Detecting Filled Pauses in Tutorial Dialogs

Gaurav Garg[1] **and Nigel Ward**

The University of Texas at El Paso

Technical Report UTEP-CS-06-32

July 28, 2006

Abstract

As dialog systems become more capable, users tend to talk more spontaneously and less formally. Spontaneous speech includes features which convey information about the user's state. In particular, filled pauses, such as *um* and *uh*, can indicate that the user is having trouble, wants more time, wants to hold the floor, or is uncertain. In this paper we present a first study of the acoustic characteristics of filled pauses in tutorial dialogs. We show that in this domain, as in other domains, filled pauses typically have flat pitch and fairly constant energy. We present a simple algorithm based on these features which detects filled pauses with 80% coverage and 67% accuracy. Analysis of the prediction failures shows that some are due to filled pauses of unusual types and related phenomena: filled pauses marking a change of state, cases where uncertainty is marked by lengthening a vowel in a word, and filled pauses which seque directly into a word.

[1]also with the Indian Institute of Technology, Kharagpur

[1]This research was sponsored in part by National Science Foundation Grant No. 0415150.

1 The Significance of Filled Pauses

Unlike read or laboratory speech, spontaneous and conversational speech contains high rate of disfluencies (Shriberg, 2001), including filled pauses, repetitions, repairs, and false starts. Filled pauses, our focus here, can be classified into unlexicalized types (e.g., *uh*, *um*) and lexicalized types (e.g., *well*, *like*) (O'Shaughnessy, 1992).

This report focuses on unlexicalized filled pauses as they appear in tutorial dialogs. Our aim is to build a tutorial system that is sensitive to and responsive to the user's internal state. For purposes of classifying the user's state, filled pauses provide important information. They can indicate that the user is having trouble recalling an answer or thinking of a word for something. Filled pauses are also signal that the user is not sure about the correct answer.

Our specific focus here is on recognizing user-produced filled pauses in memory-drill dialogs. In these dialogs the "learners" were given 3 sets of tasks, namely 1. Mention 8 consecutive exits on Interstate 10, 2. Mention 10 South American countries, and 3. name the 6 colleges of UTEP. The tutor then gave hints, to help reinforce the student's knowledge. These dialogs are representative of the dialogs seen in some forms of studying. Dialogs were collected with 16 subjects (12 male and 4 female), working with one exemplary tutor. There were a total of 106 minutes of dialog, containing 368 filled pauses. Details appear in (Hollingsed, 2006 in preparation).

The most common fillers in this corpus are *um* and *uh*.

2 Previous Work on Filled Pauses

In general, linguistic and psychological work on filled pauses has addressed issues such as their psychological and communicative significance O'Connell & Kowal (2005), and their linguistic and phonetic properties as acoustic objects Ward (2005 to appear). Computational work on filled pauses has focused mainly on the problem of detecting them: that is, given a speech signal, to automatically detect which sounds are filled pauses, rather than words. The is done primarily for the sake of improving the performance of automatic speech recognizers, since within-utterance filled pauses can cause significant difficulties for recognizers, which usually make no provision for them. Filled pauses may also be useful as an interface feature (Goto *et al.*, 1999).

The problem of detecting filled pauses has been addressed using various approaches, and using features of three types. One is the use of lexical information, in the form of a statistical language model. Second is the use of prosodic features, that is pitch, duration, energy. Third is the use of spectral information. It has been shown that lexical information is useful for detecting filled pauses in conversation, however we do not expect that it to be very useful for the learner's utterances in our tutorial dialogs, where multi-word utterances by the student are rare. Instead, students are largely just responding to questions and hints and volunteering answers. Spectral information, as a way to detect the absence of phonetic change over an extended span, is also useful in general, however lack of spectral change generally correlates highly with lack of change in energy, so we do not use that spectral information explicitly here.

Thus in this report we examine only prosodic features. In general, it is known that flat pitch and near-constant energy for an extended period of time are good indicators of filled pauses.

