FORMANT-BASED TECHNIQUE FOR AUTOMATIC FILLED-PAUSE DETECTION IN SPONTANEOUS SPOKEN ENGLISH

Kartik Audhkhasi[*]

Dept of Electrical Engineering
Univ of Southern California
Los Angeles, CA, USA
kartik.audhkhasi@usc.edu

Kundan Kandhway, Om D. Deshmukh, Ashish Verma

IBM India Research Lab
Vasant Kunj Institutional Area
New Delhi, India
{kkandhwa,odeshmuk,vashish}@in.ibm.com

ABSTRACT

Detection of filled pauses is a challenging research problem which has several practical applications. It can be used to evaluate the spoken fluency skills of the speaker, to improve the performance of automatic speech recognition systems or to predict the mental state of the speaker. This paper presents an algorithm for filled pause detection that is based on the premise that the vocal tract characteristics, and hence the formants, are stable during the production of a filled pause. The performance of the proposed algorithm is evaluated on real-life recordings of call center agents where the locations of the filled pauses are hand labeled. The proposed algorithm outperforms a standard cepstral stability based filled pause detection algorithm and a standard pitch-based detection technique.

Index Terms— Filled pause, vowel lengthening, fluency evaluation, spectral features

1. INTRODUCTION

Spontaneous speech, by definition, is not prepared well in advance. The speaker is expected to formulate his/her thoughts on the fly. The lack of preparation and/or inadequate knowledge of the language leads to several disfluencies in the output speech signal. Structurally, the disfluent speech can be split into three components [1]: the reparandum, the edit phrase and the alteration. The reparandum is the part of the speech that the speaker intends to replace. The edit phrase is the region between the reparandum and the beginning of the replacement of the reparandum. The edit phrase typically consists of unfilled or filled pauses (e.g., 'ahh', 'umm') or discourse markers (e.g., 'like', 'you know'). The alteration marks the resumption of fluency. Removal of the reparandum and the edit phrase restores the fluency in the spoken utterance. The disfluencies can be categorized in several classes based on the relative composition of these three components. Some of the common types of disfluencies are: repetitions (the reparandum and the alteration contain the same set of one or more words), substitutions, deletions, insertions and fresh starts. In majority of disfluent speech regions, either the edit phrase consists of a filled pause (e.g., 'ahh', 'umm', 'ohh') and/or the final vowel of the reparandum is lengthened. Authors in [2] report that, in the conversational Switchboard database, about 39.7% of the disfluencies contain a filled pause. Thus, detection of vowel lengthening and filled pauses is an important step towards locating the disfluent regions and evaluating the overall spoken fluency skills of a speaker. Moreover, the performance of the Automatic Speech Recognition (ASR) systems can be adversely affected by the presence of filled pauses in speech signals [3].

Previous efforts in detecting filled pauses were focused towards improving the performance of ASR systems in spontaneous speech. The method proposed in [3] splits the speech signal in segments based on a cepstral difference function. Various acoustic and prosodic features like segment duration, cepstrum-based spectral stability measure and spectral center of gravity are computed at the segment level to classify each segment as either filled pause or non-filled pause. This algorithm is used as a preprocessor to remove the detected filled pauses from the speech signal before it is passed to the ASR system. Authors in [4] propose a decision-tree model to discriminate boundaries following a filled pause from all the other word boundaries. The boundary information is obtained from automatic forced-alignments on the speech data. The features include the slope of fundamental frequency (F0) before and after the boundary, the difference of F0 across the boundary and other duration and voicing based features. Authors in [5] detect filled pauses by locating regions where the F0 transition and the spectral envelope deformation are small. The F0 transition and spectral envelope deformation are defined as the slopes of the straight lines that optimally fit the estimates of F0 and the spectral envelope respectively. This algorithm is used to add more interactive functions to a desktop ASR system [6].

Findings in [7] suggest that the intonation patterns in filled pauses are related to the intonation patterns in the speech just preceding the filled pause and hence the variations in the F0 value over a filled pause are related to the F0 of the surrounding speech. Thus, the F0 variations over a filled pause need not be small and detecting filled pauses using F0-based features warrants analyzing the F0 variation over a longer region.

This paper presents a filled pause detection method that directly tries to capture one of the robust and unique characteristics of filled pauses: the relative stability of the vocal tract shape during the production of filled pauses.

