A REAL-TIME FILLED PAUSE DETECTION SYSTEM FOR SPONTANEOUS SPEECH RECOGNITION

Masataka Goto *Katunobu Itou* *Satoru Hayamizu*
Machine Understanding Division, Electrotechnical Laboratory
1-1-4 Umezono, Tsukuba, Ibaraki 305-8568 JAPAN.
{goto, kito, hayamizu}@etl.go.jp http://www.etl.go.jp/~goto/

ABSTRACT

This paper describes a method for automatically detecting filled (vocalized) pauses, which are one of the hesitation phenomena that current speech recognizers typically cannot handle. The detection of these pauses is important in spontaneous speech dialogue systems because they play valuable roles, such as helping a speaker keep a conversational turn, in oral communication. Although a few speech recognition systems have processed filled pauses within subword-based connected word recognition or word-spotting frameworks, they did not detect the pauses individually and consequently could not consider their roles. In this paper we propose a method that detects filled pauses and word lengthening on the basis of small fundamental frequency transition and small spectral envelope deformation under the assumption that speakers do not change articulator parameters during filled pauses. Experimental results for a Japanese spoken dialogue corpus show that our real-time filled-pause-detection system yielded a recall rate of 84.9% and a precision rate of 91.5%.
Keywords: Filled pause, Hesitation, Spontaneous speech

1 INTRODUCTION

The goal of this research is to improve the computer's ability to understand speech to a degree that will make possible natural multimodal communication between humans and computers. This requires the computer to recognize the audio signals comprising the spontaneous speech uttered by a speaker thinking about speech contents on the fly. Hesitation phenomena, such as filled (vocalized) or unfilled (silent) pauses, word lengthening, restarts, and false starts, occur frequently in such speech. As an initial step toward dealing with those natural and inevitable phenomena, in this paper we concentrate on two frequent phenomena, filled pauses and word lengthening, because these phenomena play the same valuable roles in oral communication, such as helping a speaker hold a conversational turn and express mental and thinking states. In order to improve speech dialogue systems, we think that it is important to make good use of such roles without simply neglecting those phenomena.

Typical HMM-based speech recognizers accept only fluent read or planned speech without hesitation phenomena and have difficulty in dealing with spontaneous speech. The phone model, for example, does not work well when applied to speech with filled pauses and word lengthening because the duration of a phone tends to lengthen suddenly, and the language model is not effective enough to deal with filled pauses because the pauses can be inserted at almost arbitrary word positions. A few previous speech recognition systems [1][2][3] have partly handled filled pauses within subword-based connected word recognition or word-spotting frameworks. One HMM-based recognizer [2], for example, added several frequent filler words to the system vocabulary and another one [3] regarded filler words as out-of-vocabulary words and dealt with them by using a subword-unit based decoder for processing unknown words. Those systems, however, did not detect filled pauses individually and could not consider the roles of these pauses.

We therefore believe that it is necessary to detect filled pauses (fillers) and word lengthening in spontaneous speech by using bottom-up acoustical analysis. Previous investigations [4][5] of the prosodic features of filled pauses suggested the feasibility of detecting the pauses. The report of Quimbo *et al.* [5], in particular, supported the bottom-up approach of analyzing prosodic features by pointing out that human beings can, from prosodic cues, recognize filled pauses in speech that is in unfamiliar foreign language. In those investigations, however, a computational system of automatically detecting filled pauses was not built.

In this paper we propose a method for detecting filled pauses and word-lengthening phenomena in spontaneous speech. Since both these hesitation phenomena have similar acoustical features and can be considered to have the same functions in terms of oral communication, in the rest of this paper we use the term "filled pause" for both. In the following sections, we first discuss the roles of filled pauses and describe the algorithm of the method. We then show experimental results obtained using our real-time system based on the proposed method. Finally, we discuss applicability of the method in a speech recognition framework.

2 IMPORTANCE OF FILLED PAUSES

In this research, we hypothesize that the essential reason that filled pauses are inevitable in spontaneous utterances is that they are uttered when the thinking process cannot keep up with the speaking process. When the speed of speaking becomes faster than the speed of preparing its content, a speaker uses filled or unfilled pauses until the next speech content resulting from the thinking process arrives at the speaking process.

The primary utility of detecting filled pauses is that it makes it possible to improve the performance of speech recognizers by avoiding the application of typical HMM-based recognition to the filled-pause periods in spontaneous speech. Furthermore, this detection enables speech dialogue systems to make use of the following two important functions of filled pauses, which functions were also discussed in [6][7].

