30, 31] algorithms. We are not able to pinpoint specific audio regions for fillers.

We compute textual features as a representation of the whole video to be used in statistical analysis and as input to our linear models. We compute *f_num, f_um, f_uh, f_start, f_mid, f_uncertain and f_stance* by counting each respectively, and then normalising by the total number of tokens in each video. Both *sen_len and f_cog* are normalised by the average sentence length of the video.

RQ1: Are the different functions of fillers correlated differently to the FOAK?

We use a Pearson's correlation coefficient between all the filler features and the label of Confidence, as shown in Figure 2. We take the pValues for correlation between Confidence and other features, where $p < .001$.

Given Table 2, many of the features show strong correlation with Confidence, which is consistent with previous studies as discussed in Section 1. In figure 2, and we see that there

Interjection	Feature name	pValue
Basic	*f_num, f_uh, f_um*	$\boldsymbol{P_s}$**<.001**
Implicature	*f_uncertain, f_mid, f_start, rate*	$\boldsymbol{P_s}$**<.001**
	f_stance	.066
	f_cog	.05
	sen_len	.26

Table 2. pValues for correlation between Confidence and filler features. Highlighted values indicate statistical significance at $p < .001$.

Model	Features	MSE
RV baseline		1.95
RF	Basic	1.11
	Imp	0.85
	Basic + imp	**0.84**
Ridge	Basic	0.89
	Imp	0.84
	Basic + imp	**0.83**

Table 3. Results of the models described in Section 4. *Imp.* stands for implicature features.

exists a negative correlation between the label of Confidence and the *f_num, f_mid*. The negative correlation with *f_start* is smaller than *f_mid*, meaning that there is a distinction between the placement of the fillers. These negative correlations could reflect that listener's are typically not aware that fillers have been used, unless overused, or used in the wrong context [32]. The higher the filler count is, the more conscious the listener becomes of them, and this will impact the Confidence score. Interestingly, *f_uncertain* is negatively correlated with confidence, indicating that listeners do perhaps perceive these filler-stutter occurrences as speech disturbances [15]. Hence, the functions of fillers do correlate differently with the FOAK.

RQ2: Can fillers be informative in the prediction of the listener's FOAK?

We use the original standard training, testing and validation folds provided in the CMU-Multimodal SDK [33]. Our baseline is a *Random Vote* (RV); where 100 random draws respecting the train dataset balance were made. We use a mean squared error (MSE) to evaluate our models. For RV, the MSE is averaged over these 100 samples. We take 2 classic Machine learning algorithms, that is *Random Forest* (RF) and *Ridge regression* (RR), with respective hyper-parameter searches on the validation set. We choose RR as we have multi-collinear features, and both RF and RR have easy interpretability for feature importance.

The main experiments and their results are listed in Table 3. Please refer to Section 2 for the description of the features. In order to get further insight into which features seem relevant to our task, we utilise a RF feature importance, as shown in Table 4. In Table 3, the MSE for the models that use basic filler features, is lower than our baseline in predicting the FOAK. We can see that the impact of adding the implicature features

Rank	Feature	Importance
1	rate	0.210
2	f_num	0.110
3	**sen_len**	0.104
4	**f_mid**	0.102
5	**f_cog**	0.088

Table 4. Top 5 features calculated for the RF model.

decreases the MSE.

Looking at Table 4, we see that the rate of speech (*rate*) is the most important feature. In [28], it was observed that faster speakers "get ahead of themselves"; using more deletions in their speech and beginning anew, whereas slower speakers take more time to plan, increasing hesitations such as repetitions. We interpret that slower speakers may have more moments of hesitations in the videos, and *rate* may be a better indicator of this than *f_uncertain*. However, since rate is also shown as an important general speech feature in personality computing [19], it is not surprising that *rate* is ranked highest.

