Disfluencies as Extra-Propositional Indicators of Cognitive Processing

Kathryn Womack
Dept. of ASL
& Interpreting Edu.
kaw8159@rit.edu

Wilson McCoy
Dept. of Interactive
Games & Media
wgm4143@rit.edu

Cecilia Ovesdotter Alm
Dept. of English
coagla@rit.edu

Cara Calvelli
College of Health
Sciences & Tech.
cfcscl@rit.edu

Jeff B. Pelz
Center for
Imaging Science
pelz@cis.rit.edu

Pengcheng Shi
Computing &
Information Sciences
spcast@rit.edu

Anne Haake
Computing &
Information Sciences
anne.haake@rit.edu

Rochester Institute of Technology

Abstract

We explore filled pause usage in spontaneous medical narration. Expert physicians viewed images of dermatological conditions and provided a description while working toward a diagnosis. The narratives were analyzed for differences in filled pauses used by attending (experienced) and resident (in-training) physicians and by male and female physicians. Attending physicians described more and used more filled pauses than residents. No difference was found by speaker gender. Acoustic speech features were examined for two types of filled pauses: nasal (e.g. *um*) and non-nasal (e.g. *uh*). Nasal filled pauses were more often followed by longer silent pauses. Scores capturing diagnostic correctness and diagnostic thoroughness for each narrative were compared against filled pauses. The number of filled and silent pauses trends upward as correctness scores increase, indicating a tentative relationship between filled pause usage and expertise. Also, we report on a computational model for predicting types of filled pause.

1 Introduction

Although they are often not consciously realized, disfluencies are common in everyday speech. In an overview of several studies, Fox Tree (1995) estimates that approximately 6% of speech is disfluent. Disfluencies include filled pauses, silent pauses, edited or repeated words, and sounds such as clearing one's throat or click noises. Disfluencies affect the way that listeners comprehend speech in learning situations (Barr, 2003), formulate opinions of the speaker as being more or less fluent (Lövgren and van Doorn, 2005), and even parse grammatically complex sentences (Bailey and Ferreira, 2003).

Since disfluencies are generally absent in written text, they are irrelevant when analyzing text for extra-propositional meaning, such as uncertainty or modality (Vincze et al., 2008, for example). In contrast, when studying meaning in spoken language, disfluencies provide information about a speaker's cognitive state. For example, they might indicate cognitive load, uncertainty, confidence, thoughtfulness, problems in reasoning, or stylistic preferences between individuals or groups of individuals. We study filled pauses (e.g. *um* and *uh*) and leave other disfluency types for future work.

The presence of filled pauses could indicate context-dependent facets of cognitive reasoning processes. We examine filled pauses present in the speech of highly-trained dermatologists who were shown images of dermatological conditions and asked to provide a description and diagnosis. We look at the difference between two different types of filled pauses: those with nasal consonants, such as *um*; and those without nasal consonants, such as *uh*. We build a computational model to confirm findings that nasal and non-nasal filled pauses differ by prosodic and contextual features. In addition, we first compare whether there is a difference between filled pause use for variables such as level of physician expertise and gender. We also examine the relationship of correctness in the diagnostic process with respect to filled pause use.

There is evidence that filled pauses indicate cognitive processing difficulties and could change the

Proceedings of the ACL-2012 Workshop on Extra-Propositional Aspects of Meaning in Computational Linguistics (ExProM-2012), pages 1–9, Jeju, Republic of Korea, 13 July 2012. ©2012 Association for Computational Linguistics

speaker's intended meaning or the listener's perceived meaning of an utterance. However, such implicit meanings are severely understudied in previous work, especially in specialized, high-stakes domains such as medical diagnostics. Little is understood about what factors impact the linguistic behavior of using certain filled pauses rather than others, and how the use of filled pauses differs based on level of expertise, gender, or diagnostic correctness. Looking into these differences is useful to form a better understanding of the relationship between language and specialized decision-making processes. More specifically, it is necessary to improve the understanding of how speakers' use of filled pauses differs based on the context of speech and how they change the meaning and reception of speech in extra-propositional ways.

2 Previous Work

Filled pauses in English include monosyllables with and without nasal consonants, such as *um* and *uh* respectively. Filled pauses are most common in unstructured, spontaneous speech, but they are also present in prompted, structured speech; and occur in both monologues and dialogues.

