Pauses in Deceptive Speech

Stefan Benus[*]*, Frank Enos*[*]*, Julia Hirschberg*[*] *& Elizabeth Shriberg*[§]

[*]Department of Computer Science, Columbia University, New York, USA
[§]SRI International, Menlo Park, USA & ICSI, Berkeley, USA
sbenus@cs.columbia.edu

Abstract

We use a corpus of spontaneous interview speech to investigate the relationship between the distributional and prosodic characteristics of silent and filled pauses and the intent of an interviewee to deceive an interviewer. Our data suggest that the use of pauses correlates more with truthful than with deceptive speech, and that prosodic features extracted from filled pauses themselves as well as features describing contextual prosodic information in the vicinity of filled pauses may facilitate the detection of deceit in speech.

1. Introduction

Everyday spontaneous human communication is rich in various types of disfluencies. Pauses, whether vocalized or silent, are among the most common speech disfluencies. Pauses tend to occur at salient points in discourse, affect both rhythmical and intonational aspects of speech, and can convey a wide variety of intentional and unintentional communicative messages (e.g. [19], [16], [4], [20]). In this paper we examine the use of filled and silent pauses as cues to the detection of deception in speech. We use a new corpus of deceptive and non-deceptive speech ([12]), as well as new features of filled pauses, and test previous hypotheses in the literature that suggest that pauses provide useful predictors of deception. Specifically, we are interested in determining: a) whether the use of silent and filled pauses can aid the detection of deception, b) whether there are differences among *um*, *uh*, and the silent pause in cuing deceptive speech, and c) whether prosodic features of filled pauses facilitate the detection of deception

1.1. Previous Research

In the literature, filled pauses such as *um* and *uh* have been found to signal the length of the delay of upcoming speech ([18], [5]), to mark speakers' intentions to assume and hold the floor in dialogues ([19]), to facilitate the perception of upcoming linguistic material ([11], [10]), to signal discourse structure ([16]) to aid in the management of interpersonal communication ([2], [4]), to signal the strength of the preceding intonational boundaries ([20]), to correlate intonationally with preceding speech ([17]), and influence syntactic parsing ([8]). Silent pauses in pre-focal position have been shown to add emphasis but also to signal non-assertiveness and to strengthen listeners' perception of question intonation ([11]). Several studies have argued that pragmatic factors such as the speaker's comfort with the topic, honesty ([9]), or certainty about their answers ([3]) can be signaled by pauses as well.

Previous research in deceptive speech provides conflicting evidence for the importance of pauses as cues to speaker deceptiveness. On the one hand, the construction of deceptive utterances is assumed to require increased cognitive load compared to the formulation of truthful utterances ([24]). Pauses, both silent and filled, are thus hypothesized to be automatic reactions to speech-planning problems arising from the increase in cognitive load required for deception ([11], [14]). In perception studies of subjects asked to detect deception in others' speech, this hypothesis has been supported in studies such as [9] where filled pauses following a direct question were perceived as signaling subjects' discomfort with the topic or the preparation of a dishonest answer. In production studies such as [21] and [22], subjects who were instructed to deceive police officers in mock interrogations used more filled pauses (*ums* and *uhs*) than subjects instructed to tell the truth, and people who fabricated more complex lies were observed to use more speech disturbances than those who fabricate simple ones.

On the other hand, a recent meta-study of 120 independent sample groups ([6]) has found that speech disturbances have little predictive power as cues to deceit. The effect of filled pauses was negligible when results from all samples were combined. Surprisingly, however, when subjects were given explicit incentive to deceive, deceptive speech contained fewer filled pauses than truthful speech, although the effect was not statistically significant. This is consistent with the hypothesis of some practitioners (c.f. [15]) that deceptive speech is more careful or planned, which in turn predicts fewer pauses compared to non-deceptive speech.

Turning now to the relationship between deception and response latency, lies are predicted to be preceded by longer latencies ([7], [24]). This prediction stems from the observation that deception correlates with attempts of the subjects to control their behavior, the amount of time they spend thinking, and their feelings of guilt ([6]). However, latency was not a significant factor in determining deception, although lies were preceded by slightly longer latencies ([6]).