We would like to focus on the two sets of features that are highlighted in the Table 4, which are the synactic feature *f_mid*, and the cognitive load measured by *f_cog and sen_len*. These are both features that relate to planning and structure of the speech. Given the top 5 features, we interpret that the functions of fillers that are given importance by the models are the *structural functions* of fillers, and not the *metacognitive functions*, such as (*f_stance, f_uncertain*). FOAK could be related to how well formulated the speaker is, and this could be captured by our model. We interpret that the impression the listener has of the speaker could be based on how well structured their argument was.

5. CONCLUSIONS

In this paper we presented a feature study of the role of fillers in the prediction of the FOAK. The contributions of this paper are: 1. To introduce a new and challenging task, that is, FOAK/Confidence prediction, which we think has widespread importance and applicability, 2. To design filler features and observe their role in the FOAK, in the context of free speech, and 3. To highlight that deeper representations of fillers could be beneficial to SLP/affective computing tasks. Future work will be dedicated to expand our features set by mixing filler features with text embedding representations and acoustic features to develop a FOAK prediction system. Given our results, we would also like to investigate the role of discourse features in FOAK prediction. For the acoustic filler features, this will require a corpus with precise speech and text alignment. It would be interesting to see if other SLP tasks benefit from a more fine-grained representation of fillers at a syntactic level, especially if it is able to capture structural elements of a speaker's argument, given that spoken language often lacks structure.

6. REFERENCES

[1] Tanvi Dinkar, Ioana Vasilescu, Catherine Pelachaud, and Chloé Clavel, "Disfluencies and teaching strategies in social interactions between a pedagogical agent and a student: Background and challenges," 2018.

[2] Herbert H Clark and Jean E Fox Tree, "Using uh and um in spontaneous speaking," *Cognition*, vol. 84, no. 1, pp. 73–111, 2002.

[3] Etsuko Yoshida and Robin J Lickley, "Disfluency patterns in dialogue processing," in *DiSS-LPSS Joint Workshop 2010*, 2010.

[4] Susan E Brennan and Maurice Williams, "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*, vol. 34, no. 3, pp. 383–398, 1995.

[5] Martin Corley, Lucy J MacGregor, and David I Donaldson, "It's the way that you, er, say it: Hesitations in speech affect language comprehension," *Cognition*, vol. 105, no. 3, pp. 658–668, 2007.

[6] Joseph P Pickett, *The American heritage dictionary of the English language*, Houghton Mifflin Harcourt, 2018.

[7] Vicki L Smith and Herbert H Clark, "On the course of answering questions," *Journal of memory and language*, vol. 32, no. 1, pp. 25–38, 1993.

[8] H Paul Grice, Peter Cole, Jerry Morgan, et al., "Logic and conversation," *1975*, pp. 41–58, 1975.

[9] Martin Corley and Oliver W Stewart, "Hesitation disfluencies in spontaneous speech: The meaning of um," *Language and Linguistics Compass*, vol. 2, no. 4, pp. 589–602, 2008.

[10] Heather Pon-Barry, "Prosodic manifestations of confidence and uncertainty in spoken language," in *Ninth Annual Conference of the International Speech Communication Association*, 2008.

[11] Marc Swerts and Emiel Krahmer, "Audiovisual prosody and feeling of knowing," *Journal of Memory and Language*, vol. 53, no. 1, pp. 81–94, 2005.

[12] Xiaoming Jiang and Marc D Pell, "The sound of confidence and doubt," *Speech Communication*, vol. 88, pp. 106–126, 2017.

[13] Tobias Schrank and Barbara Schuppler, "Automatic detection of uncertainty in spontaneous german dialogue," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[14] Jeroen Dral, Dirk Heylen, and Rieks op den Akker, "Detecting uncertainty in spoken dialogues: an exploratory research for the automatic detection of speaker uncertainty by using prosodic markers," in *Affective Computing and Sentiment Analysis*, pp. 67–77. Springer, 2011.