Much research has been done into hedging, negation, and other propositional features that change the meaning or modality of phrases (Morante and Sporleder, in press). Less research has been done into the usage of filled pauses and their relationship to certainty and speculation. It has been shown that disfluencies are used to indicate uncertainty in speakers' forthcoming statements or to indicate that the speaker is engaged in the discourse but working to formulate their response (Brennan and Williams, 1995; Smith and Clark, 1993). These studies found that speakers less confident of their answers take longer to answer and use more disfluencies.

Recent studies have suggested that disfluencies provide meaningful information about the speaker's cognitive or linguistic processes (Arnold et al., 2003; Bortfeld et al., 2001; Corley and Stewart, 2008; Oviatt, 1995, for example), and are unintentional indications that the speaker is having difficulty formulating upcoming speech.

More specifically, it has been shown that the two major categories of filled pauses, i.e. nasal and non-nasal, are specific indicators of the level of cognitive load, with nasal filled pauses indicating higher load and non-nasal filled pauses indicating lower load. Barr (2001) performed an experiment in which a speaker described one of several visible images to a listener who then selected the image being described. In this study as well as in Barr and Seyfiddinipur (2010), listeners focused on a topic that was new to the discourse or exceptionally complex when they heard the speaker say *um*. Although they did not differentiate between nasal and non-nasal filled pauses, Arnold et al. (2003; 2007) found in similar experiments that filled pauses often preceded unfamiliar or complex objects.

There is evidence that speakers use filled pauses to indicate different processing difficulties. Clark and Fox Tree (2002) describe four different filled pauses that are annotated in the corpora they use. These are *uh*, *um*, and their elongated versions *u:h* and *u:m*. They argue that each of these corresponds to a different following pause time with *uh* being followed by the shortest pause time, then *u:h*, *um*, and *u:m* followed by the longest. It is important to note that their primary corpus is the London-Lund Corpus of Spoken English, in which the pause times were annotated based on the transcriber's estimate of pause time in units of "one light foot" or "one stress unit" (Clark and Fox Tree, 2002, p. 80) rather than measured in seconds.[1]

However, studies on filled pauses by Barr (2001) and Smith and Clark (1993) measured the duration of silent pauses in seconds and confirm that *um* was followed by longer silent pauses than *uh*. The hypothesis suggested by Barr, Clark and Fox Tree, and Smith and Clark is that *uh* indicates a minor delay and lower level of cognitive difficulty while *um* indicates a major delay due to higher level of difficulty in speech planning and production.

On the other hand, a study by O'Connell and Kowal (2005) refuted the findings of Clark and Fox Tree and showed that specific filled pauses could not predict pause time in their corpus of TV interviews. O'Connell and Kowal's corpus was six interviews conducted by various TV personnel with

[1]The difference between listeners' perception of duration and actual duration is an important one because perceptual and actual duration do not always match (Megyesi and Gustafson-Capkova, 2002; Spinos et al., 2002).

Hillary Clinton because these "professional speakers" (O'Connell and Kowal, 2005, p. 560) should be more likely to use filled pauses according to convention. However, speech in public TV interviews is likely to be pre-planned and highly self-monitored by the speakers, and it may not be appropriate to consider this situation a model for spontaneous, less formal, and less public speech. It has been shown that rate and use of filled pauses can vary widely within certain fields (Schachter et al., 1991), in situations that are more or less structured (Oviatt, 1995), and depending on the formality of the situational context (Bortfeld et al., 2001).

3 Data, Annotation, and Methods

Data were acquired from a study involving 16 dermatologists, including 12 attending physicians and 4 residents. The participants were evenly split for gender. These physicians were shown 50 images of different dermatological conditions and asked to provide a description and diagnosis of each. In a modification of the Master-Apprentice scenario (Beyer and Holtzblatt, 1997), each observer explained his or her thoughts and processes to a student who was silent. These are monologues; however, the Master has the feeling of interaction and of dialogue.

Audio of each description was recorded while eye-movements were tracked. The relationship between eye-movements and extra-propositional features will be the topic of a later study. The audio files were manually single-annotated and time-aligned at the word level in Praat, a software for acoustic and phonetic analysis (Boersma, 2001). A section of the spoken narrative with time-alignment is pictured in Figure 1. Praat and Python scripts were used to computationally extract measurements of pitch, intensity, and duration for words, silent pauses, and narratives. In total, there were 800 audio-recorded narratives. At this time, 707 of these narratives have been time-aligned and annotated and only these are used in this study.