Thus, there are mixed claims and findings about the importance of filled and unfilled pauses in signaling deception. While some studies have found differences between deceptive and non-deceptive speech with respect to different aspects of pausing, there appears to be no clear and simple result in the literature on the subject. In this paper we examine filled and silent pauses and their characteristics as cues to deception in a new corpus of deceptive and non-deceptive speech. In addition to investigating whether the presence or absence of pauses signals deception, we also focus on the following hypotheses:

1.2. Do *ums, uhs* and silent pauses behave similarly in cuing deception?

The literature that addresses differences among the two types of filled pauses and unfilled pauses provides mixed findings with respect to how the three pattern in cognitive tasks. Some studies have suggested that *uhs* pattern together with silent pauses and that both contrast with *ums*. For example, in [10],

cue words were recognized faster when preceded by *um* than by a silent pause or *uh*. In [18], *um* signaled a longer following pause than *uh* and thus it was argued that speakers consciously choose between *um* and *uh* to signal the depth of their retrieval problem. Assuming that the construction of deceptive utterances increases cognitive load and subsequent planning problems, the presence of *um* should be a better predictor for deception than *uh*.

Other studies, however, have proposed that the two filled pauses pattern together and contrast with silent pauses. For example, [3] argued that the type of pause (filled vs. silent) affects listeners' judgments of recorded speech as to whether the speakers knew the answer to a question. However, there was no significant difference between *um* and *uh*. This study concluded that, while filled pauses imply different perception than silent pauses, *um* and *uh* do not differ in their meanings.

Although these studies employed paradigms other than deception, the behavior of the three types of pauses in signaling cognitive meanings does not seem to be uniform. Hence, in an effort to shed more light on the relationship between pauses and deception, we also analyze potential cues to deception of each pause type separately.

1.3. Can prosodic features of filled pauses help in detecting deception?

To our knowledge, while the presence or absence of pauses as potential cues to deception has been investigated, the only prosodic feature that has been examined is the length of the pauses. However, several studies have investigated the link between deception and other prosodic features. For example, [6] found cross-study evidence for increase in pitch as an indicator of deception. Higher pitch is assumed to indicate increased tension on the part of deceivers. Hence, we hypothesize that filled pauses with higher pitch and intensity may occur in deceptive speech.

However, in general, clause-internal filled pauses tend to be produced with lower pitch register than surrounding phrases ([17]). Therefore, the differences in the setting of the pitch register for filled pauses and the rest of the utterance may cancel out the potential link between deception and higher pitch. Therefore, we investigate the usefulness of features extracted from the filled pause itself such as mean or maximum of pitch and intensity as well as the potential of 'dynamic' features such as the changes in the means and maxima of the filled pause and the material that surrounds it. We now describe the corpus on which we test these hypotheses.

2. Corpus and Methodology

2.1. The Corpus

The Columbia/SRI/Colorado (CSC) Deception Corpus ([12]) consists of 32 interviews averaging 30 minutes. The subjects, equally divided between males and females, were first tested in 6 areas of general knowledge and skills, and then informed of their scores. The subjects were next promised a monetary incentive if they could persuade an interviewer that their performance in the tasks was consistent with that of a target profile. (In all cases, the subjects' performance was manipulated by varying the difficulty of tasks such that their performance in fact differed substantially from the target profile on four tasks and matched on two. They were thus motivated to lie to the interviewer on four tasks.) We will refer to deception related to these tasks as **global deception**. Subjects were also asked to press a pedal invisible to the interviewer after each of their responses, to indicate if any part of their previous utterance was false or not. The data from these pedal presses will be referred to as **local deception**.

The speech of both the subject and the interviewer were recorded with a head-mounted microphone on a digital recorder in a sound-proof room. Hand transcriptions of the conversations were then aligned with the sound signal using automatic forced alignment. The speech of the subjects (approximately 7 hours in total) was segmented into sentence-like units (SUs) based on the punctuation in the transcription. Of these units, 9068 were coded for local deception, and 5435 SUs were classified as truths and 3633 as lies.

2.2. The Data

Due to the experimental design, there is more data in the category of global lies than truths. Yet, the corpus contains more locally truthful than deceptive speech. Therefore, the global bias for deception induced by the experimental setup did not prevent a general tendency of subjects to produce truthful utterances.

The data from all 32 speakers yielded 2103 tokens of *um*, and 1511 tokens of *uh*, for a total of 3614 filled pauses, which constitutes approximately 4.5% of all words in the transcripts. This rate is slightly higher than the rates found in other corpora. For example, [16] reported the rate of 3% in more controlled air-travel dialogues (AMEX) and 2% in less control conversations (Switchboard). The rate of filled pauses was higher for males than for females (5.3% vs. 3.7%), which confirms previous findings ([16]).

Some speech in the corpus related to the experimental procedure rather than to the actual paradigm and thus was categorized as 'off-talk' and not assigned a truth value. Due to minor differences in the classification of the 'off-talk' for some analyses, the number of filled pauses included in the analyses slightly varies. Out of 3614 filled pauses, 3246 (3303) were labeled for global deception and 3495 (3555) for local deception.