[15] Sunghyun Park, Han Suk Shim, Moitreya Chatterjee, Kenji Sagae, and Louis-Philippe Morency, "Computational analysis of persuasiveness in social multimedia: A novel dataset and multimodal prediction approach," in *Proceedings of the 16th International Conference on Multimodal Interaction*. ACM, 2014, pp. 50–57.

[16] François Mairesse, Marilyn A Walker, Matthias R Mehl, and Roger K Moore, "Using linguistic cues for the automatic recognition of personality in conversation and text," *Journal of artificial intelligence research*, vol. 30, pp. 457–500, 2007.

[17] Björn Schuller, Felix Weninger, Yue Zhang, Fabien Ringeval, Anton Batliner, Stefan Steidl, Florian Eyben, Erik Marchi, Alessandro Vinciarelli, Klaus Scherer, et al., "Affective and behavioural computing: Lessons learnt from the first computational paralinguistics challenge," *Computer Speech & Language*, vol. 53, pp. 156–180, 2019.

[18] Paul Ekman, Wallace V Friesen, Maureen O'Sullivan, and Klaus Scherer, "Relative importance of face, body, and speech in judgments of personality and affect.," *Journal of personality and social psychology*, vol. 38, no. 2, pp. 270, 1980.

[19] Alessandro Vinciarelli and Gelareh Mohammadi, "A survey of personality computing," *IEEE Transactions on Affective Computing*, vol. 5, no. 3, pp. 273–291, 2014.

[20] Pentti Haddington, "Stance taking in news interviews," *SKY Journal of Linguistics*, vol. 17, pp. 101–142, 2004.

[21] Esther Le Grezause, *Um and Uh, and the expression of stance in conversational speech*, Ph.D. thesis, 2017.

[22] Divya Saini, "The effect of speech disfluencies on turn-taking," 2017.

[23] Yaniv Leviathan and Yossi Matias, "Google duplex: An ai system for accomplishing real world tasks over the phone.," *Google AI Blog*, 2018.

[24] Gunnel Tottie, "On the use of uh and um in american english," *Functions of Language*, vol. 21, no. 1, pp. 6–29, 2014.

[25] Sowmya Rasipuram and Dinesh Babu Jayagopi, "Automatic assessment of communication skill in interface-based employment interviews using audio-visual cues," in *2016 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*. IEEE, 2016, pp. 1–6.

[26] Elizabeth Shriberg, "To 'errrr'is human: ecology and acoustics of speech disfluencies," *Journal of the International Phonetic Association*, vol. 31, no. 1, pp. 153–169, 2001.

[27] Alexandre Garcia, Slim Essid, Florence d'Alché-Buc, and Chloé Clavel, "A multimodal movie review corpus for fine-grained opinion mining," *CoRR*, vol. abs/1902.10102, 2019.

[28] Elizabeth E Shriberg, "Phonetic consequences of speech disfluency," Tech. Rep., SRI INTERNATIONAL MENLO PARK CA, 1999.

[29] John J Godfrey, Edward C Holliman, and Jane McDaniel, "Switchboard: Telephone speech corpus for research and development," in *[Proceedings] ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing*. IEEE, 1992, vol. 1, pp. 517–520.

[30] Jiahong Yuan and Mark Liberman, "Speaker identification on the scotus corpus," *Journal of the Acoustical Society of America*, vol. 123, no. 5, pp. 3878, 2008.

[31] R. M. Ochshorn and M. Hawkins, "Gentle forced aligner [computer program]," .

[32] Gunnel Tottie, "On the use of uh and um in american english," *Functions of Language*, vol. 21, no. 1, pp. 6–29, 2014.

[33] Amir Zadeh, "Cmu-multimodal sdk," `https://github.com/A2Zadeh/CMU-MultimodalSDK/blob/master/mmsdk/mmdatasdk/dataset/standard_datasets/POM/pom_std_folds.py`.