Speech Communication 83 (2016) 10–20

Contents lists available at ScienceDirect

Speech Communication

journal homepage: www.elsevier.com/locate/specom

Whisper to normal speech conversion using pitch estimated from spectrum

Hideaki Konno[a,*], Mineichi Kudo[b], Hideyuki Imai[b], Masanori Sugimoto[b]

[a] Hakodate Campus, Hokkaido University of Education, 1–2 Hachiman-cho, Hakodate, Hokkaido 040–8567, Japan
[b] Graduate School of Information Science and Technology, Hokkaido University, Sapporo, Hokkaido 060–0814, Japan

ARTICLE INFO

Article history:
Received 15 December 2015
Accepted 2 July 2016
Available online 5 July 2016

Keywords:
Whispered speech
Pitch
Mel-scaled filter bank
Principal component analysis
Multiple regression analysis

ABSTRACT

We can perceive pitch in whispered speech, although fundamental frequency (F_0) does not exist physically or phonetically due to the lack of vocal-fold vibration. This study was carried out to determine how people generate such an unvoiced pitch. We conducted experiments in which speakers uttered five whispered Japanese vowels in accordance with the pitch of a guide pure tone. From the results, we derived a multiple regression function to convert the outputs of a mel-scaled filter bank of whispered speech into the perceived pitch value. Next, using this estimated pitch value as F_0, we constructed a system for conversion of whispered speech to normal speech. Since the pitch varies with time according to the spectral shape, it was expected that the pitch accent would be kept by this conversion. Indeed, auditory experiments demonstrated that the correctly perceived rate of Japanese word accent was increased from 55.5% to 72.0% compared with that when a constant F_0 was used.

1. Introduction

People sometimes whisper to keep a conversation amongst a few people so as not to disturb others in public or at meetings. Whispering has become more popular recently due to the widespread use of smart phones. Some people, whose larynx has been affected by an accident or disease, can only converse by whispering. There is a need for an effective method to convert whispered speech to normal speech to make the whispered speech more intelligible over the phone or to improve communication for impaired persons. However, conversion of whispered speech to normal speech is difficult because some information is obscured or missing in whispered speech. A major problem is that distinct pitch, i.e., tone height, is not maintained in whispered speech.

Pitch[1], as one of the attributes of sensation, conveys speech prosody such as intonation and accent of pitch-accent languages such as Japanese. The pitch of normally phonated speech corresponds to its fundamental frequency (F_0), which is the number of vocal-fold vibrations per second. However, whispered speech has no F_0 because the sound source of whispering is turbulent airflow (Laver, 1994) and the vocal folds do not vibrate. Nevertheless, whispered speech can communicate the prosody (Heeren and van Heuven, 2014; Heeren and Lorenzi, 2014; Kong and Zeng, 2006; Tartter and Braun, 1994) and convey the sensation of pitch.

Pitch-accent languages (e.g., Japanese) and tone languages (e.g., Chinese) have lexically distinct words that differ only in pitch (McCawley, 1968). Using such words in four pitch-accent or tone languages, (Jensen, 1958) reported the correct recognition rates of the whispered word meaning. The rates were 53% to 73% in Norwegian, 100% in Swedish, 71% to 85% in Slovenian, and 73% to 88% in Chinese (Mandarin). As for Japanese whispered words, (Sugito et al., 1991) reported an approximately 90% accuracy in perceptually recognizing correct accents. In Mandarin, 72% of whispered words were judged with a correct tone (Kong and Zeng, 2006). Thus, whispered speech is considered to carry pitch-accent or tonal cues.

Recently, various techniques for converting whispered speech to normal speech have been developed with the aim of improving intelligibility and naturalness (Huang et al., 2012; McLoughlin et al., 2015; Morris and Clements, 2002; Sharifzadeh et al., 2010; Tran et al., 2010), and conversion from other kinds of speech, e.g., body-conducted speech, to whispered speech has also been studied (Hirahara et al., 2010; Toda et al., 2012).

Two things are necessary for conversion of whispered speech to normal speech: modification of the vocal tract characteristics and

* Corresponding author.
E-mail address: konno.hideaki@h.hokkyodai.ac.jp (H. Konno).

[1] ANSI (1994) defines pitch as "that attribute of auditory sensation in terms of which sounds may be ordered on a scale extending from low to high."

http://dx.doi.org/10.1016/j.specom.2016.07.001
0167-6393/

generation of a F_0 contour. The vocal tract characteristics have to be modified because they are different in whispered speech and normally phonated speech even if the same phonemes are being uttered. In whispered speech, perceived pitch may be generated by changing the vocal tract characteristics. Therefore, we need to recover the vocal tract characteristics in a synthetic system. Another ingredient for normal speech synthesis is the F_0 contour, which conveys intonation and pitch accent in normally phonated speech. The methodology of generating the F_0 contour is classified into five groups: 1) the use of a fixed F_0 for any word (Sharifzadeh et al., 2010), 2) use of a contour model of F_0 for individual words (Huang et al., 2012), 3) estimation of the F_0 contour on the basis of the power of the speech (Morris and Clements, 2002), 4) application of the F_0 contour estimated from the normal speech which has the same phrase as the whispered speech (Morris, 2003; Toda et al., 2012; Tran et al., 2010), and 5) estimation of the F_0 contour from the spectral shape including formants (McLoughlin et al., 2013; 2015). Note that the first two groups give the same fixed F_0 contour regardless of the uttered words or only a slight variation of the F_0 contour. The third group does not correct pitch contour unless appropriate feedback control of speech gain is made. The fourth group needs pairs of normal and whispered speech of the same sentence for designing the system. This is sometimes too costly and impractical, e.g., in cases in which the normal voice has been lost due to disease. Therefore, in this study, we used the fifth method in which a dynamic pitch contour is generated depending on the uttered whispered speech.