We automatically extracted standard features such as mean, maximum and minimum of F_0 and intensity from each filled pause. We then normalized these values by calculating z-scores for individual speakers to minimize the effect of anatomical and physiological factors of acoustic measures. To investigate the potential effect of deception on changes in F_0 and intensity in the vicinity of filled pauses, we also extracted dynamic prosodic features in the following way. We located the pause-defined units (PDU) that contain an FP, the following PDU if the filled pause was followed by a silent pause, and the preceding PDU if the filled pause was turn-internal and was preceded by a silent pause. From the stylized F_0 and raw intensity of these units we then automatically extracted various targets (e.g. maximum, first F_0 peak, etc.) and calculated the ratios between the targets of the filled pauses and those in the surrounding material. To obtain more reliable dynamic features, we hand-corrected the stylized F_0 contours for spurious or missing targets in the subset of the corpus (7 interviews). This gave us information about 485 filled pauses.

Finally, we also extracted both turn-internal and turn-initial silent pauses. The ratio of turn-internal pauses over fluent transitions between word pairs was 20.2% (pause/all-

transitions) or 24.3% (pause/non-pause). Turn-initial silent pauses, or latencies, were extracted in those turns that followed a direct question from the interviewer. This provided us with 3116 latency tokens for the analysis.

3. Analysis and Results

3.1. Presence vs. absence of pauses

Subjects used filled pauses significantly more frequently in locally truthful than in locally deceptive statements, $\chi^2(1, N = 76635) = 20.515$, $p < 0.001$). The same generalization was observed in the subset of filled pauses that occurred turn-initially, $\chi^2(1, N = 3803) = 31.47$, $p < 0.001$. This finding corroborates the findings in [6]. Note, however, that subjects in this experiment had little time to plan their responses, since the interviews occurred just after the tasks they performed. The frequency of filled pauses in global lies was not significantly different from the frequency in global truths, $\chi^2(1, N = 73800) = 0.251$, ns., $\chi^2(1, N = 3450) = 1.54$, ns. in turn-initial position.

Turn-internal silent pauses also occurred more frequently in truthful than in deceptive speech. This was the case both locally, $\chi^2(1, N = 74585) = 45.27$, $p < 0.001$, and globally, $\chi^2(1, N = 71879) = 24.80$, $p < 0.001$. This result was confirmed by calculating the temporal distance between each pair of consecutive pauses within a turn. One-way ANOVA showed that silent pauses in local truths were closer in time to each other than in those in local lies, $F(1, 14954) = 16.002$, $p < 0.001$. Silent pauses were also systematically longer in lies than in truths, but this effect was not significant.

The length of turn-initial silent pauses was **not** a significant predictor of deception in our corpus. We tested a) latency for all responses, b) latency when the response began with a filled pause, and c) total latency calculated as the sum of the raw latency, the length of a turn-initial filled pause if present, and the length of a following silent pause if present. None of these measures showed a significant effect of deception either locally or globally. However, the latency to response **was** systematically longer before deceptive utterances than before truthful ones; mean difference was around 20ms. The global deception factor did not affect latencies in any systematic pattern.

Hence, in terms of the distribution of filled and silent pauses in the corpus, we find that indeed there are significantly fewer pauses in lies than in truths and that there is a tendency for latencies to be longer before lies than before truthful statements.

3.2. Differences between *um* and *uh*

Examining *um* vs. *uh* in our corpus, we first find that *um* was more likely to be followed by a silent pause than was *uh*, $\chi^2(1, N = 3614) = 301.64$, $p < 0.001$. The length of silent pause following turn-initial *um* was also significantly greater than the length of silent pause following turn-initial *uh*, $F(1,1196) = 93.49$, $p < 0.001$; mean difference 455ms. Latencies preceding turns that began with *um* were also significantly longer than those preceding turns that began with *uh*, $F(1, 1196) = 16.38$, $p < 0.001$; mean difference 149ms. As expected, given the segmental difference, *ums* were also significantly longer than *uhs*, $F(1,3612) = 885.8$, $p < 0.0001$; mean difference 255ms.

In terms of prosodic differences between the two filled pauses, *ums* were significantly louder, had a greater intensity range, and lower minimum pitch than *uhs,* $F(1,3612) = 86.633$, $p<0.0001$ for maximum intensity, $F(1,3612) = 6.283$, $p = 0.012$ for mean intensity, $F(1, 3517) = 13.833$, $p = 0.0002$ for minimum intensity. In general, therefore, in our corpus *ums* are louder, longer, they tend to be preceded by longer latencies, and they are more likely to be followed by longer silent pauses than *uhs*.