- Communicative functions

  In spoken dialogue, a speaker uses filled pauses to keep a conversational turn while taking enough thinking time to prepare a subsequent utterance. On the other hand, a listener hearing filled pauses usually waits for the speaker's subsequent utterance without interrupting the turn.

- Affective and cognitive functions

  For achieving smooth dialogue by sharing mental states among interlocutors, a speaker unconsciously uses filled pauses to express mental states such as diffidence, anxiety, hesitation, and humility and also to express different thinking states, such as retrieving information from memory and seeking an expression appropriate for a listener. On the other hand, a listener interprets filled pauses as indicators for inferring speaker's mental and thinking states. In addition, filled pauses sometimes enable a listener to predict the speaker's subsequent utterance to some extent.

3 DETECTION METHOD

The basic idea of our method is to find acoustical features of filled pauses in speech signals by using frequency analysis. If filled pauses are, as described in the previous section, uttered while the speaking process is waiting for the next speech content from the thinking process, a speaker cannot change articulator parameters during the filled pauses because subsequent utterances have not yet been prepared. Our method hence assumes that a filled pause contains a continuous voiced sound of an unvaried phoneme, because such a sound is uttered when the vocal cords are vibrated with almost constant articulator parameters (i.e., with a constant vocal-tract shape). Typical Japanese fillers such as /ee-/, /maa-/, and /ano-/ as well as most word-lengthening sounds satisfy this assumption.

Our method accordingly detects filled pauses on the basis of the following two features:

1. Small F0 (fundamental frequency) transition

   When the tension of the vocal cords is unvaried under constant articulator parameters, the F0 of the voice remains almost constant.

2. Small spectral envelope deformation

   When the vocal tract shape is unvaried under constant articulator parameters, the spectral envelope forming the formants remains almost constant. When the deformation of the envelope is evaluated, it is necessary to eliminate the air flow's amplitude modulation, since the air flow from the lungs may vary.

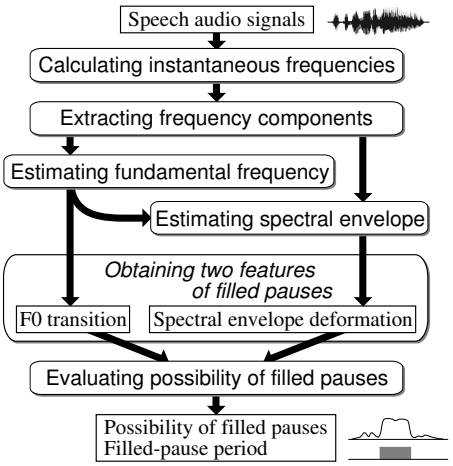

Figure 1: Overview of our filled-pause-detection method.

In the following, we describe the main procedure of our method (Figure 1).

3.1 Calculating Instantaneous Frequencies

The first step is to calculate the *instantaneous frequency* [8], the rate of change of the phase of a signal, of filter-bank outputs by using the short-time Fourier transform (STFT) whose output can be interpreted as a collection of uniform-filter outputs. When the STFT of a signal $x(t)$ with a window function $h(t)$ is defined as

$$X(\omega, t) = \int_{-\infty}^{\infty} x(\tau) h(\tau - t) e^{-j\omega\tau} d\tau = a + jb, \qquad (1)$$

the instantaneous frequency $\lambda(\omega, t)$ is given by this equation [8]:

$$\lambda(\omega, t) = \omega + \frac{a\frac{\partial b}{\partial t} - b\frac{\partial a}{\partial t}}{a^2 + b^2}. \qquad (2)$$

In our current implementation the input signal is digitized at 16 bit / 16 kHz, and then the STFT with a 1024-sample Hanning window is calculated by using the Fast Fourier Transform (FFT). Since the FFT frame is shifted by 160 samples, the discrete time step (1 *frame shift*[1]) is 10 ms.

3.2 Extracting Frequency Components

The extraction of frequency components is based on the the mapping from the center frequency ω of an STFT filter to the instantaneous frequency $\lambda(\omega, t)$ of its output [9][10][11]. By finding fixed stable points of the mapping, we can extract a set $\Psi_f(t)$ of instantaneous frequencies of the frequency components by using the following equation [10]:

$$\Psi_f(t) = \{\ \psi\ \mid\ \lambda(\psi, t) - \psi = 0,\ \frac{\partial}{\partial\psi}(\lambda(\psi, t) - \psi) < 0\}. \qquad (3)$$

By calculating the power of those frequencies which is given by the STFT power spectrum at $\Psi_f(t)$, we can define the power distribution function $\Psi_p(\omega, t)$ as

$$\Psi_p(\omega, t) = \begin{cases} |X(\omega, t)| & \text{if } \omega \in \Psi_f(t) \\ 0 & \text{otherwise.} \end{cases} \qquad (4)$$

[1]The term *time* in this paper is the time measured in units of frame shift.