3 Preliminary Analysis

In the tutorial corpus filled pauses differ prosodically from other utterances. Firstly the duration of vowels in the filled pauses in tutorial system is generally longer; in effect the speaker utters them to fill silence. Hence the vowel is lengthened. Also the pitch or F0 is found to be stable in contrast to the pitch variations characteristic of spontaneous speech.

Based on this analysis, we recognized unlexicalized filled pauses (hereafter called "fillers" for short) by the following prosodic features:

1. The pitch of filler lies along the median of the pitch of the user across all his utterances.
2. A fillers has flat pitch.
3. The energy of a filler is stable or falling towards the end of their utterance.

4 The Algorithm

This section gives our prosody-only approach for detection of unlexicalized filled pauses in tutorial dialogs.

The algorithm works in four steps. It first identifies regions of speech and then applies filters to these regions to winnow out utterances with the properties mentioned above. In this section the algorithm itself is explained. The specific values used for the parameters are given in a later section.

4.1 Determination of Vowel Regions

In English, or at least in our data, filled pauses are vowels. As an energy threshold for determining whether a sound is a vowel or something else, we simply take the average of the energy for the track. Using this we discover regions where energy is continuously above the threshold. Figure 1 shows a typical filler, *uh* appearing at time 36050ms to 37000ms.

4.2 Pitch-Based Filtering

For each such region the pitch is examined. The pitch values used are the "interpolated pitch" values computed by the Didi suite, which in turn are computed using the algorithm of Hirose, Fujisaki and Seto (Hirose *et al.*, 1992). The reason for interpolation is that the pitch detection is noisy, but we know that the pitch of a user cannot change abruptly from one frame to next frame.

Our first filter on the pitch checks that the pitch of a region is in the mid-range, neither low or high. In these dialogs, the pitch of fillers is near the overall median pitch; perhaps this is because during the utterance of filler words the larynx tends to be in a relaxed state. Figure 1 shows that the pitch of a filler *uh* lies along the median pitch, which is represented by the horizontal dotted line. In our algorithm we specify an interval in which pitch should lie in order to qualify as a potential filler utterance; this is defined as lying within a certain percentage above/below the median pitch for that speaker. Since the pitch range of male speakers' fillers seems to be generally greater than that of female speaker, we use different values for men and women: `DEV_PITCH_MALE` and `DEV_PITCH_FEMALE`, respectively.