Now turning to the relationship between filled pause type and deception, we found a significant correlation between filled pause type and local deception: *ums* correlated with lying, $r(3555) = –0.04$, $p = 0.023$. The correlation between filled pause type and global deception, however, tended in the opposite direction (lies correlated with *uhs)* but was not significant, $r(3303) = 0.03$, $p = 0.086$. The difference between the patterns for global and local lies may be attributed to the fact that most of the local lies were also classified as global lies but many local truths were not classified as global truths. Hence, there seems to be a tendency for *uhs* to occur in utterances that were locally truthful but the subjects were expressing a global lie.

3.3. Acoustic features of filled pauses and deception

In general, the factor of deception showed some effect on the prosodic features of filled pauses. When *um* and *uh* were pooled, filled pauses in global truths were longer than in lies, $F(1,3301) = 5.471$, $p = .019$. However, *uhs* were longer in local lies than truths, $F(1,1509) = 7.069$, $p = .008$. The data indicate that *ums* in deceptive speech are louder than in true statements. Maximum intensity in global lies was greater than in truths, $F(1, 1943) = 5.583$, $p = .018$. Furthermore, *ums* in a turn-internal position had significantly greater mean intensity in local lies than in truths, $F(1,1360) = 6.809$, $p = .009$.

Although the speaker-normalized mean and maximum F_0 values of filled pauses themselves did not correlate significantly with deception, several generalizations were observed in the subset of the corpus hand-corrected for F_0 targets. Most crucially, the degree of pitch reset of the filled pause correlated with local deception. The down-step from the preceding material into the filled pause as well as the up-step from the filled pause into the following material were greater in deceptive than in truthful utterances, $F(1, 263) = 11.02$, $p = 0.001$) and $F(1, 485) = 5.03$, $p = 0.025$ respectively.

In perceptually salient, turn-initial filled pauses this pattern is also observed: the up-step between the filled pause and the following material was greater for deceptive than for truthful speech, $F(1, 284) = 6.11$, $p = 0.014$. Interestingly, turn-initial filled pauses in these seven interviews had greater normalized mean and maximum pitch when they occurred in locally truthful than in locally deceptive speech, $F(1,284) = 4.91$, $p = 0.027$.

3.4. Machine learning experiments

To see whether the differences we have observed between filled pauses in deceptive and non-deceptive speech can provide useful predictors of deception, we next performed machine learning experiments on our corpus, using the static and dynamic features alone, and using them in conjunction with other potential predictors. We used three of the classifiers implemented in the WEKA software package ([23]): logistic regression, rule-induction (Ripper) and tree-generation (C-4.5). In the first experiment we extracted 65 prosodic features from those filled pauses that were labeled for local deception and that were longer than 30 ms and from their context. This

resulted in 3502 data points. The baseline error for this task, when we predict the majority class of local truth, is 36.1%. We divided the data 90%/10% into training and test sets respectively. The best result was achieved using a logistic regression learner that gave an error rate of 35.8%, a very small improvement over the baseline.

Following the observation in [12] that the best detection of deception is achieved with the combination of prosodic, lexical and subject-dependent features, in the second experiment we appended the filled pause features to the features extracted from all sentence-like units (SUs) of the CSC corpus. Out of the total of 9068 SUs in the corpus, 2730 contained at least one filled pause; SUs with more than one filled pause were assigned the features from the first filled pause. The baseline error for this task was 40.1%, predicting 'True' for each SU. The best result with all the features combined (filled-pause, acoustic, lexical, subject-dependent) was achieved with the C4.5 classifier that reduced the error to 32,2%. By comparison, the error of the same classifier in the experiment with the filled pause features omitted was 33.5%. Hence, the addition of the filled pause features resulted in an improved prediction of deception.

4. Discussion and Conclusions

Our data show that in general, the use of pauses correlates more with truthful than with deceptive speech. This was the case for both silent and vocalized pauses. Hence, this result supports the hypothesis that subjects monitor their speech more during lying than during truth-telling even though they did not have time to plan their deceptive utterances in advance. The assumption that the rate of pausing is greater in deceptive speech due to increased cognitive load associated with lying is not directly supported in our overall data. Yet, some support for this assumption was found in the relationship between deception and pause type. Local deception does correlate with the use of *um* more than with the use of *uh*, and *um* is longer, tends to be preceded by longer latencies and is surrounded by more silent pauses.