2. PROPOSED FILLED-PAUSE DETECTION METHOD

In spontaneous speech, there is sometimes a time lag between the end of the current spoken sentence and the beginning of the next one. Many speakers fill this gap by extending the last vowel (e.g., 'theee') and/or by using filled pauses like 'ahh's and 'umm's while the next thought or the next set of words is chosen. As the next word is still to be chosen, there is minimal coarticulation effect and the articulators do not change their positions during the filled pauses (authors

[*]The work was performed during the summer internship at IBM India Research Lab.

in [6] also make similar observations). As a result, the vocal tract characteristics do not change during the production of filled pauses. This translates in the vocal tract resonances, i.e. the formants, remaining stable over the entire duration of the filled pause. Fig. 1(a) show the spectrogram along with the first two formants of the utterance 'I bought ahh'. Note that the formants are stable over the filled pause duration (0.6 - 1.0 sec) whereas they vary considerably during normal speech (0.05 - 0.22 sec and 0.3 - 0.4 sec). The proposed algorithm utilizes this unique characteristic of the filled pauses and vowel lengthening (jointly referred to as filled pauses in the rest of the paper) to automatically detect them in spontaneous speech. The analysis in the algorithm is restricted to the first two formants mainly because they capture the stability adequately and the higher formants are more difficult to track accurately.

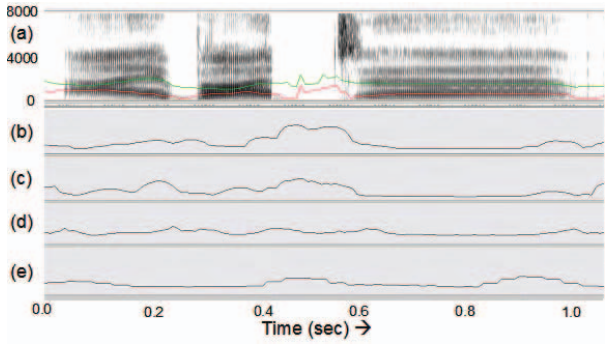

Fig. 1. *(a) Spectrogram of the utterance 'I bought ahh'; The first two formants are overlaid. The two normal-speech vowels are between 0.05-0.22 sec and 0.3-0.4 sec and the filled pauses is between 0.6-1.0 sec. (b) F1SD; (c) F2SD; (d) Cepstral instability measure C_i; (e) F0SD*

The various steps involved in the proposed algorithm are shown in Fig. 2. The speech signal is passed through a simple energy-based Voice Activity Detector (VAD) to identify silent and low energy regions. These regions are excluded from the rest of the analysis. The energy threshold is set in such a way that the weak fricatives are also excluded from the analysis. Spectral characteristics of filled pauses almost never resemble those of weak fricatives and the formant tracks in weak fricatives are also quite unreliable.

In all the experiments presented here, the wavesurfer [8] formant tracker is used to compute the formants. The formants are computed at a frame rate of 10 ms. The stability of each formant is analyzed separately. The stability of a formant at a given frame is quantified by computing the Standard Deviation (SD) of the formant value over a window of W frames centered on the current frame. As the window length is reduced, more normal-speech frames exhibit less SD increasing the rate of false alarms. For longer windows, the SD of filled-pause frames increases which leads to more misses. Fig. 3 compares the distribution of the SD of the first formant (F1SD) for filled-pause frames and normal-speech frames in the training data. The distribution of F1SD for frames corresponding to normal speech has a long tail that extends over much higher values ($>$ 190 Hz). The distributions plotted in the figure are truncated at 190 to highlight the main differences in the two distributions at the lower SD values. Note that the F1SD for filled pause frames is much lower than that for normal speech frames. For example, about 78.7% of the filled pause frames have F1SD below 40 Hz while only about 19.5% of the normal speech frames have F1SD below 40 Hz. Similar trends in

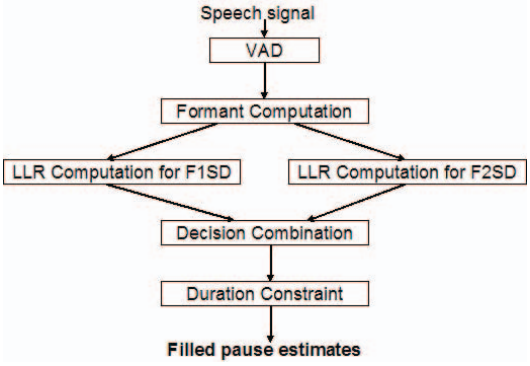

Fig. 2. *Flow chart of the proposed filled pause detection algorithm.*

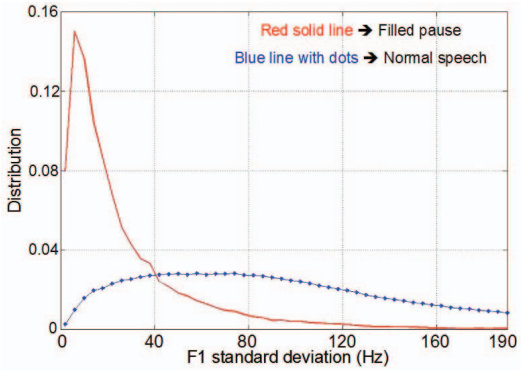

Fig. 3. *Comparison of the distribution of the standard deviation for frames in filled pauses (red curve) and for frames in normal speech (blue-solid-dots line).*

