Proposal of Acoustic Measures for Automatic Detection of Vocal Fry

Carlos Toshinori Ishi, Hiroshi Ishiguro and Norihiro Hagita

Intelligent Robotics and Communication Laboratories

ATR, Kyoto, Japan
carlos@atr.jp, ishiguro@ams.eng.osaka-u.ac.jp, hagita@atr.jp

Abstract

Vocal fry is a voice quality that often appears in relaxed voices indicating low tension, or in pressed voices expressing attitudes/feelings of surprise, admiration and suffering. We propose a set of acoustic measures for automatically detecting vocal fry segments in speech utterances. In order to deal with vocal fry utterances with very low fundamental frequencies, where classic short-term analysis methods become problematic, a glottal pulse synchronized method is proposed. The acoustic measures are based on power, periodicity and similarity properties of vocal fry signals. The basic idea is to scan for power peaks in a “very short-term” power contour, check for periodicity properties and evaluate a similarity measure between power peaks for deciding the possibility of vocal fry pulses. Sub-harmonic properties are also taken into account in the periodicity analysis. Evaluation of the proposed measures in automatic detection resulted in 73.3 % correct detection, with an insertion error rate of 3.9 %.

1. Introduction

We are analyzing acoustic features which are related with paralinguistic information in speech, for purposes of automatic extraction in a human-machine communication scenario. Besides the largely used prosodic features (such as pitch, power and duration), recent studies have reported that voice quality information (due to modality in the laryngeal voice source, e.g., breathiness, creakiness, harshness, etc.) also takes an important role in the perception of paralinguistic information.

Vocal fry (henceforth VF), creak, creaky voice, glottal fry, pulse register, and laryngealization are terminologies found in the literature for a voice quality characterized by a train of relatively discrete laryngeal (or glottal) excitations (or pulses of the literature for a voice quality characterized by a train of expressive emphasizing utterances as a pressed voice (called “rikimi” in Japanese) that carries paralinguistic information primarily associated with feelings/attitudes of surprise, admiration and suffering [9]. VF segments in such pressed voices are often observed to have very low fundamental frequencies.

Further, the irregular characteristics of VF segments can cause severe errors in pitch determination algorithms [10] which are important for prosodic information extraction. Thus, knowledge about the location of VF could be used to indicate paralinguistic information as well as to improve pitch determination.

There are many studies reporting physiological, perceptual and acoustic properties of VF in several research areas [1-6,11-16]. Many of them report qualitative or descriptive analyses of acoustic features that are related with different voice qualities. However, only a few evaluate their performance for automatic detection purposes. In this paper, we discuss the problems related with automatic detection of VF segments, and propose a set of acoustic measures, focusing on the problems caused by very low fundamental frequencies.

2. The problems related with automatic detection and the proposed method

The fundamental frequency ranges for VF are reported as being consistently lower than 100 Hz, with averages around 24 to 52 Hz [2,3], and does not differ for males and females [1]. The glottal pulses in VF can be associated with two, or even three pulses in a rapid succession followed by a period of significant vocal tract damping [3,11].

Many acoustic analyses have been conducted in temporal, spectral and cepstral domains [5,6,12]. Usual analysis methods evaluate periodicity (or harmonicity) properties using a short-term analysis frame with fixed length.

A problem of using fixed length arises in all analysis domains when VF segments have very low fundamental frequencies (i.e., very large inter-pulse time intervals). In a standard (commonly used) analysis frame length around 25 to 32 ms, it might happen that only one (or none) glottal pulse(s) lies within the analysis frame in VF segments. The presence of at least two glottal pulses within the analysis frame would be necessary to appear some harmonic structure in the spectrum, or to appear autocorrelation peaks, reflecting some short-term periodicity between glottal pulses. A simple approach could be taken by just increasing the analysis frame length. In our previous study [6], autocorrelation-based periodicity analysis was conducted using an adaptively variable frame length. However, such solution solves only part of the problem, since more than 2 glottal pulses with different inter-pulse intervals may also be present within a large analysis frame. This also would disturb the harmonic structure in the spectrum, or reduce the magnitude of autocorrelation (or cepstral) peaks.

To resolve the frame length problem, we decided to realize a glottal pulse-synchronized processing, when no periodicity can be detected within the (fixed length) analysis frame. For that purpose, candidates for glottal pulses are detected based on the damping and low fundamental frequency properties of VF.
The idea is based on the observations that damping in large inter-pulse intervals is characterized by an up and down movement in the amplitude envelope, or in a local power contour, of the speech signal.

Another problem regarding automatic detection is that most of acoustic analyses evaluate temporal or spectral features of pre-segmented voiced parts of the speech signal. In a real problem of automatic detection from the whole speech utterance including consonants and also non-speech segments (background noise), many insertion errors might occur since such segments also usually have aperiodic characteristics. Thus, the problem is how to discriminate between the aperiodicities caused by VF, and those caused by consonants or background non-speech signals.

Regarding this problem, we propose the estimation of a similarity measure between successive (or close) glottal pulses. The measure is based on an assumption that the vocal tract configuration does not change much between two successive glottal excitations, such that their vocal tract responses are expected to be similar. Figure 1 shows a block diagram of the proposed method.