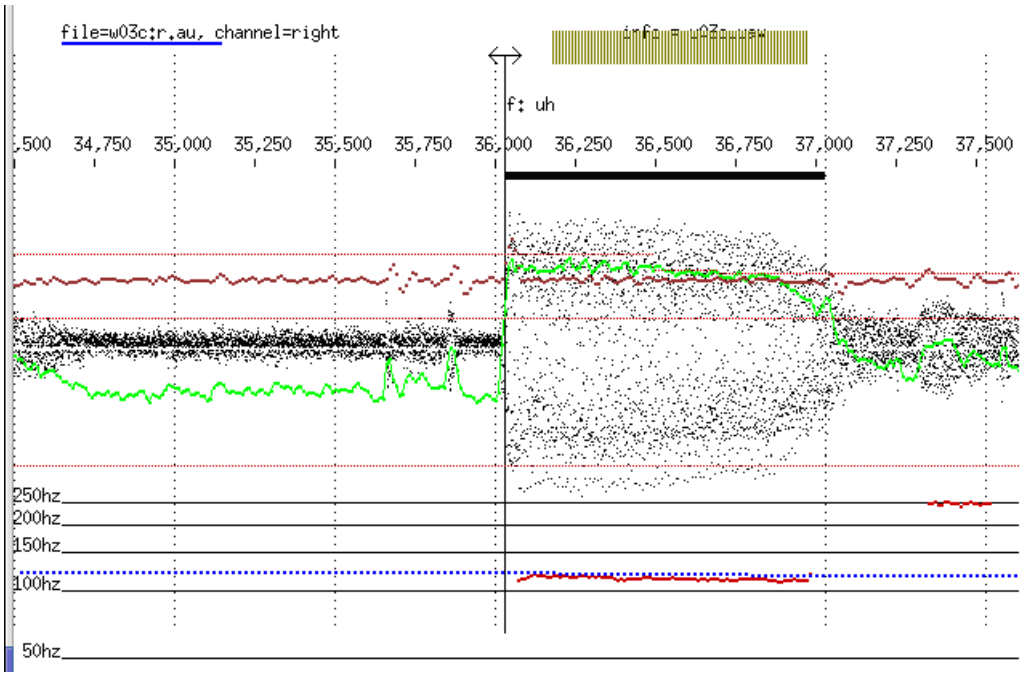

Figure 1: A Typical Filter as Seen in Didi (Ward & Tsukahara 2000).

Our second filter on the pitch checks that the pitch is flat throughout the region. This is done by a constraint on the slope, where the slope is calculated over the previous 200 ms. Specifically, the slope of the pitch must always lie between `MIN_PITCH_SLOPE_THRESHOLD` and `MAX_PITCH_SLOPE_THRESHOLD`.

4.3 Energy-Based Filtering

At the start of a filler the energy rises quickly and then remains stable or falls steadily (see Figure 2) through the utterance. This is detected by measuring the slope of the log energy between every pair of adjacent 10 ms frames. This must not be rising in the body of the filler. Specifically, the algorithm checks that the change in log energy is less than `MAX_ENERGY_SLOPE` everywhere in the last half of the utterance.

4.4 Merging Adjacent Fillers

After applying the above filters, regions having a duration greater than `MIN_FILLER_DURATION` are marked as fillers. Since fillers are utterances used by speaker to fill silence and to indicate that he/she is still thinking, two fillers uttered at short interval signal the same aspect of user feeling. So two fillers which are less than `MIN_GAP_BTW_FILLERS` apart are merged into one filler.

5 Implementation

This algorithm is implemented as a part of the `didi` suite in `/home/users/guarav/aiz suite`. To invoke the algorithm to detect the fillers, `dede` should be run with the `-dfl` or `-dfr` option, depending upon whether the user utterances are in the left track or in the right track (where dfl

