Buletinul Ştiinţific al Universităţii "Politehnica" din Timişoara

Seria ELECTRONICĂ şi TELECOMUNICAŢII
TRANSACTIONS on ELECTRONICS and COMMUNICATIONS

Tom 53(67), Fascicola 2, 2008

Robust Detection of Filled Pauses in Spontaneous Conversational Speech

Marcel Gabrea[1]

Abstract – Most automatic speech recognition work has concentrated on read speech, whose acoustic aspects differ significantly from speech found in actual dialogues. A primary difference between read speech and spontaneous speech concerns a high rate of disfluencies (e.g., filled pauses, repetitions, repairs, false starts). Filled pauses (e.g., "uh," "um"), unlike silences, resemble phones as part of words in continuous speech. In this paper the problem of detection of filled pauses in spontaneous speech and how this can be useful in automatic speech recognition are considered. The acoustic aspects of filled pauses in a widely-used SWITCHBOARD database are examined here, from the point of view of identifying them acoustically using a combination of duration, fundamental frequency and spectra.

Keywords: automatic speech recognition, conversational speech, SWITCHBOARD, disfluencies, filled pauses.

I. INTRODUCTION

The growing advance of automatic speech recognition has addressed the increasing need for studying variabilities in the behavior for a large multispeaker database of telephone bandwidth spontaneous speech. Spontaneous or conversational speech differs from read speech in several ways, the most obvious difference concerning hesitation phenomena. Spontaneous speech is punctuated with and interrupted by a wide variety of seemingly meaningless words (e.g., "uh," "um") as well as false starts, silent pauses and lengthened words.

Restarts are interruptions in the flow of speech, where the speaker reiterates a portion of the speech immediately preceding, with or without a change. They consist of those instances in which a speaker begins an utterance and then restarts the utterance (there may or may not be a pause before the restart) or in which an utterance is begun and then abandoned. When an utterance is begun and then abandoned, it is generally followed by a pause which may then be followed be a new utterance or a complete stop in the conversation.

Pauses are simple interruptions in the flow of speech, where a significant delay occurs in the delivery of the speech, and they are ever-present.

Pauses can be subdivided into either filled pauses or unfilled pauses. Filled pauses may be categorized as either unlexicalized (e.g., "uh," "um") or lexicalized (e.g., "well", "like", "you know"). The specific interruption phenomena studied here are unlexicalized filled pauses.

A primary application of this study lies in improving the performance of automatic speech recognizers, for applications that must accept an input of spontaneous speech (e.g., verbal conversations with computer databases). Within-utterance filled pauses can cause significant difficulties for automatic speech recognizers, which usually make no prevision for them at random locations and cause difficulties in having a proper interpretation in the language-model component. Automatically locating filled pauses could help automatic recognizers avoid textual errors in the output. For such purposes, we wish to eliminate filled pauses so that the recognizer will operate on only a sequence of desired words.

Acoustical analyses of disfluencies with a view toward speech recognizers have only been done in the last few years [3] - [14].

A research group at SRI tried to automatically locate filled pauses [7] - [10]. It appears that their work has not examined direct filled-pause detection from speech, but rather by augmenting more general ASR methods. By using SWITCHBOARD conversations they assumed knowledge of the word boundaries and examined detection of filled pauses through the use of a more general ASR system, including relevant language models. The model is based on a generalization of the standard N-gram language model.

In [11] the different functions of filled pauses are investigated and they show that by modeling them appropriately can lower the perplexity of the neighboring words. Possible models include extending and reducing the N-gram, using a class grammar in conjunction with word N-gram, and removing the disfluency markers from the word history.

For the work presented in [12], acoustic and prosodical cues to self-repairs are identified, based on

[1] École de technologie supérieure, Département de génie électrique,
1200 Notre-Dame Ouest, H3C 1K3 Montréal,
e-mail marcel.gabrea@etsmtl.ca

an analysis of the ATIS database and methods are proposed for exploiting these cues for repair detection and correction. These are examined in a statistical model of repair site detection and a prosodically labeled corpus of repair utterances was used.