Four transcribers worked independently on time-alignment, and they were given instructions by one coordinator. Every spoken token was included in the transcriptions, including filled pauses, extra-linguistic sounds such as clicks, repairs, and silent pauses. Annotators were instructed to mark only

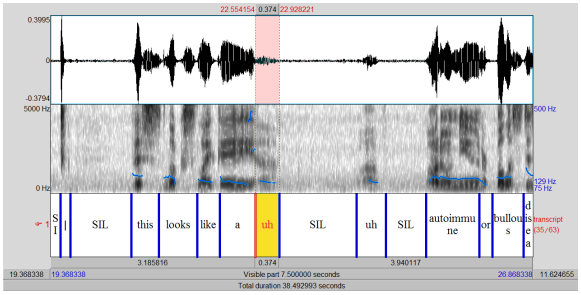

Figure 1: Screenshot of the program Praat which was used to time-align each narrative and extract acoustic prosodic information about the physicians' speech.

silent pauses that were longer than 30 milliseconds, because it has been shown that pauses under 20-30 ms are not consistently perceived by listeners in discourse (Kirsner et al., 2002; Lövgren and van Doorn, 2005).

After word-level time-alignment, each narrative was independently annotated by three expert dermatologists who did not participate in the original data elicitation procedure. Each narrative was examined for *medical lesion morphology* (the description of the condition), *differential diagnosis* (possible diagnostic conditions), and *final diagnosis* (the diagnosis that the observer found most likely). These independent experts annotated the physicians' diagnostic correctness for the three steps of the diagnostic process. They annotated medical lesion morphology as *correct*, *incorrect*, *correct but incomplete*, or *none*, indicating that no medical morphology was given. Final diagnosis was labeled as *correct*, *incorrect*, or *none*, and differential diagnosis was rated as *yes*, *no*, or *no differential given*. An analysis of the annotated data set is discussed by McCoy et al. (Forthcoming 2012).

4 Results and Discussion

4.1 Types of Filled Pauses

Nasal filled pauses included *hm* and *um* and non-nasal filled pauses included *ah*, *er*, and *uh*. We analyzed nasal and non-nasal filled pauses as groups rather than each individual filled pause because the number of filled pauses within each category was not balanced. Higher token counts of *uh* and *um* were identified, with fewer *ah*, *er*, and *hm* filled pauses. In comparing use of nasal and non-nasal filled pauses,

FPs	No.	Dur.	St. Dev.	%
hm	78	0.48 s	0.20	2%
um	1439	0.51 s	0.19	36%
Total (nasal)	1517	0.50 s	0.19	38%
ah	23	0.46 s	0.23	1%
er	9	0.26 s	0.09	<1%
uh	2401	0.36 s	0.16	61%
Total (non-nasal)	2433	0.36 s	0.16	62%
Total (all)	3950	0.42 s	0.19	100%

Table 1: Total number of each type of filled pause (FPs) with mean duration in seconds, standard deviation of the mean duration, and percentage of all filled pauses.

we considered all 707 narratives. The number of tokens and average duration for each filled pause is given in Table 1.

The average filled pause duration was slightly longer for nasal than for non-nasal, likely due to the segmental quality.

In total, 38% of the filled pauses in our data set are nasal. However, observers vary widely in their individual usage, from one observer who used 22 non-nasal (10%) and 189 nasal (90%) filled pauses to an observer at the other extreme who used 562 non-nasal (97%) and only 19 nasal (3%) filled pauses. Some people seem to have a tendency to use one type of filled pause over the other.

Clark and Fox Tree (2002) found that nasal filled pauses were more often followed by silent pauses and that those silences were on average longer than that of non-nasal filled pauses. Our data are consistent with this as shown in Tables 2 and 3,[2] and Figure 2. Of the total nasal filled pauses, 70% were followed by a silent pause, whereas only 41% of non-nasal filled pauses were followed by a silent pause.