In terms of prosodic features, we found more cue value in loudness than in simple pitch related features. Our pilot results also suggest, however, that in addition to the static features, it is promising to investigate the prosodic relationship of filled pauses to that of surrounding material. Moreover, results from the machine learning suggest that the combination of static and dynamic features extracted from the filled pauses with other prosodic, lexical and subject-dependent features can improve results

Finally, we have found that speaker-dependent lexical habits such as the use of filled pauses or cue phrases (e.g. *now* or *well*) proved to be helpful in detecting deception in speech ([12]). Hence, our next step is to identify the most common cue phrases used by individual speakers and investigate the usefulness of static and dynamic features extracted from these phrases in detecting deception.

5. References

[1] Bailly, G.; Aubergé, V., 1997. Phonetic representation for intonation. In *Progress in Speech Synthesis,* J. Ph. van Santen (ed.). New York: Springer, 435-441.

[2] Bortfeld, H.; Leon, S.; Bloom, J.; Schober, M.; Brennan, S., 2001. Disfluency Rates in Conversation: Effects of Age, Relationship, Topic, Role, and Gender. *Language and Speech* 44(2), 123-147.

[3] Brennan, S.; Williams, M., 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.

[4] Clark, H. H., 1994. Managing problems in speaking. *Speech Communication* 15, 243-250.

[5] Clark, H.; Fox Tree, J., 2002. Using uh and um in spontaneous speech. *Cognition*, 84, 73–111.

[6] DePaulo, B. M.; Lindsay, J.; Malone, E.; Muhlenbruck, L.; Charlton, K.; Cooper, H., 2003. Cues to deception. *Psychological Bulletin*, 129(1):74–118.

[7] Ekman, P., 1992. Telling lies: clues to deceit in the marketplace, politics, and marriage. Norton, New York.

[8] Ferreira, F; Lau, F.; Bailey, K., 2004. Disfluencies, language comprehension, and Tree Adjoining Grammars. *Cognitive Science* 28(5), 721-749.

[9] Fox Tree, J., 2002. Interpreting Pauses and Ums at Turn Exchanges. *Discourse Processes* 34(1), 37-55.

[10] Fox Tree, J., 2001. Listeners' uses of um and uh in speech comprehension. *Memory and Cognition* 29, 320-326.

[11] Goldman-Eisler, F., 1968. *Psycholinguistics:Experiments in Spontaneous Speech.* New York: Academic Press.

[12] Hirschberg, J. et al., 2005. Distinguishing deceptive from non-deceptive speech. *Proceedings of 9th Interspeech*, 1833-6.

[13] House, D., 2003. Perceiving question intonation: the role of pre-focal pause and delayed focal peak. *Proceedings of 15th ICPhS*.

[14] Oomen, C.C.E.; Postma, A., 2001. Effects of divided attention on the production of filled pauses and repetitions. *Journal of Speech, Language, & Hearing Research*, 44, 997-1004.

[15] Reid, J. E.; and Associates. 2000. *The Reid Technique of Interviewing and Interrogation*. Reid, John E. and Associates, Inc., Chicago.

[16] Shriberg, E., 2001. To "Errrr" is Human: Ecology and Acoustics of Speech Disfluencies. *Journal of the International Phonetic Association* 31(1), 153-169.

[17] Shriberg, E.; Lickley, R., 1993. Intonation of clause-internal filled pauses. *Phonetica* 50, 172-179.

[18] Smith, V. L.; Clark, H. H., 1993. On the course of answering questions. *Journal of Memory and Language* 32, 25-38.

[19] Stenström, A., 1990. Pauses in monologue and dialogue. In *London-Lund Corpus of Spoken English: Description and Research,* J. Svartvik (ed.). Lund: Lund University Press.

[20] Swerts M., 1998. Filled pauses as markers of discourse structure. *Journal of pragmatics* 30, 485-496.

[21] Vrij, A.; Heaven, S., 1999. Vocal and verbal indicators of deception as a function of lie complexity. *Psychology, Crime, & Law*, 5, 203-315.

[22] Vrij, A. ; Winkel, F. W., 1991. Cultural patterns in Dutch and Surinam non-verbal behavior: Analysis of simulated police citizen encounters. *Journal of Nonverbal Behavior*, 15, 169–184.

[23] Witten, I. H.; Frank, E., 2005. Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann, San Francisco.

[24] Zuckerman, M.; Koestner, R.; Driver, R., 1981. Beliefs about cues associated with deception. *Journal of Nonverbal Behavior*, 6, 105-114.