In this study, we first analyzed the relationship between the perceived pitch and short-term amplitude spectra in five whispered Japanese vowels. To achieve this goal, taking the same point of view as (Moore, 2013) who stated that "assigning a pitch value to a sound is generally understood to mean specifying the frequency of a pure tone that has the same subjective pitch as the sound," we conducted an experiment in which speakers uttered a whispered vowel in such a way that the pitch of the vowel is intended to be the same as that of a guide pure tone with a fixed frequency. We connected the observed spectral information to the guide pitch caused by the pure tone's frequency and quantitatively evaluated the relationship. Based on the results of this analysis, we derived a regression formula to estimate the pitch value from the spectrum of whispered vowels. Next, we applied this estimation technique to the conversion of whispered speech to normal speech. To the best of the authors' knowledge, this is the first trial to use an estimated pitch value in this type of conversion. It is expected that the pitch-accent clarity is improved by this method, because the estimated pitch value recovers the dynamics of pitch of the same word in normal speech. We confirmed the effectiveness of this method in auditory experiments.

The contribution of this paper is fourfold: (1) a systematic method for collecting whispered vowels having a certain pitch has been established, (2) a regression function for estimating the pitch value from the short-term spectrum of whispered speech has been derived, (3) the estimated pitch value is utilized for conversion of whispered speech to normal speech, especially aiming at recovery of the accent of words, and (4) the developed conversion system does not require any extra information other than the input whispered speech, unlike the conversion systems developed in many previous studies.

2. Related studies

First, we will review studies on the pitch of whispered speech and then studies on conversion of whispered speech to normal speech.

There have been many studies on the differences between normal speech and whispering. Many of those studies were experimental observations. Whispered vowels with different levels of perceived pitch have been analyzed. According to Meyer-Eppler (1957), in five whispered German vowels, as the pitch level increases, a specific range of the spectrum rises upward in one group of vowels, and only the power in a high-frequency region increases in another group. In five whispered Japanese vowels, (Hirahara, 1991) showed that the degree of increase in formant frequencies varies over vowels when the pitch increases. A significant change of F_1 (the first formant frequency) and F_2 (the second formant frequency) on Japanese vowel /a/ was reported by Higashikawa et al. (1996), and, similarly, a large change of F_2 was reported for Mandarin (Chen and Zhao, 2008).

A speech perceptional approach has also been taken to reveal the relationship between the pitch of whispered vowels and formant frequencies. An auditory experiment showed that the pitch of the whispered vowel /a/ can be increased by increasing F_1 and F_2 (Higashikawa and Minifie, 1999).

In their previous studies, the authors analyzed the mechanism underlying the generation of perceived pitch in whispered speech and in particular, its relationship with the spectrum. In the first report (Konno et al., 1994), in explaining the phenomenon in which sounds with different formant frequencies can be perceived as the same vowel, it was pointed out that the perceived pitch might affect the phonemic quality of those vowels. The authors later showed that both formant frequencies in a low-frequency region and spectral tilt affect the perception of pitch in whispered vowels (Konno et al., 1996). Those studies showed that not only formants, but also spectral shape is influential in the perception of pitch in whispered speech; however, the degree of influence was not determined. Recently, the authors provided preliminary results regarding the degree of influence obtained from statistical analysis of whispered Japanese vowels uttered according to a guide tone with a specific pitch (Konno et al., 2013). In this study, we investigated the quantitative relationship between spectral information and pitch in more detail.

The following is a review of methods for converting whispered speech to corresponding normal speech. The typical conversion systems are shown in Fig. 1 and Table 1.

Conversion of whispered speech to normal speech requires 1) modification of vocal tract information and 2) generation of the fundamental frequency F_0 (Fig. 1). Modification of vocal tract information is typically carried out by shifting formant frequencies and altering formant bandwidths or by spectrum estimation using a Gaussian mixture model.

There are mainly two types of methods for F_0 generation. In one approach (Type A in Fig. 1), the formants or the gain obtained from the whispered speech are used for generating F_0 (e.g., McLoughlin et al., 2015; Morris and Clements, 2002; Sharifzadeh et al., 2010). This approach typically needs additional information such as the rule for F_0 generation as a function of formants or gain. In general, it is difficult to construct such a rule appropriately. In the other approach (Type B in Fig. 1), a model of F_0 is constructed in advance in a feature space, such as a Gaussian mixture model in the cepstral coefficient space (Toda et al., 2012; Tran et al., 2010) or a jump Markov linear system with speech gain and linear prediction cepstral coefficients (LPCC) (Morris, 2003) , and F_0 is generated by referring to the extracted spectral information of the whispered speech. In this case, we need the pairs of normal speech and whispered speech of the same sentences with time mapping of phonemes to train the model. We summarize the methodology in Table 1.

In the approach proposed in this study, we do not need any *a priori* rules as in the Type A approach or any extra data other than whispered speech (unlike in Type B approaches). By our approach, whispered speech's original intonation, accent and tone are expected to remain preserved in the converted normal speech.