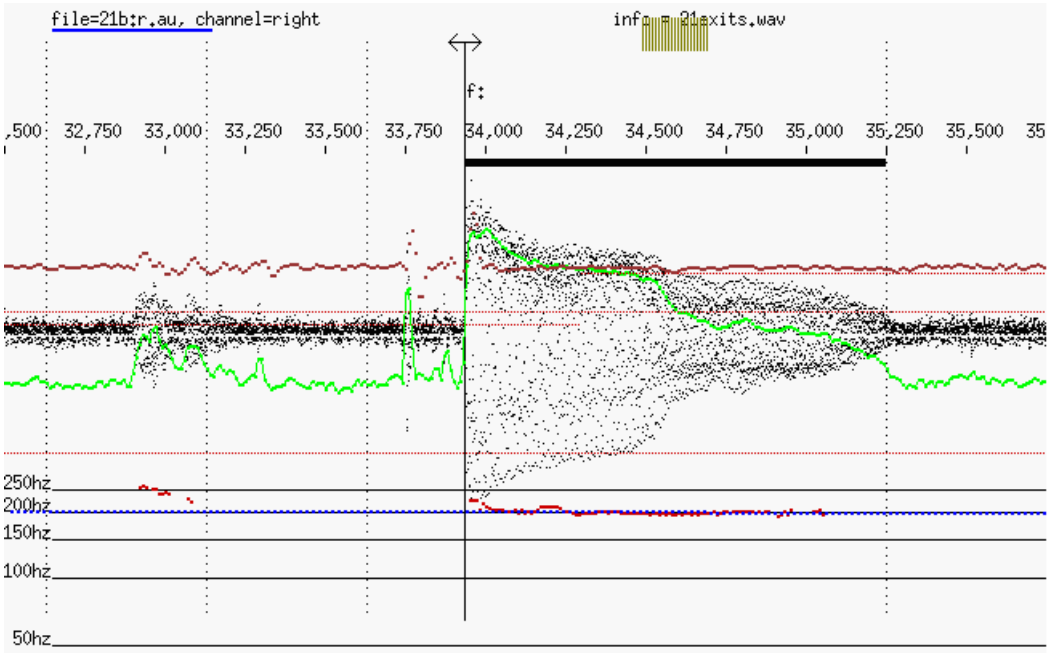

Figure 2: A Filler, *am*, with Decreasing Energy

Parameter	Value
DEV_PITCH_MALE	12% of median pitch in Hz
DEV_PITCH_FEMALE	10% of median pitch in Hz
MIN_PITCH_SLOPE_THRESHOLD	- 0.1% of median pitch/ms
MAX_PITCH_SLOPE_THRESHOLD	+0.1% of median pitch/ms
MIN_GAP_BTW_FILLERS	500 ms
MIN_FILLER_DURATION	200 ms
MAX_ENERGY_SLOPE	0.10 log energy/ms

Table 1: Parameter Values

stands for Detecting Fillers in Left track and dfr stands for Detecting Fillers in Right track). The output shows all the detected fillers, giving their start and end times in milliseconds, and also the coverage and the accuracy for each speech file.

6 Parameter Values

The values of the parameters have to be set appropriately. We tuned them to maximize performance in the corpus. The measure of performance is coverage and accuracy. The parameter values giving the best performance are shown in Table 1, where:

DEV_PITCH_MALE is the percentage of median pitch for male speakers to determine the upper and lower bounds of the range of pitch allowed of fillers. Increasing this range increases the coverage but decreases the accuracy.

DEV_PITCH_FEMALE is the same, but for female speakers.

`MIN_PITCH_SLOPE_THRESHOLD` is the lowest allowable pitch decrease. Increasing the threshold increases the coverage but decreases the accuracy.

`MAX_PITCH_SLOPE_THRESHOLD` is, similarly the maximum pitch rise.

`MIN_GAP_BTW_FILLERS` is maximum time gap between two fillers for them to be merged into one.

`MIN_FILLER_DURATION` is the minimum duration of a region to be accepted as filler.

7 Results

To evaluate it, the algorithm was run on the dataset consisting of all 106 minutes of tutorial dialog, and performance compared to the 368 fillers, which were manually labeled by Tasha Hollingsed. Coverage obtained was 80.2% with accuracy of 67.6%.

8 Error Analysis

Although there were some clear errors, there were also cases which were counted as such but which may not have been, since the labeling of fillers in the corpus was done without reference to any definition or using clear criteria.

8.1 Type I Errors

There were two major reasons why the algorithm sometimes failed to identify a filler.

First, in some cases the filler had an expressive intonation contour, which was not flat or did not lie along the median. Figure 3 shows such a filler, *um*.