In [13] the author suggests using the juncture phenomena as cues to the early detection of disfluency by listeners. His work is based on the occurrence of juncture phenomena between words in fluent speech, which are usually absent at the interruption point in disfluent utterances.

In [14] a review the acoustic and linguistic properties of children's speech are presented. The verbal child-machine spontaneous interaction is reviewed and results from recent studies are presented. Age trends of acoustic, linguistic and interaction parameters are discussed, such as sentence duration, filled pauses, politeness and frustration markers, and modality usage. The implications for acoustic modeling, linguistic modeling and spoken dialogue systems design for children are discussed.

The algorithm presented in this paper was developed by examination and analysis of many utterances from different speakers. A large database of spontaneous speech (i.e., SWITCHBOARD) was analyzed in terms of duration and fundamental frequency. All phases of the task used automatic F0, energy, duration, and spectral estimation directly from speech signals in conjunction with a simple expert system.

This paper is organized as follows. We present in section II the speech database SWITCHBOARD. Section III is concerned with the presentation of the automatic filled pause localization. The experimental results and conclusions are the subject of the last section.

II. DATABASE: SWITCHBOARD

A. *Introduction*

SWITCHBOARD is a corpus of spontaneous conversations which includes about 2430 conversations averaging 6 minutes in length and addresses the growing need for a large multispeaker database of telephone bandwidth speech [1]. It was collected at Texas Instruments with funding by DARPA and contains over 240 hours of recorded speech, and about 3 million words of text, spoken by over 500 speakers of both sexes from every major dialect of American English. Switchboard was collected directly from T1 lines without human intervention. The waveform files were recorded, with no degradation due to the collection system into two channels at an 8 kHz sample rate and with 8-bit mu-law quantization, exactly as read from the digital line. The speech was fully transcribed, and the transcription conventions documented. Court reporters produced most of the verbatim transcripts, following a manual prepared specifically for the project. Each transcript is accompanied by a time alignment file which was accomplished with supervised phone-based speech recognition [2] and which estimates the beginning time and duration of each word in the transcripts in centiseconds. For more details, see [1].

B. *Filled pauses in SWITCHBOARD*

In the approximately 2000 conversations examined, there were 65347 "uh" filled pauses and each conversation averaged about a dozen or more filled pauses. The estimation of duration of these filled pauses is directly derived from the timing files. They are word segmented and formatted as follows: Starting Time - Duration -Word [1]. The minimum duration of "uh" is found to be about 20 ms, whereas its maximum duration is 1200 ms. Average duration for "uh" filled pauses was a mean of about 165.45 ms and a median of about 180 ms. Figure 1 shows the histogram for the "uh" filled pause duration.