The mean duration of silent pauses following nasal filled pauses was 1.5 s while non-nasal was 1.1 s, which indicates a difference significant enough that it could be recognized by a listener. These findings show that nasal filled pauses are good indicators of continuing delay, which supports Clark and Fox Tree's hypothesis that nasal and non-nasal filled pauses are used to indicate different levels of difficulty in speech planning. Taken with the results of experiments by Barr (2001) that nasal filled pauses are more often used before a topic that is relatively

[2]The data were analyzed using two-sample t-tests assuming unequal variances.

	Nasal (*hm*, *um*)	Non-nasal (*ah*, *er*, *uh*)	*p*
Dur. of FPs	0.50 s	0.36 s	< 0.01
Dur. of FPs + SILs	2.46 s	1.37 s	< 0.01
No. of FPs	1517	2433	n/a

Table 2: Mean duration in seconds of filled pauses (FPs), and mean duration of the filled pause including the span of any preceding and following silences. If there were no silences, only the duration of the filled pause was used to calculate the mean.

	Nasal (*hm*, *um*)	Non-nasal (*ah*, *er*, *uh*)	*p*
Dur. of pre. SILs	1.19 s	1.15 s	0.4
No. of pre. SILs	1167	1197	n/a
Dur. of foll. SILs	1.50 s	1.07 s	< 0.01
No. of foll. SILs	1059	1006	n/a

Table 3: Mean duration in seconds of silent pauses (SILs) preceding filled pauses, silent pauses following filled pauses, and the number of tokens for each. Durations were only considered if there was a silence, so the number of silences was different for each calculation.

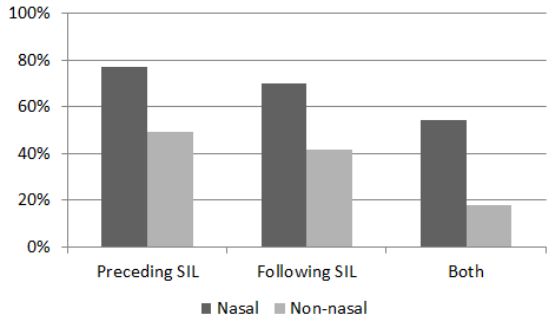

Figure 2: The percentage of nasal and non-nasal filled pauses with a preceding silent pause, following silent pause, and a silent pause both preceding and following.

complex or new to discourse, it seems that nasal filled pauses indicate a higher level of cognitive difficulty than non-nasal filled pauses.

In their previously-mentioned study, Clark and Fox Tree also found that nasal filled pauses were more often preceded by delays and that those delays were longer. Similarly, in our data 77% of the nasal filled pauses were preceded by silences, compared with 49% of non-nasal.

No difference was found in the mean duration of preceding silences, however. Although this conclusion is tentative, it seems that the duration of the preceding pause could be the maximum length of silence a speaker feels is permissible before needing to indicate their continuing participation in the discourse. This supports Jefferson's (1989) findings of a "standard maximum silence" of around 1 second in discourse. At that point, the speaker could need to signal that they have more to say, using a nasal filled pause if they anticipate a long delay or a non-nasal filled pause if they anticipate a shorter delay. The longer duration of surrounding silent pauses for nasal filled pauses also supports the conclusion that they indicate higher cognitive load and more pre-planning. This critical finding highlights the importance of considering filled pauses in computational modeling and hint at their potential usefulness across phenomena of extra-propositional meaning.

4.2 Gender

Traditional stereotypes have held that women are less confident speakers than men. When women and men use the same number of hedge words or modifiers, women are judged more harshly as sounding passive or uncertain (Bradley, 1981). Although different rates and ratios of filled pauses were identified, Acton (2011), Binnenpoorte et al. (2005), and Bortfeld et al. (2001) all found that women used a lower rate of filled pauses than men. Acton also found that women consistently used a higher ratio of nasal filled pauses.

Our data were analyzed at the level of diagnostic narrative based on the means of: number of filled pauses, filled pauses per second, the percentage of filled pauses (i.e. the rate per 100 words), the number of nasal filled pauses, and the percentage of nasal filled pauses. The difference between the means was not statistically significant, confirmed by the computed p-score.[3] Hence, our data do not support a difference in men's and women's use of filled pauses.

There are several possible explanations for this. For example, it has been shown that women tend to be more conscious of their speaking style than men because they are aware of the stereotyping mentioned previously (Gordon, 1994), and they may make more effort to speak clearly. Acton (2011) and Bortfeld et al. (2001) noted different usage of filled pauses by men and women in different situations. Whereas our results point to gender neutrality and refute the common gender bias as well as findings of previous studies, we recognize that our results could reflect that this study involved a largely homogeneous professional and educational group. The studies mentioned thus far used corpora consisting of casual conversations in various situations with individuals of various backgrounds. Further research into gender differences in expert fields could clarify this factor further.