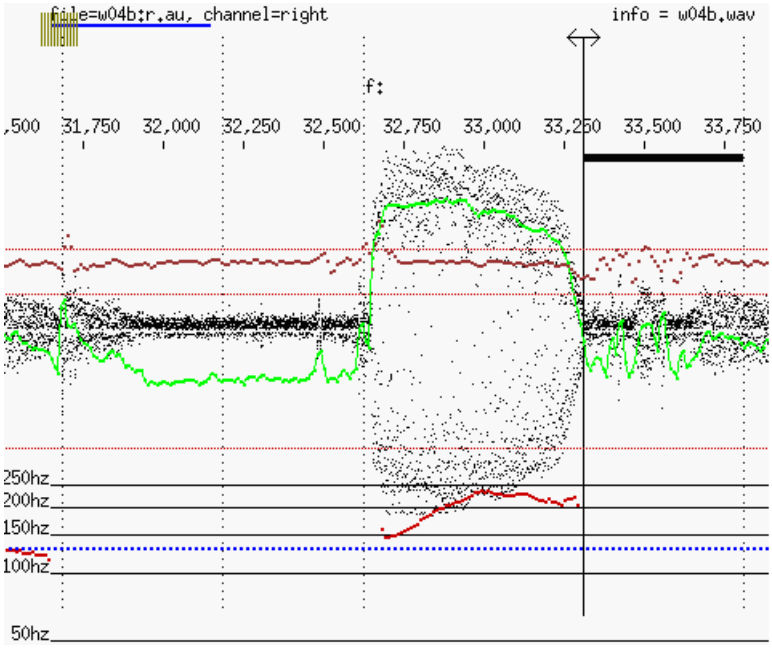

Figure 3: A Filler with a Non-flat Intonation

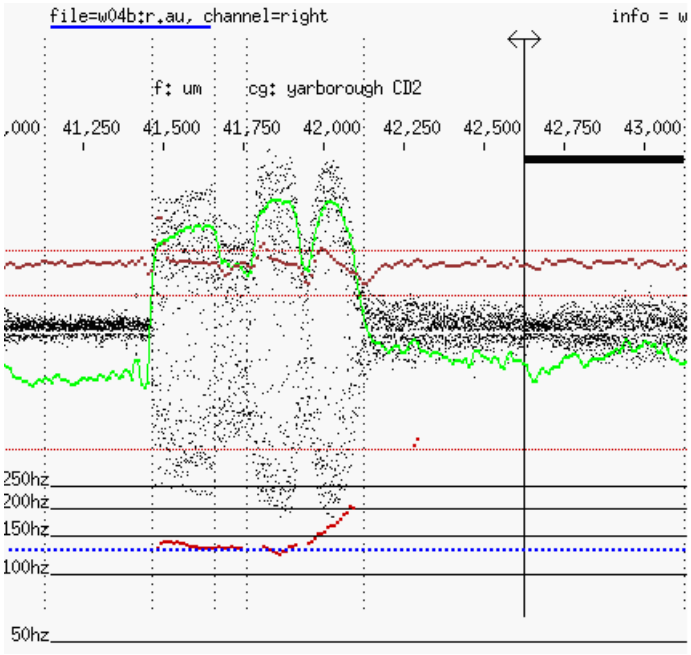

Figure 4: A Short Filler

Second, in some cases the filler duration was smaller than `MIN_FILLER_DURATION`. Figure 4 shows such a short *um*.

8.2 Type II Errors

There were two major reasons why the algorithm sometimes identified as a filler a speech segment that was not labeled as such.

First, there were cases where the vowel at the start of the word was elongated. For example, in the utterance *Lomaland*, the vowel /o/ was elongated by the user.

Second, vowels towards the end of words were also sometimes elongated. For example, in the utterance *Paraguay*, the vowel /u/ was elongated by the user.

References

Goto, Masataka, Katunobu Itou, & Satoru Hayamizu (1999). A Real-time Filled Pause Detection System for Spontaneous Speech Recognition. In *Eurospeech '99*, pp. 227–230.

Hirose, Keikichie, Hiroya Fujisaki, & Shigenobu Seto (1992). A scheme for pitch extraction of speech using autocorrelation function with frame length proportional to the time lag. In *1992 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. I–149–152.

Hollingsed, Tasha (2006, in preparation). The Value of Being Responsive in Spoken Tutorial Dialogs. UTEP Computer Science Department Masters Thesis.

O'Connell, Daniel C. & Sabine Kowal (2005). *Uh* and *Um* Revisited: Are They Interjections for Signaling Delay? *Journal of Psycholinguistic Research*, 34:555–576.