3.3 Estimating Fundamental Frequency

To estimate the F0 of a speaker's voice in real-world audio signals containing background noises or music, we find the most predominant harmonic structure in the extracted frequency components by using a comb-filter-like analysis. The basic idea is to evaluate the possibility $P_{F0}(F,t)$ of the F0 at frequency F at time t. Hereafter, we use the frequency unit *cent* to denote the log-scale frequency. In this paper the frequency f_{Hz} in Hz is converted to the frequency f_{cent} in cent as follows:

$$f_{cent} = 1200 \log_2 \frac{f_{Hz}}{\mathrm{REF}_{Hz}} \quad (5)$$

$$\mathrm{REF}_{Hz} = 440 \times 2^{\frac{3}{12}-5}. \quad (6)$$

The F0 possibility $P_{F0}(F,t)$ is defined as

$$P_{F0}(F,t) = \int_{-\infty}^{\infty} p(x;F)\ \Psi'_p(x,t)\ dx, \quad (7)$$

where the unit of frequencies x and F is cent, $p(x;F)$ denotes a filter function, and $\Psi'_p(x,t)$ is the same as the power distribution $\Psi_p(\omega,t)$ (Equation (4)) except that the frequency unit is cent. The $p(x;F)$ passes harmonic components of the F0 F and is given by

$$p(x;F) = \sum_{h=1}^{\mathrm{N}} \mathrm{H}^{h-1}\ G(x;F + 1200 \log_2 h, \mathrm{W}_1) \quad (8)$$

$$G(x;m,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-m)^2}{2\sigma^2}\right), \quad (9)$$

where N (8, in our current implementation) is the number of harmonics considered, H (0.97) is an amplitude attenuation factor, and W_1 (20 cent) is the standard deviation of the Gaussian distribution $G(x;m,\sigma)$.

Finally, the frequency $F_{F0}(t)$ of the F0 is determined by finding the frequency that maximizes $P_{F0}(F,t)$:

$$F_{F0}(t) = \underset{F}{\mathrm{argmax}}\ P_{F0}(F,t). \quad (10)$$

3.4 Estimating Spectral Envelope

For robustness in real-world environments, the spectral envelope is estimated by using only local information on the harmonic structure of the obtained F0 $F_{F0}(t)$. The power $Pow(k,t;F_{F0}(t))$ of k-th harmonic component of $F_{F0}(t)$ is extracted by finding the local-maximum power calculated with a Gaussian kernel around each F0's multiple:

$$Pow(k,t;F_{F0}(t)) = \max_x\ G(x;F_{F0}(t) + 1200 \log_2 k, \mathrm{W}_2)\ \Psi'_p(x,t) \quad (11)$$

where W_2 (35 cent) is the standard deviation of the Gaussian distribution.

We then estimate the spectrum envelope in the linear-scale frequency by linear interpolation of the adjacent $Pow(k,t;F_{F0}(t))$. This envelope is calculated under the upper-limit frequency (3200 Hz) that covers the first and second formant frequencies of Japanese vowels. To make use of the global envelope deformation as the filled-pause feature, the method resamples the interpolated envelope at low frequency resolution ξ (200 Hz) and obtains the spectral envelope $Env(n,t)$ at frequency $n\xi$ where $1 \leq n \leq \mathrm{N}_{\max}$ (15). Then the $Env(n,t)$ is normalized so that it satisfies $\Sigma_{n=1}^{\mathrm{N}_{\max}} Env(n,t) = 1$ in order to compensate for the amplitude modulation of lungs' air flow.

3.5 Obtaining Two Features of Filled Pauses

The two filled-pause features the method obtains are the amount $A_f(t)$ of the F0 transition that indicates how much the F0 changes and the amount $A_s(t)$ of the spectral envelope deformation that indicates how much and how unevenly the spectral envelope changes.

The amount $A_f(t)$ is defined as the absolute value of the slope b_{F0} of a straight line obtained by least-squares fitting of short-term transition of the log-scale F0 $F_{F0}(t)$. The b_{F0} is obtained by minimizing

$$err_{F0}{}^2 = \Sigma_{\tau=0}^{\mathrm{Period}_{\mathrm{F0}}-1}(F_{F0}(t-\tau) - (a_{F0} + b_{F0}\tau))^2 \quad (12)$$

over a_{F0} and b_{F0}, where $\mathrm{Period}_{\mathrm{F0}}$ (5 frame shifts) is the fitting period.