4.3 Level of Expertise

Our data were analyzed based on the means per narrative, similar to Section 4.2, but comparing levels of expertise (attending versus resident physicians). Attending physicians' narratives had a longer mean duration and significantly more words. Attending physicians also used more filled pauses, a higher rate of filled pauses per 100 words, and a higher percentage of nasal filled pauses (see Table 4).[4]

One probable explanation for the difference is that the experienced attendings noticed more about the image, leading them to give more information about their thought processes and go into more detail than residents. It is possible also that the attendings' experience could have provided them with a larger conceptual space and options to explore. This explains the longer narrative time and the higher number of words used. Many of the dermatological terms used are highly complex and may require explanation on the part of the observer, and other stud-

[3]The mean of each category was determined for each observer, and then analyzed using a two-sample t-test. In total, we had 355 narratives from males and 352 from females.

[4]These results were calculated using the mean of each observer and each narrative. A paired t-test was used to compare means for residents on each image against means for attendings on each image.

For Narratives	Attendings' Means	Residents' Means	p
Total Dur.	46.1 s	33.8 s	< 0.01
No. of Words	85.7	50.9	< 0.01
No. of FPs	6.3	1.9	< 0.01
% FPs	8%	4%	< 0.01
% Nasal FPs	0.4%	0.2%	< 0.01

Table 4: Analysis considered, at the narrative level, attending and resident physicians' mean total duration, number of words (including filled and silent pauses), number of filled pauses (FPs), percentage of filled pauses of total words (total words includes pauses; without pauses, this rate would be higher), and percentage of nasal filled pauses of total filled pauses.

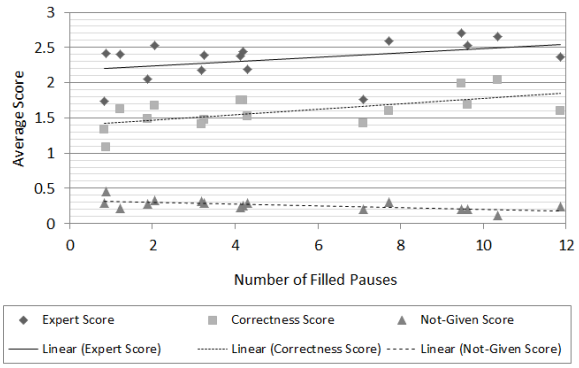

Figure 3: Average number of filled pauses per narrative by observer (y-axis) against the holistic expert score, correctness score, and not-given score (x-axis).

ies have found that the filled pause rate increases as the utterance length increases (Oviatt, 1995; Bortfeld et al., 2001), so one would expect to see more filled pauses used in longer descriptions.

One issue with our data is that the number of attending physicians and the number of resident physicians is not balanced. We had 592 narratives done by 12 attendings and 115 done by 4 residents. All values were calculated using means so the values are not weighted based on the number of narratives analyzed. However, we have previously mentioned that personal preference plays a role in the usage of filled pauses, and we have a wider variety of attending observers than resident observers. It could be that our resident observers happened to be the kinds of people who do not use many filled pauses.

4.4 Diagnostic Correctness

Three scores were determined for each narrative. The first score was the *holistic expert score* provided by the expert annotators, based on "relevancy, thoroughness, and accuracy" of each narrative from 1 to 3 with 3 being the best. The second score was an overall *correctness score* which spanned from 0 to 3, with one-third of a point given per independent annotator for each step (i.e. medical lesion morphology, differential diagnosis, and final diagnosis) if *correct* and $\frac{1}{3} * 0.5$ points given for *correct but incomplete*. The last score was the *not-given score* which, similar to the correctness score, spanned from 0 to 3 with one-third of a point given per annotator for each step if the original observer did not provide that information.[5]

Correlation between these three scores and the number or rate of words, filled pauses, and silent pauses was not strong enough to make predictions, indicating that more factors than just the scores should be considered. However, certain trends were evident. As the holistic expert and correctness scores improved, the means of narratives' total duration in seconds and total number of words also increased. This finding, combined with the fact that experienced physicians spoke more and had higher average correctness and expert scores, indicates that verbal behavior can reflect both heightened conceptual knowledge and level of expertise.