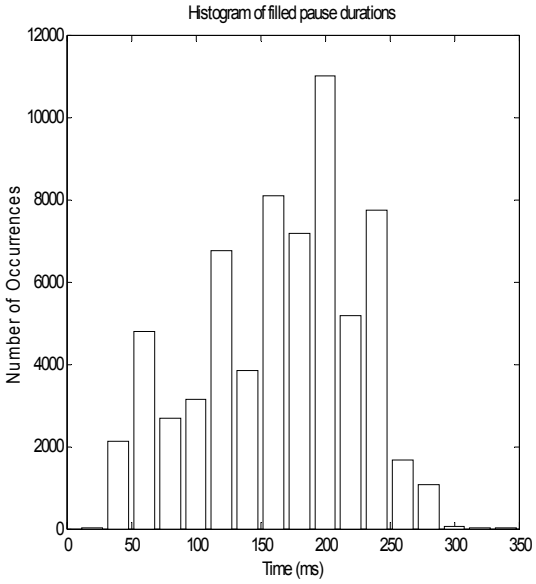

Fig. 1. Histogram of filled pauses durations

The "uh" filled pauses resemble short, simple words in continuous, spontaneous speech but could be distinguished by analyzing the silence periods (if any) adjacent to the filled pause. In the approximately 2000 conversations examined, 35156 "uh" filled pauses were preceded by a silence (mean of about 379.69 ms and median of about 240 ms), 54706 were followed by a silence (mean of about 522.12 ms and median of about 460 ms) and 54706 (83.72%) had an adjacent pause (for 72.88% this pause exceeded 100 ms). Figure 2 shows the histogram of silent pauses adjacent to a filled pause.

The presence of a filled pause is located as a long, steady vowel with low fundamental frequency F0 relative to the calculated average F0 for each speaker during a conversation. In comparing with the ATIS-database [2] the F0 was not very low. Studying the pitch pattern of such pauses shows that the filled pause tends to show a slightly falling pitch pattern when such a pause is preceded by a vowel but

SWITCHBOARD filled pauses include many examples with rising F0. Figure 3 shows an example at a turn-talking point.

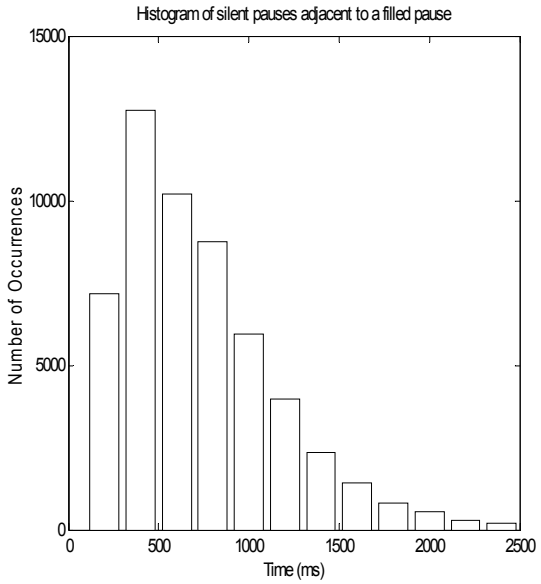

Fig. 2. Histogram of silent pauses adjacent to a filled pause

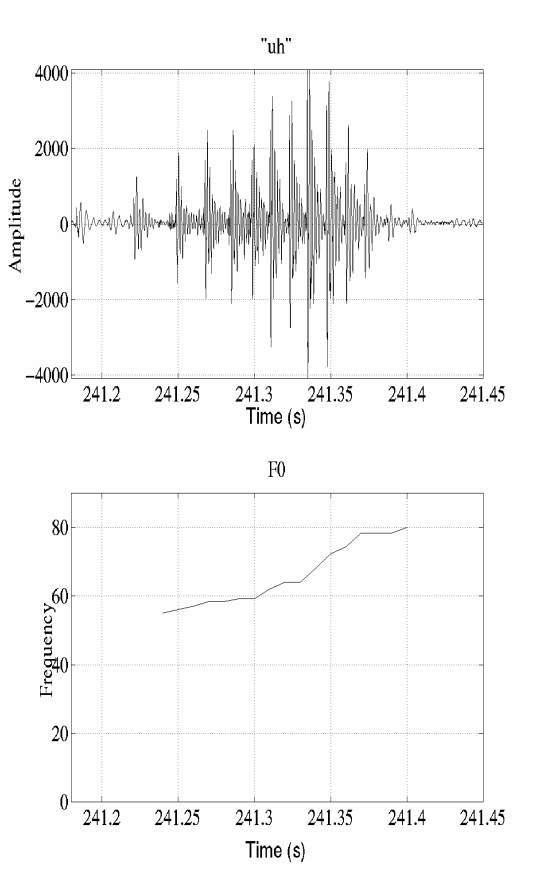

Fig. 3. The time-domain waveform of a filled pause "uh" and the corresponding pitch contour

The filled pauses have a spectrum of a steady central vowel with little spectral change. A typical example of a filled pause spectrum is illustrated in Figure 4. This figure shows "uh" after the vowel "I" uttered by a male speaker.