The amount $A_s(t)$ is defined as

$$A_s(t) = \left(\frac{1}{\mathrm{N}_{\max}}\Sigma_{n=1}^{\mathrm{N}_{\max}} b_s(n)^2\right)\left(\frac{1}{\mathrm{N}_{\max}}\Sigma_{n=1}^{\mathrm{N}_{\max}} err_s(n)^2\right), \quad (13)$$

where $b_s(n)$ is the slope of a straight line that is similarly obtained by least-squares fitting of n-th-harmonic's short-term transition of the log-scale power of the envelope Env(n,t). The method minimizes the fitting error $err_s(n)$

$$err_s(n)^2 = \Sigma_{\tau=0}^{\mathrm{Period}_{\mathrm{s}}-1}(10 \log_{10} Env(n,t-\tau) - (a_s(n) + b_s(n)\tau))^2 \quad (14)$$

over $a_s(n)$ and $b_s(n)$, where $\mathrm{Period}_{\mathrm{s}}$ (10 frame shifts) is the fitting period.

3.6 Evaluating Possibility of Filled Pauses

On the basis of short-term averages $S_i(t)$ $(i = f,s)$ of the two obtained features, the possibility $P_{fp}(t)$ of filled pauses is defined as

$$P_{fp}(t) = \exp\left(-\frac{(\mathrm{R}\ S_f(t) + (1-\mathrm{R})\ S_s(t))^2}{\mathrm{W}^2}\right) \quad (15)$$

$$S_i(t) = \frac{1}{\mathrm{Period}_{\mathrm{fp}}}\Sigma_{\tau=0}^{\mathrm{Period}_{\mathrm{fp}}-1} A_i(t-\tau) \quad (i = f,s) \quad (16)$$

where R (0.034) and W (0.575) are empirical constant values and $\mathrm{Period}_{\mathrm{fp}}$ (10 frame shifts) is the averaging period.

When this possibility is high enough for a certain period of time, the method judges that a filled pause is uttered. We calculate the accumulated sum $Sum_{fp}(t)$ of $P_{fp}(t)$ as long as $P_{fp}(t) > e^{-1}$. When $Sum_{fp}(t) > \mathrm{Th}_{\mathrm{fp}}$, where $\mathrm{Th}_{\mathrm{fp}}(7e^{-1})$ is a constant threshold, the current time t is judged to be within the filled-pause period.

4 EXPERIMENTAL RESULTS

A real-time filled-pause-detection system based on the above method has been implemented and tested on a Japanese spontaneous speech corpus consisting of 100 utterances by five men and five women (10 utterances per subject). Each utterance contained at least one filled pause. Those utterances were excerpted from a spontaneous speech dialogue corpus [12] collected using a Wizard of OZ system and were automatically segmented by detecting each silence interval longer than 300 ms.

In our experiment the recall rate (the number of filled pauses detected correctly / the total number of filled pauses) was 84.9 percent (107 / 126) and the precision rate

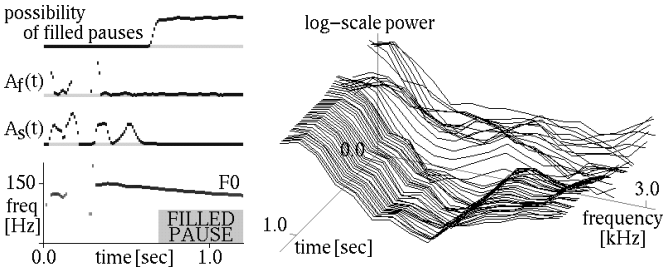

Figure 2: An example of the F0 and intermediate results (left) and the corresponding spectral envelope (right) for part of a male spontaneous utterance /iqkaini-/.

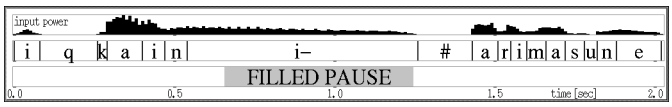

Figure 3: An example of the hand-labeled phone sequence and the detected filled-pause period for a male spontaneous utterance /iqkaini-#arimasune/.

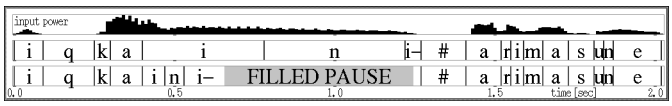

Figure 4: An example of the original alignment result and the alignment result improved by the detected filled-pause period.

