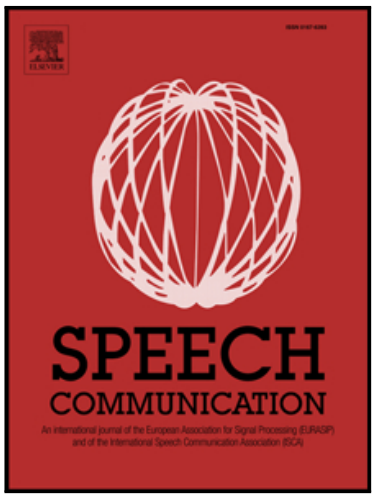

Detecting Breathing Sounds in Realistic Japanese Telephone Conversations and Its Application to Automatic Speech Recognition

Takashi Fukuda, Osamu Ichikawa, Masafumi Nishimura

PII: S0167-6393(17)30237-6
DOI: 10.1016/j.specom.2018.01.008
Reference: SPECOM 2530

To appear in: *Speech Communication*

Received date: 30 June 2017
Revised date: 29 January 2018
Accepted date: 31 January 2018

Please cite this article as: Takashi Fukuda, Osamu Ichikawa, Masafumi Nishimura, Detecting Breathing Sounds in Realistic Japanese Telephone Conversations and Its Application to Automatic Speech Recognition, *Speech Communication* (2018), doi: 10.1016/j.specom.2018.01.008

Detecting Breathing Sounds in Realistic Japanese Telephone Conversations and Its Application to Automatic Speech Recognition

Takashi Fukuda[a], Osamu Ichikawa[a], Masafumi Nishimura[b]

[a]*IBM Research AI, Nihonbashi Hakozaki-cho, Chuo-ku, Tokyo, 103-8510, Japan*
[b]*Shizuoka University, Suruga-ku, Shizuoka, 422-8017, Japan*

Abstract

Non-verbal sound detection has long attracted attention in the speech analytics field. Although detecting laughter, coughs, and lip smacking has been well studied in the literature, breath-event detection has not been investigated much despite the need for doing so. Breath events are highly correlated with major prosodic breaks, meaning that the positions of breath events can be used as a delimiter of utterances in combination with a voice activity detection (VAD) technique. Silence intervals approximately 20 ms long right before and after breathing sounds, called "edges", are clearly observed in speech signals. In the literature, capturing the edges is shown to be very effective in reducing false alarms in the detection of breath events. However, the edges often disappear when breaths are taken in spontaneous speech. In this work, we focus on the robustness of breath-event detection in spontaneous speech. The breath detection method we have developed leverages acoustic information that is specialized for breathing sounds, leading to a two-step approach that can detect breath events with an accuracy of 97.4%. We also propose splitting unsegmented speech signals into semantically grouped utterances by leveraging the breath events. The speech segmentation based on accurate breath-event detection provided a 3.8% relative error reduction in automatic speech recognition (ASR).

Keywords: Breath-event detection, spontaneous speech, speech phrasing, voice activity detection, automatic speech recognition.

1. Introduction

Many applications of speech technology including automatic speech recognition (ASR), speech analytics, and digital speech communication equipment have recently been commercialized by speech solution vendors to take advantage of the progress being made with speech technology. Typical examples of the applications include a voice search used in mobile phones (Sainath et al., 2017), a voice control for car navigation systems (Wang et al., 2008), and a spoken dialog system for use with robots (Williams and Young, 2007). Applying speech technology to the clinical and healthcare fields has also attracted attention (Saon and Chien, 2012).

In addition to these applications, techniques for telephone conversations have recently been attracting attention and the related technologies have been used in many business situations, such as for decoding operators' utterances in call centers and automatically displaying relevant information searched from internal/external Web sites with decoded results. Although the recent focus in using speech technology for telephone conversations is ASR, also attracting attention is the exploitation of non-verbal sounds that are predefined events in speech signals such as laughter, coughs, and lip smacking (Laskowski, 2009; Knox and Mirghafori, 2007; Matos et al., 2006; Drugman et al., 2011; Mesaros et al., 2010). Although there have been various studies on the automatic detection of predefined events, little investigation has been done on the detection of breath events. In particular, there has been almost no research on either the detection of breath events observed in spontaneous speech or that explicitly leveraging breath-event information for ASR, in spite of its great potential for the analysis of paralinguistic information.

This paper focuses on the automatic detection of inhaled breath events that are clearly observed in realistic telephone conversations (Fukuda et al., 2011). The definition of the breath events in this paper will be shown in Section 2.1. The target language in this paper is Japanese[1]. Though there are also some studies that focus on the analytics of "exhaled" breath events, which seems effective mainly in the clinical field (Amann et al., 2014; Konvalina and Haick, 2014), detecting the exhaled breath events is beyond the scope of this paper. Prior works on breath-event detection include the automatic segmentation of a continuous speech signal, where breath events serve as natural delimiters in utterances (Price et al., 1989; Wightman and Ostendorf, 1994). Price et al. (1989) proposed using a Gaussian mixture model (GMM) to discriminate breath and non-breath events. Wightman and Ostendorf (1991) improved breath-event detection performance by using cepstral coefficients and an algorithm based on Bayesian discrimination. Ruinskiy and Lavner (2007) proposed an effective breath-event detection algorithm based on template matching and the accurate detection of very

*Corresponding author.
Email address: fukuda1@jp.ibm.com (Takashi Fukuda)
URL: http://researcher.ibm.com/researcher/ (Takashi Fukuda)

[1]Timings of taking breaths during conversations are speaker dependent.