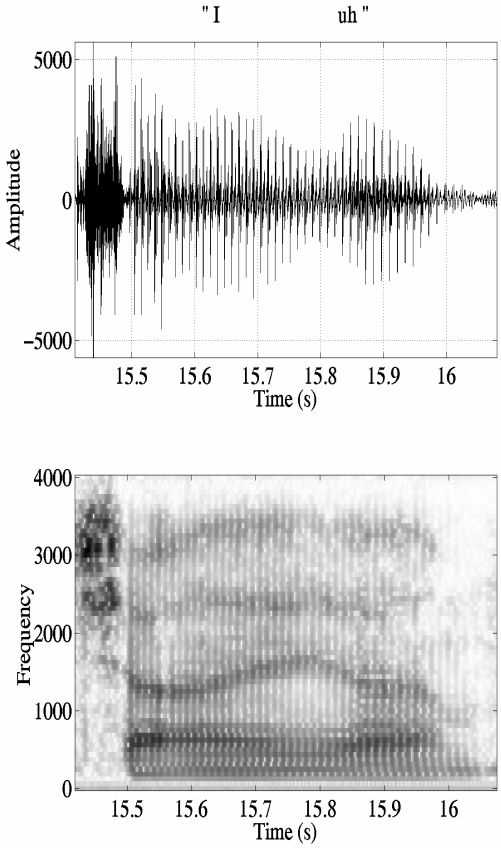

Fig. 4. The time-domain waveform of a filled pause "uh" uttered by a male speaker after the vowel "I" and the corresponding spectrogram

III. PROPOSED METHOD

The algorithm presented in this paper was developed by examination and analysis of many utterances from different speakers.

All phases of the task used automatic F0, energy, duration, and spectral estimation directly from speech signals in conjunction with a simple expert system.

The presence of filled pauses was estimated and located as long, steady vowels with low F0 relative to the calculated average F0 for the speaker during all the conversation and a spectrum of a steady central vowel with little spectral change (a cepstral distance was used) and often bordered by silence.

In the first step silent pauses and voiced zones were easily located using a simple algorithm with two energy and two zero crossing thresholds.

In the second step F0 estimates were provided using the autocorrelation method. The concepts of dynamic programming and pattern matching are used in postprocessing technique for pitch period contour tracking [15] [16]. Only the voiced zones with a low F0 are retained.

An automatic segmentation process [17] was used in a third step as a way of delimiting vowel-sized units in voiced zones with a low F0.
A cepstral distance was used in the fourth step to select the candidates with a spectrum of a steady central vowel.
In the last step are retained the candidates bordered by silence with a minimal duration.

IV. ACOUSTICAL ANALYSIS RESULTS

The algorithm proposed in section III has been tested over 2000 conversations from the SWITCHBOARD corpus. The energy, zero crossing and F0 thresholds used in the first three steps were estimated after a preliminary analysis of conversation waveforms. The spectrum of a steady central vowel with little spectral change used in the forth step was a mean filled pause spectrum in Case 1 or a spectrum corresponding to the first filled pause in the file in Case 2. Performance metrics included recall (RC): disfluencies detected / disfluencies, false alarms (FA): others called disfluencies / others and accuracy (AC): correct classifications / all data points. Results obtained after the last step are presented in Table 1 and after the second step in Table 2 for different recall rates. A future goal is to improve recognition performance by integrating the detection of filled pauses with a language model.