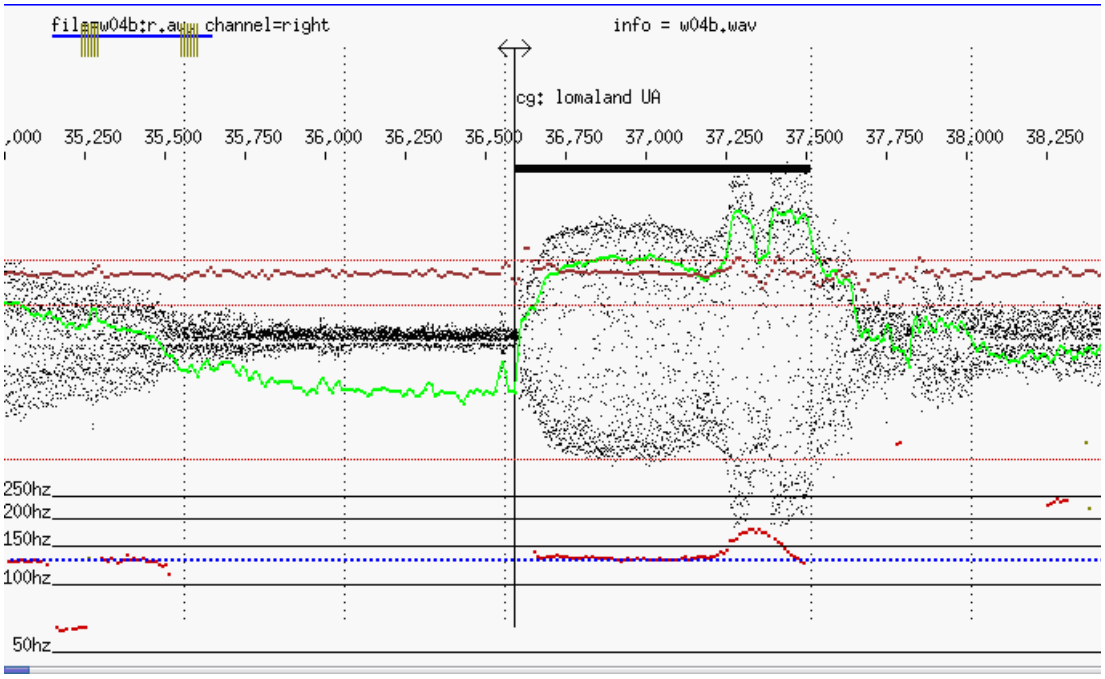

Figure 5: Detecting the Initial Vowel of the Word *Lomaland* as a Filler

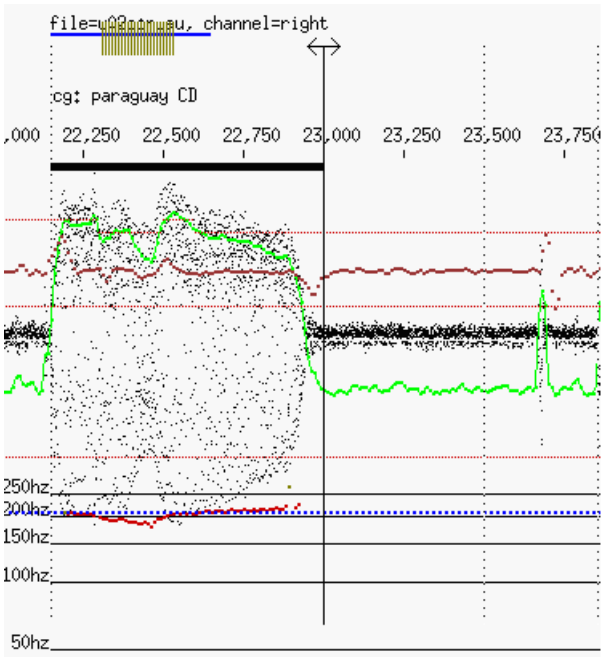

Figure 6: Detecting the Last Syllable of the Word *Paraguay* as a Filler

O'Shaughnessy, Douglas (1992). Recognition of Hesitations in Spontaneous Speech. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. I–521–524.

Shriberg, Elizabeth (2001). To 'errr' is Human: Ecology and acoustics of speech disfluencies. *Journal of the International Phonetic Association*, 31:153–169.

Ward, Nigel (2005, to appear). Non-Lexical Conversational Sounds in American English. *Pragmatics and Cognition*.

Ward, Nigel & Wataru Tsukahara (2000). Prosodic Features which Cue Back-Channel Feedback in English and Japanese. *Journal of Pragmatics*, 32:1177–1207.