The number of filled pauses per narrative, number of silent pauses per narrative, and the total duration of filled and silent pauses (per narrative) also increased as the holistic expert and correctness scores improved and the not-given score decreased. The graph of filled pauses in Figure 3 indicates that the increase in the number of filled and silent pauses involve more cognitive processing. That the not-given score tends to inversely decrease could indicate very little cognitive processing (e.g., if an observer was so unsure that they did not even hazard a guess).

The number and percentage of nasal filled pauses, as opposed to non-nasal filled pauses, increased at

[5]There was not a strong correlation between the holistic expert, correctness, and not-given scores, but each score measured different criteria. The mean holistic expert score was 2.3 with a standard deviation of 0.5; the mean correctness score was 1.6 with a standard deviation of 0.8; and the mean not-given score was 0.26 with a standard deviation of 0.16.

a slightly higher rate as the holistic expert and correctness scores increased. This could indicate that nasal filled pauses indicate a higher cognitive load and therefore more consideration in the decision-making process. However, as discussed in Section 4.1, this corpus has more non-nasal than nasal filled pauses and some observers have a particular preference, so this would need to be controlled and investigated further.

5 Computational Model of Filled Pauses Based on Speech Features

A computational model was developed to classify filled pauses as either nasal or non-nasal,[6] based on features discussed in our analysis and in previous work. This model performs above a majority class baseline, supporting our findings that there are differences between the two types of filled pauses, given the features that we have examined, which can be captured by a computational model.

The features considered for classification were total duration and number of words in the narrative; duration, intensity, mean pitch, minimum pitch, and maximum pitch of the filled pause;[7] the filled pause's time and word position in the narrative; time and word position as a percentage of the total narrative; and length of silent pauses[8] on each side of the filled pause. The CFS subset evaluation features selection algorithm was first applied. The filled pause duration, maximum pitch, left silence length, and right silence length were maintained as features for classification; other features were not used further.

The widely used J48 decision tree algorithm in Weka[9] was used to classify our data, which allowed us to visualize our model. The experimental approach was guided by the relatively small size of the dataset. We wanted to avoid over- or under-interpretation of results based on just a small held-out test set. The data were shuffled and partitioned differently during tuning and testing to ensure distinct identities of the data splits so that parameters were not tuned on test folds. The algorithm's parameters were tuned using 5-fold cross-validation; the best-performing fold's parameters were chosen. The data were then shuffled anew and split into 10 folds with each fold being the test set for one experimental run. Results are reported on the final 10-fold cross-validation case.

		Predicted	
		Nasal	Non-nasal
Actual	Nasal	900	617
	Non-nasal	462	1971

Table 5: Confusion matrix of classification results.

The baseline for this model was 62% because the majority class, non-nasal filled pauses, comprised that percentage of the data set. Our model correctly classified 73% of the instances, performing 11% above the baseline. A confusion matrix of the classifier output is shown in Table 5. The model performs best for non-nasal filled pauses, likely because they are more common.

The output of the decision tree indicated that duration of the filled pause was the most important feature. As discussed in Section 4.1, this corresponds with our previous statistical findings as well as those of Clark and Fox Tree (2002) that there is a difference in duration of filled pauses. The next most important features were the left and right silence lengths, also supported by our analysis as well as by Clark and Fox Tree (2002) and Barr (2001). The last selected feature was the maximum pitch of the filled pause, possibly due to phonemic qualities.

This computational model mirrors the findings of Section 4.1 that the duration of filled pauses and of surrounding silent pauses are a differentiating factor between nasal and non-nasal filled pauses and that the contextual surroundings of each filled pause type are different. The finding that the two distinct types of filled pauses behave differently in this domain could also aid language processing systems for clinicians in the medical field. Further research into filled pause and other speech phenomena in each step of the diagnostic process (i.e. medical lesion morphology, differential diagnosis, and final diagnosis) could also be explored in future work.

[6] We also made a fine-grained model to classify specific filled pauses *ah*, *er*, *hm*, *uh*, and *um*. It had 70% accuracy but was generally unable to identify the least-often occurring *ah*, *er*, and *hm* filled pauses, so it is not reported on here.