distributions are observed for the SD of the second formant (F2SD): i.e., the F2SD values for frames corresponding to filled pauses are typically much lower than the F2SD values corresponding to frames from normal speech.

The distributions of normal-speech F1SD and filled-pause F1SD are discretized using 51 bins. 50 equi-width bins cover the $[0-200]$ range and the last bin covers all the SD values above 200. This discretization leads to 51-point probability mass functions (p.m.f.), $f^{FP}_{F1SD}(i)$ and $f^{NS}_{F1SD}(i), i = 1 \ldots 51$, of F1SD for filled-pause and normal-speech frames respectively. If the observation, x, lies in bin i, then the Log Likelihood Ratio (LLR) is computed as follows:

$$L_{F1SD}(x) = log\left[\frac{f^{FP}_{F1SD}(i)}{f^{NS}_{F1SD}(i)}\right] \quad (1)$$

If $L_{F1SD}(x)$ is greater than a threshold τ_{F1}, then the observation x which lies in bin i, is said to have come from a filled pause frame. Otherwise, the observation is said to have come from a normal-speech frame. The $L_{F1SD}(x)$ value can also be used as a decision confidence measure for soft-decision analysis. The relative values of precision and recall can be varied systematically by varying the threshold τ_{F1}. Similar LLR computations are performed for F2SD.

During evaluation, for a given frame, the F1SD and F2SD are computed as mentioned above and the corresponding LLR bins are identified. The frame is marked as likely-filled-pause if L_{F1SD} and L_{F2SD} are greater than the corresponding thresholds τ_{F1} and τ_{F2}

respectively. Thus, the stability decisions from F1 and F2 are combined using a simple frame-wise logical AND operation.

In the next step, a continuity constraint is used to remove short regions from being wrongly identified as filled-pauses. All the regions where the number of contiguous likely-filled-pause frames is less than an optimal threshold D are identified and the corresponding frames are marked as normal speech frames. The remaining likely-filled-pause frames are the final detected filled pauses. This continuity constraint greatly reduces the number of false alarms. The threshold D can be adjusted to vary the false alarms and misses (i.e., precision and recall). All the above thresholds were chosen by optimizing the filled-pause/normal-speech separation on the training data.

The next section describes our implementation of two standard filled pause detection techniques presented in literature.

3. STANDARD FILLED PAUSE DETECTION TECHNIQUES

3.1. Cepstral variation based technique

To compare the performance of formants with that of other spectral features in capturing the spectral stability, a filled pause detection technique based on cepstral variation [3] is implemented. The cepstral features used are the 13 Mel Frequency Cepstral Coefficients (MFCCs) and are computed using HTK [9]. The cepstral instability, C_i, at frame i is computed as:

$$C_i = \frac{1}{||\underline{M_i}||} \sum_{j=i-\frac{N}{2}}^{i+\frac{N}{2}} ||\underline{M_j} - \underline{\mu_i}|| \quad (2)$$

$$\underline{\mu_i} = \frac{1}{N+1} \sum_{j=i-\frac{N}{2}}^{i+\frac{N}{2}} \underline{M_j}$$

where, $\underline{M_i}$ is the MFCC vector computed at frame i, $\underline{\mu_i}$ is the mean of the MFCC vectors corresponding to the $\pm N/2$ frames adjacent to the ith frame and $||.||$ is the Euclidean norm. Thus, C_i, the cepstral instability at frame i can be defined as the normalized average of the individual Euclidean distances between the MFCC vectors corresponding to frames that are at the most $\pm N/2$ frames away from the current frame and the mean of these vectors. Frames corresponding to filled pauses exhibit lower C_i values as compared to those for normal speech frames. The distribution of the C_i values for filled pause frames and normal speech frames is discretized using 51 bins. 50 equi-width bins span the $[0-1.5]$ range and one bin is used for C_i values greater than 1.5. The rest of the training and evaluation steps are identical to those used in the proposed technique (ref section 2).

3.2. Pitch-based technique

It has been observed [6] that filled pauses are typically spoken with minimal intonation implying that the F0 remains almost flat during the filled pauses. The present implementation of the pitch-based filled pause detection technique identifies filled pause regions by detecting regions with minimal F0 variations.