(the number of filled pauses detected correctly / the total number of filled pauses detected) was 91.5 percent (107 / 117). Figure 2 shows an example of intermediate results of evaluating the possibility of filled pauses and Figure 3 shows an example of the result of correctly detecting a filled pause.

The main reasons for the recall-rate errors (misses) were too short duration of filled pauses like a short /e-/, too large F0 changes, and disorder of harmonic components of hoarse voices. The recall rate was exceptionally low (53.8 percent) for a particular male subject who tended to speak with a low-frequency hoarse voice. On the other hand, the precision-rate errors (false alarms) mainly occurred at continuous unvaried voiced sounds within words uttered with a flat F0. Such sounds tended to occur at successive similar vowel sounds caused by the target undershoot of phone transition, while our method rejected typical successive similar vowel sounds because their F0 changed enough in usual.

5 APPLICABILITY IN SPEECH RECOGNITION FRAMEWORK

As a preliminary step toward making use of the filled-pause detection method in a speech recognition framework, we tested how much phone alignment can be improved by using the detected filled-pause periods. In this preliminary test, the input was matched with a phone HMM connected according to the given correct phone sequence corresponding to the input. In general, a filled pause has a bad influence on the phone alignment determined by the Viterbi algorithm. To improve this alignment, we used the detected filled-pause period for dynamic phone-duration control: during the detected period we inhibited the transition from a vowel phone to the next phone.

Figure 4 shows an example of an original bad alignment result and the alignment result improved by using the filled-pause detection. Results like this suggest that when utterances contain filled pauses the performance of a typical HMM-based speech recognizer can be improved by using the filled-pause detection method.

6 CONCLUSION

We have described a method for detecting filled pauses and word-lengthening phenomena by finding a continuous voiced sound of an unvaried phoneme. The method is based on two acoustical features, small F0 transition and small spectral envelope deformation, which are estimated by identifying the most predominant harmonic structure in the input. Experimental results for a Japanese spontaneous speech corpus show that our system can detect, in real time, filled pauses with a recall rate of 84.9 percent and a precision rate of 91.5 percent.

We plan to apply our method to a speech recognizer by using not only the filled-pause period (discrete judgement) but also the filled-pause possibility value (continuous judgement). Future work will also include application of our method to English filled pauses and integration of the method with a speech dialogue system to make full use of the valuable functions of filled pauses.

REFERENCES

[1] W. Ward. Understanding spontaneous speech: The Phoenix system. In *Proc. of ICASSP 91*, pp. 365–367, 1991.

[2] S. Nakagawa and S. Kobayashi. Phenomena and acoustic variation on interjections, pauses and repairs in spontaneous speech (*in Japanese*). *J. Acoust. Soc. Jpn. (J)*, 51(3):202–210, 1995.

[3] A. Kai and S. Nakagawa. Investigation on unknown word processing and strategies for spontaneous speech understanding. In *Proc. of Eurospeech '95*, pp. 2095–2098, 1995.

[4] D. O'Shaughnessy. Recognition of hesitations in spontaneous speech. In *Proc. of ICASSP 92*, pp. I–521–524, 1992.

[5] F. C. M. Quimbo, T. Kawahara, and S. Doshita. Prosodic analysis of fillers and self-repair in Japanese speech. In *Proc. of ICSLP 98*, 1998.

[6] Y. Takubo. Towards a linguistic model of speech performance (*in Japanese*). *Journal of Information Processing Society of Japan*, 36(11):1020–1026, 1995.

[7] R. L. Rose. *The communicative value of filled pauses in spontaneous speech*. PhD thesis, University of Birmingham, 1998.

[8] J. L. Flanagan and R. M. Golden. Phase vocoder. *The Bell System Technical Journal*, 45:1493–1509, 1966.

[9] F. J. Charpentier. Pitch detection using the short-term phase spectrum. In *Proc. of ICASSP 86*, pp. 113–116, 1986.

[10] T. Abe, T. Kobayashi, and S. Imai. The IF spectrogram: a new spectral representation. In *Proc. of ASVA 97*, pp. 423–430, 1997.

[11] H. Kawahara, H. Katayose, R. D. Patterson, and A. de Cheveigné. Highly accurate F0 extraction using instantaneous frequencies *(in Japanese). Tech. Com. Psycho. Physio., Acoust. Soc. of Japan, H-98-116*, pp. 31–38, 1998.

[12] K. Itou, T. Akiba, O. Hasegawa, S. Hayamizu, and K. Tanaka. A Japanese spontaneous speech corpus collected using automatically inferencing Wizard of OZ system. *J. Acoust. Soc. Jpn. (E)*, 20(3), 1999.