[7] Pitch features were extracted considering gender: 75-300 Hz for men and a 100-500 Hz for women.

[8] If there was no silence, the value was 0.

[9] See http://www.cs.waikato.ac.nz/ml/weka/.

6 Conclusion

The results of this study underscore the need for further research into the production of disfluencies, especially in decision making situations and in specialized fields such as dermatology. Future work will further explore their connection with highly relevant extra-propositional meaning phenomena in diagnostic verbal behaviors such as certainty, confidence, correctness, and thoroughness.

This study has shown that the two main types of filled pauses, nasal and non-nasal, differ in their usage. Nasal filled pauses are more likely to be preceded and followed by silent pauses, and these following silent pauses are more likely to be longer. These findings are reinforced by the computational model which identified the duration of the filled pause, duration of surrounding silences, and pitch as important for classification of filled pause type.

That longer and more frequent silent pauses surround nasal filled pauses supports the hypothesis that nasal filled pauses indicate a higher level of cognitive load (Clark and Fox Tree, 2002) or a topic that is new to the discourse or unusually complex (Barr, 2001; Barr and Seyfiddinipur, 2010).

The lack of differences in use of filled pauses by speaker gender given the differences found by Acton (2011), Binnenpoorte et al. (2005), and Bortfeld et al. (2001) shows that more research is needed to understand gender variation in speech.

Another finding was that level of expertise influenced the use of filled pauses and overall narrative length. On average, attending physicians spoke longer, said more, used more filled pauses, and had a higher percentage of nasal filled pauses. Attending physicians also had slightly higher holistic expert and correctness scores and were more likely to provide medical lesion morphology, differential diagnosis, and final diagnosis. We believe that attending physicians likely noticed more about the images due to their experience.

The differences by level of expertise (in our study, between attending and resident physicians) need to be verified and compared with more data and in non-medical fields. The differences could also be related to teaching experience of the attending physicians, so further research could compare experienced physicians who are also teachers with those who are not, and if their speaking style affects students' comprehension. In general, differences in linguistic behaviors in relation to levels of expertise deserve more research, and might have long-term implications for development of clinical decision-support and training systems.

The information used by the physicians in our study was limited; they were only shown images of dermatological conditions without being able to examine the patient, run diagnostic tests, or have a patient history. This may have changed their the behavior, along with factors such as the difficulty of diagnosis of each image and their role in the Master-Apprentice scenario. Understanding how these variables affect the diagnostic process of physicians could help us understand how disfluencies are impacted by the contexts of diagnostic decision-making.

The differences found between the use of filled pauses based on level of expertise and on the correctness of narratives seem to indicate that filled pauses could provide partial information about the experts' decision-making process as well as level of confidence and certainty. This is especially important in the medical domain in order to understand how physicians' verbal behaviors are interpreted by other physicians as well as by patients and students.

We recently collected a similar, larger data set and we plan to further examine differences based on expertise in this new corpus. In the recent data collection, observers were also asked to rate their level of certainty about the diagnosis. This provides the opportunity to examine the relationship between disfluencies and certainty. We have eye-tracking data for both studies and future work will also look at eye-movements in relation to the use of filled and silent pauses, certainty, expertise level, and cognitive load.

Acknowledgements

Supported in part by NIH 1 R21 LM010039-01A1, NSF IIS-0941452, RIT GCCIS Seed Funding, and RIT Research Computing (http://rc.rit.edu). We thank Lowell A. Goldsmith, M.D. and the anonymous reviewers for their comments, and Dr. Rubén Proaño for input on statistical analysis.

References

Eric K. Acton. 2011. On gender differences in the distribution of um and uh. *University of Pennsylvania Working Papers in Linguistics*, 17(2).

Jennifer E. Arnold, Maria Fagnano, and Michael K. Tanenhaus. 2003. Disfluencies signal theee, um, new information. *Journal of Psycholinguistic Research*, 32(1):25–36.

Jennifer E. Arnold, Carla L. Hudson Kam, and Michael K. Tanenhaus. 2007. If you say thee uh you are describing something hard: The on-line attribution of disfluency during reference comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 33(5):914–930.

Karl G.D. Bailey and Fernanda Ferreira. 2003. Disfluencies affect the parsing of garden-path sentences. *Journal of Memory and Language*, 49:183–200.