The pitch tracker provided by wavesurfer is used to estimate the F0 values. The stability of F0 at a given frame is quantified by computing the standard deviation of the F0 values (F0SD) over a window of 6 frames centered on the current frame. The rest of the training and evaluation steps are identical to those used by the proposed formant based technique. A window of 6 frames was found to be optimal for pitch-based discrimination between filled-pause frames and normal-speech frames.

4. DATABASE

The proposed algorithm and the other two standard algorithms are trained and evaluated using real life recordings of 96 candidates who were interviewed for call center agent positions at IBM Daksh's Gurgaon India call center facility. Each candidate was asked to speak for about one minute on one of the topics from a pre-selected set. These recordings were made as part of the larger project to automatically evaluate spoken English skills of the speakers [10]. Of these 96 recordings, 50 recordings are randomly chosen to form the training data. The rest of the recordings are used as the test data. The start and the end frames of the filled pauses in this data are hand labeled. Every label was cross-checked by one labeler. Each filled pause in the test data is further categorized as *robust* or *non-robust*. Robustness of a filled pause is identified by its perceptual prominence which includes factors like energy during the filled pause, duration of the filled pause and its proximity to a silence region. The robust filled pauses in the entire database were identified by only one labeler to maintain consistency across the database. [1]

The training database consists of 538 filled pauses. The number of filled pauses per speaker varies from 3 to 29 with an average of 10.8 filled pauses per speaker. The average duration of filled pauses is 263.4 ms (minimum and maximum durations of the filled pauses are 85 and 784 ms, respectively). 77.3% of the filled pauses are longer than 200 ms and 99.6% are longer than 100 ms. Only 55.0% of the filled pauses had silence in the immediate neighbourhood prompting the authors to not use the 'silence-border' criterion used by other researchers [3, 11] to detect filled pauses. The 46 recordings in the test data contain 484 filled pauses of which 192 were labeled as robust.

As mentioned earlier, the various thresholds used in the three algorithms were optimized on the training data. The optimal values of the threshold are: $W = 11$, $N = 10$, $\tau_{F1} = 0$, and $\tau_{F2} = 0$.

5. RESULTS

Fig. 1(a) shows the spectrogram of the utterance 'I bought ahh'. F1SD and F2SD, which capture the variations in the first and the second formant respectively, are plotted in Fig. 1(b) and 1(c) respectively. F1SD and F2SD are low for filled pauses frames and relatively high for normal speech frames. Fig. 1(d) plots the cepstral instability (C_i) measure for the utterance. The C_i measure is low for filled pauses but the difference between the C_i values for filled pause frames and the C_i values for the normal-speech frames is not as pronounced as the difference between the F1SD (or the F2SD) values for the two classes. Similar observations can be made about the F0SD contour which is plotted in Fig. 1(e).

The evaluation criteria used to compare the performance of the three filled-pause detection techniques are recall (ratio of total number of correctly detected filled pauses and total number of filled pauses) and precision (ratio of total number of correctly detected filled pauses and total number of detected filled pauses). Table 1 presents the variation in the precision and recall values for all the

[1] To encourage future evaluations of filled pause detection techniques on a common database, authors would be happy to share this database and the labels with other researchers in this area.

Table 1. *Precision (and recall) values for the three methods as the duration threshold is varied.*

Duration threshold (D)	Proposed method	Cepstral stability method (based on [3])	Pitch-based method (based on [6])
10	0.64 (0.73)	0.44 (0.68)	0.38 (0.54)
11	0.68 (0.67)	0.50 (0.64)	0.40 (0.47)
12	0.73 (0.63)	0.54 (0.58)	0.42 (0.43)
13	0.77 (0.59)	0.60 (0.55)	0.46 (0.39)
14	0.81 (0.55)	0.65 (0.51)	0.48 (0.35)
15	0.84 (0.50)	0.68 (0.47)	0.52 (0.32)
16	0.86 (0.45)	0.75 (0.43)	0.55 (0.28)

three methods as the duration threshold is varied. Several important observations can be made from the table. The performance of the proposed technique is considerably better than that of the other two methods. For all the three methods, as the duration threshold is increased, the precision increases at the expense of the recall. The performance of the pitch-based technique is the lowest which confirms the findings in [7] that the filled pauses need not exhibit a flat intonation pattern. The precision-recall curves for all the three techniques are plotted in Fig. 4. It can be deduced from the figure that for a given precision, the proposed technique leads to the highest recall as compared to the other two methods and for a given recall value the proposed technique leads to the highest precision.