Table 1. Results obtained after the last step of the algorithm (%)

CASE 1			CASE 2		
RC	FA	AC	RC	FA	AC
71.43	1.44	98.07	71.43	1.70	97.81
85.71	1.83	97.81	78.57	3.01	96.65
92.86	2.36	97.43	85.71	3.80	96.01
92.86	3.67	96.27	92.86	4.85	95.11

Table 2. Intermediary results obtained after the second step of the algorithm (%)

CASE 1			CASE 2		
RC	FA	AC	RC	FA	AC
71.43	21.36	78.51	85.71	25.16	75.03
85.71	25.43	74.77	85.71	24.64	75.42
92.86	32.63	67.82	85.71	27.52	72.72
92.86	33.16	67.31	92.86	32.63	67.82

REFERENCES

[1] Godfrey J. J., Holliman E. C., and McDaniel J. "SWITCHBOARD Telephone Speech Corpus for Research and Development". IEEE International Conference on Acoustics. Speech, and Signal Processing. San Francisco, 1992, Vol. I, pages 517-520.

[2] Wheatley B., Doddington G., Hemphill, C., Godfrey, J. J., Holliman, E. C., McDaniel, J., and Fisher D. "Robust Automatic Time Alignment of Orthographic Transcriptions with Unconstrained Speech". IEEE International Conference on Acoustics, Speech, and Signal Processing. San Francisco, 1992, Vol. I, pages 533-536.

[3] O'Shaughnessy D. "Recognition of Hesitations in Spontaneous Speech". IEEE International Conference on Acoustics, Speech, and Signal Processing. San Francisco, 1992, Vol. I, pages 521-524.

[4] O'Shaughnessy D. "Locating disfluencies in Spontaneous Speech: An Acoustical Analysis". Proceedings of Eurospeech Conference. Berlin, 1993, pages 2187- 2190.

[5] O'Shaughnessy D. "Better Detection of Filled Pauses in Spontaneous Speech". 138th Meeting of Acoustical Society of America, Columbus, 1999.

[6] O'Shaughnessy D. "Better Detection of Hesitations in Spontaneous Speech". Workshop on Disfluency in Spontaneous Speech, Berkeley, 1999, pages 39-42.

[7] Shriberg E. "Disfluencies in SWITCHBOARD". International Conference on Spoken Language Processing. Philadelphia, 1996, Vol. Addendum, pages 11-14.

[8] Shriberg E., Bates, R., and Stolcke A. "A Prosody-only Decision-tree Model for Disfluency Detection". Proceedings of Eurospeech Conference, Rhodes, 1997, pages 2383-2386.

[9] Shriberg E. "Phonetic Consequences of Speech Disfluency". International Congress of Phonetic Sciences, San Francisco, 1999.

[10] Stolcke A. and Shriberg E. "Statistical language modeling for speech disfluencies". Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP-96, vol. 1, pages 405 – 408.

[11] Man-Hung Siu and Ostendorf M. "Modeling disfluencies in conversational speech". International Conference on Spoken Language Processing. Philadelphia, 1996, Vol. 1, pages 386-389.

[12] Nakatani C. H. "A Corpus-based Study of Repair cues in Spontaneous Speech". *Journal of the Acoustical Society of America*. 93(3):1603-1616, 1994.

[13] Lickley R. J. "Juncture Cues to Disfluency". International Conference on Spoken Language Processing. Philadelphia, 1996, pages. 2478-2481.

[14] Potamianos A. and Narayanan S. "A Review of the Acoustic and Linguistic Properties of Children's Speech". IEEE 9th Workshop on Multimedia Signal Processing, MMSP-2007, pages 22 – 25.

[15] Secrest B. G., and Doddington, G. R., "An Integrated Pitch Tracking Algorithm for Speech Systems". IEEE International Conference on Acoustics, Speech, and Signal Processing. pages 1352-1355, Boston, 1983.

[16] Ney H. "A Dynamic Programming Technique for Nonlinear Smoothing". IEEE International Conference on Acoustics, Speech, and Signal Processing. pages 62-65, 1981.

[17] Mermelstein P. "Automatic Segmentation of Speech into Syllabic Units". *Journal of the Acoustical Society of America*. 58:880-883, Oct. 1975.