Dale J. Barr and Mandana Seyfiddinipur. 2010. The role of fillers in listener attributes for speaker disfluency. *Language and Cognitive Processes*, 25(4):441–455.

Dale J. Barr. 2001. Trouble in mind: Paralinguistic indices of effort and uncertainty in communication. *Oralité and gestualité: Communication Multimodale, Interaction*, pages 597–600.

Dale J. Barr. 2003. Paralinguistic correlates of conceptual structure. *Psychonomic Bulletin & Review*, 10(2):462–467.

Hugh Beyer and Karen Holtzblatt. 1997. *Contextual Design: Defining Customer-Centered Systems*. Morgan Kaufmann.

Diana Binnenpoorte, Christophe Van Bael, Els den Os, and Lou Boves. 2005. Gender in everyday speech and language: A corpus-based study. *Interspeech*, pages 2213–2216.

Paul Boersma. 2001. Praat, a system for doing phonetics by computer. *Glot International*, pages 341–345.

Heather Bortfeld, Silvia D. Leon, Johnathan E. Bloom, Michael F. Schober, and Susan E. Brennan. 2001. Disfluency rates in conversation: Effects of age, relationship, topic, role, and gender. *Language and Speech*, 44(2):123–147.

Patricia Hayes Bradley. 1981. The folk-linguistics of women's speech: an empirical investigation. *Communication Monographs*, 48(1):78–91.

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*, 34:383–398.

Herbert H. Clark and Jean E. Fox Tree. 2002. Using uh and um in spontaneous speaking. *Cognition*, 84:73–111.

Martin Corley and Oliver W. Stewart. 2008. Hesitation disfluencies in spontaneous speech: The meaning of um. *Lang. and Linguistics Compass*, 2(4):589–602.

Jean E. Fox Tree. 1995. The effects of false starts and repetitions on the processing of subsequent words in spontaneous speech. *Journal of Memory and Language*, 34:709–738.

Elizabeth Gordon. 1994. Sex differences in language: Another explanation? *American Speech*, 69(2):215–221.

Gail Jefferson. 1989. Notes on a possible metric which provides for a 'standard maximum' silence of approximately one second in conversation. In Derek Roger and Peter Bull, editors, *Conversation*, chapter 8, pages 166–196. Multilingual Matters, Clevedon, UK.

Kim Kirsner, John Dunn, Kathryn Hird, Tim Parkin, and Craig Clark. 2002. Time for a pause. *Proc. of the 9th Australian Int'l. Conf. on Speech Science & Tech.*, pages 52–57.

Tobias Lövgren and Jan van Doorn. 2005. Influence of manipulation of short silent pause duration on speech fluency. *Proceedings of DiSS05*, pages 123–126.

Wilson McCoy, Cecilia Ovesdotter Alm, Cara Calvelli, Jeff Pelz, Pengcheng Shi, and Anne Haake. Forthcoming-2012. Linking uncertainty in physicians' narratives to diagnostic correctness. *Proc. of the ExProM 2012 Workshop*.

Beata Megyesi and Sofia Gustafson-Capkova. 2002. Production and perception of pauses and their linguistic context in read and spontaneous speech in Swedish. *ICSLP 7*.

Roser Morante and Caroline Sporleder. in press. Modality and negation: An introduction to the special issue. *Computational Linguistics*.

Daniel C. O'Connell and Sabine Kowal. 2005. uh and um revisited: Are they interjections for signaling delay? *Journal of Psycholinguistic Research*, 34(6):555–576.

Sharon Oviatt. 1995. Predicting and managing spoken disfluencies during human-computer interaction. *Computer Speech and Language*, 9:19–35.

Stanley Schachter, Nicholas Christenfeld, Bernard Ravina, and Frances Bilous. 1991. Speech disfluency and the structure of knowledge. *JPSP*, 60(3):362–367.

Vicki L. Smith and Herbert H. Clark. 1993. On the course of answering questions. *Journal of Memory and Language*, 32:25–38.

Anna-Marie R. Spinos, Daniel C. O'Connell, and Sabine Kowal. 2002. An empirical investigation of pause notation. *Pragmatics*, 12(1):1–9.

Veronika Vincze, György Szarvas, Richárd Farkas, György Móra, and János Csirik. 2008. The bioscope corpus: Biomedical texts annotated for uncertainty, negation, and their scopes. *BMC Bioinformatics*, 9.