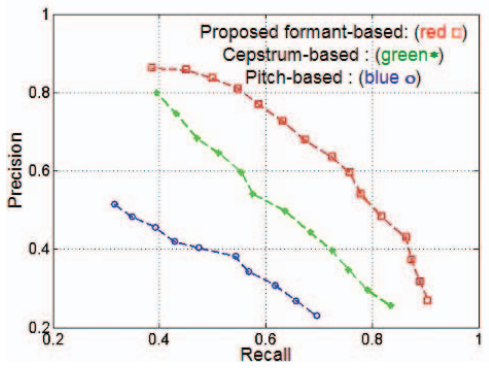

Fig. 4. *Precision-Recall curve for the proposed technique (red squares), the cepstrum-based method (green filled-circles) and the pitch-based method (blue open-circles) as the duration threshold (D) is varied.*

The performance of the individual formants in detecting the filled pauses was also analyzed. It was observed that the second formant typically leads to lesser false alarms (i.e., higher precision) as compared to the first formant. This behaviour is expected as the second formant is known to be a good indicator of the place of articulation and thus more affected by coarticulation [12]. The combination of the two formants leads to a much better performance than the individual performance of either of the formants.

As mentioned earlier, the main goal of the proposed technique for filled pause detection is to automatically quantify the spoken fluency skills of a speaker. In such cases, detection of the perceptually *robust* filled pauses is more crucial than false alarms or deletions of non-robust filled pauses. At the optimal duration of $D = 12$, the proposed technique is able to detect 85% of the *robust* filled pauses. The corresponding numbers for the cepstrum-based method and the pitch-based method are 79% and 58%, respectively.

Our analysis of the errors shows that the missed filled pauses can be categorized in one of the following three cases: (a) Some speakers tend to lower there volume during a filled pause. In such cases it is difficult to accurately track the formants which results in missing the filled pause, (b) Filled pauses, sandwiched between two words, are typically co-articulated and lack formant stability and (c) The filled pauses are short and are eliminated by the duration threshold. False alarms typically include (a) tense vowels like in 'too' and 'ya', and (b) cases where similar sounding vowels occur in succession leading to a similar formant structure over a long duration.

6. DISCUSSION AND FUTURE WORK

This paper presents a technique for automatic detection of filled pauses. The proposed technique is based on the premise that the vocal tract characteristics, and hence the formant frequencies, remain stable during the production of filled pauses. It is shown that the performance of the proposed technique is better than that of cepstral-stability based and pitch-based techniques. The details of the use of the proposed technique for the evaluation of fluency of spoken English skills of a speaker are presented in [10]. The current effort is focused on combining the formant-based features with cepstral-based and pitch-based features to improve the overall accuracy of filled pause detection.

7. REFERENCES

[1] W. J. M. Levelt, "Monitoring and self-repair in speech," in *Cognition*, 1983, pp. 41–104.

[2] A. et. al. Stolcke, "Automatic detection of sentence boundaries and disfluencies based on rcognized words," *Proc. ICSLP*, pp. 2247–2250, 1998.

[3] F. Stouten and J-P Martens, "A feature-based filled pause detection technique for dutch," in *IEEE Intl Workshop on ASRU*, 2003, pp. 309–314.

[4] E. Shriberg, R. Bates, and A. Stolcke, "A prosody-only decision-tree model for disfluency detection," in *Proceedings of Eurospeech*, Rhodes, Greece, 1999, pp. 2383–2386.

[5] Masataka Goto et. al., "A real-time filled pause detection system for spontaneous speech recognition," in *Proc. of Eurospeech*, September 1999, pp. 227–230.

[6] Masataka Goto et. al., "Speech interface exploiting intentionally-controlled nonverbal speech information," in *Proc. of ACM Sypmosium on UIST*, October 2005, pp. 35–36.

[7] E. Shriberg, "Phonetic consequences of speech disfluency," in *Int. Conf. on Phonetics Sciences*, San Francisco, 1999, pp. 619–622.

[8] Wavesurfer: An open source speech tool, "http://www.speech.kth.se/wavesurfer/," .

[9] The HTK book, "http://htk.eng.cam.ac.uk," 2002.

[10] O. Deshmukh et. al., "Automatic evaluation of spoken english fluency," in *submitted to Int. Conf. on Acoustics, Speech, and Signal Processing*, 2009.

[11] M. Gabrea and D. O'Shaughnessy, "Detection of filled pauses in spontaneous conversational speech," in *Proceedings of ICSLP*, Beijing, 2000, pp. 678–681.

[12] K. Stevens, *Acoustic Phonetics*, M.I.T. Press, 1999.