![Block diagram of the basic idea of the proposed method for automatic detection of vocal fry](image)

### 3. The speech data and labels

The speech utterances used for analysis and evaluation are extracted from natural conversations between native speakers of Japanese, recorded for the JST/CREST ESP project. A small dataset of speech utterances contained in 6 conversation passages of 1 male and 3 female adult speakers is used for analysis. These are part of the analysis data used in [9]. The utterances contain VF segments with very large inter-pulse intervals in pressed voices. Another set of 60 utterances selected from data of 1 male and 1 female adult speaker is left for evaluation of automatic detection (Section 5).

A strict segmentation and categorization of vocal registers is difficult, since the transition between registers is not well defined in both physiological and perceptual viewpoints [16]. The categorization difficulty is not only between modal and VF, but also between VF and harsh voice qualities, since harsh voices also have irregularity properties in fundamental frequency. Here, a tentative labeling has been conducted by one subject with experience in voice quality, by looking at the waveforms, spectrograms and listening to the segments. Readers are invited to listen to samples of the labeled data set at: <http://www.irc.atr.jp/~carlos/vocalfry/>. Although a detailed labeling including several voice qualities has been done, for the present study, we group the segment labels in 4 global categories: “VF” (segments containing VF pulses with large inter-pulse intervals), “VF?” (segments including acoustic and perceptual features between VF and diplophonia, or between VF and modal), “R” (segments containing some irregularity in periodicity, sounding rough, but individual pulses can not be heard as in VF), and “NF” (all the rest, including modal voice, whispery voice, consonants and non-speech segments). The labeling resulted in 19 “VF?”, 30 “VF?”, and 17 “R” segments for the analysis data set, and a total of 38 “VF?”, 54 “VF?” and 91 “R” segments for the evaluation data set.

### 4. Description and analysis of the proposed acoustic measures

Before any subsequent power-related and periodicity-related analyses, the speech signal is pre-processed by band-pass filtering in a frequency range of 100 to 1500 Hz, where most information about periodicity is retained. Frequencies lower than 100 Hz are filtered out for cutting DC components and gradually rising or falling components that would disturb periodicity analysis, while frequency components above 1500 Hz are filtered out for reducing the effect of high frequency noise components.

#### 4.1. Power peak detection

Although the standard (commonly used) frame length for short-term feature extraction analysis is around 25 to 32 ms, with frame shifts of 10 or 5 ms, we decided to use a frame length of 5 ms, with frame shifts of 2.5 ms, which we call “very short-term” power analysis. These values are chosen in order to make detectable peaks and valleys in the power contour for individual glottal pulses in VF segments. As VF, consistently occurs below 100 Hz (i.e. inter-pulse intervals above 10 ms), the above conditions must be enough for detecting the power changes between individual glottal pulses in VF.

For analysis purposes, firstly, not so strict thresholds of 1 dB are imposed between the peak power and the power two frames before and after the peak, for detecting local peaks in the “very short-term” power contour. Power rising and falling features are then analyzed for the detected local power peaks. The power rising was estimated as the maximum power difference (in dB) between the peak and the four preceding frames (i.e., in an interval of about 10 ms before the peak). The power falling was estimated in the same way, for four frames following the power peak. These values are again based on the evidences that VF is consistently below 100 Hz, which means that power changes are expected to be detected within an interval of 10 ms.

Fig. 3 shows the distributions of the maximum power rising and falling values for local power peaks detected in VF segments and NF (non-fry) segments.

![Distributions of power rise and fall degrees (dB)](image)

The distributions in Fig. 3(a) show presence of large values for both powers rising and falling, reflecting the damping property of VF. NF segments show predominance of both powers rising and falling around a range of 1 to 6 dB, as shown in Fig. 3(b). However, it seems difficult to set a threshold for discriminating between VF and NF segments. Section 5 shows the performances of automatic detection setting different thresholds for these parameters.
4.2. Intra-frame periodicity (IFP)

A periodicity constraint is imposed to ensure that periodic glottal pulses are not detected within the current analysis frame, or between two excitation pulses. This is to avoid the comparison of two glottal pulses that are possibly separated by a modal segment.

Intra-frame periodicities (IFP) are obtained by doing autocorrelation analysis in a (standard) frame length of 32 ms with frame shifts of 10 ms. The IFP measure is defined as the autocorrelation value of the maximum peak, normalized by a factor of (frame length) / (frame length – lag). This normalization is for compensating the decreasing property of autocorrelation functions as the lag increases [6].

Only autocorrelation peaks whose lags are smaller than 15 ms (corresponding to F0 larger than about 66.7 Hz) are considered for periodicity analysis, which means that at least 2 complete glottal cycles are present within the analysis frame.

An important constraint of sub-harmonicity is imposed for autocorrelation peaks corresponding to F0 larger than 200 Hz. The periodicity of all sub-harmonics above 66.7 Hz is checked, in order to avoid misdetection of periodicity due to strong harmonics around the first formant, rather than a periodicity due to repetition of glottal cycles. An example showing sub-harmonic properties in the autocorrelation function is illustrated in Fig. 4, comparing VF containing only one glottal pulse within the frame and modal voice with high fundamental frequency, in vowel /e/ segments extracted from a female speaker voice. A threshold function is defined as 0.5 * (frame length – lag) / (frame length), also taking into account the decreasing property of autocorrelation functions.

For modal segments, the autocorrelation peaks of the sub-harmonics of the strong harmonic are also usually strong. The example in Fig. 4(b) shows that all sub-harmonics above 66.7 Hz (autocorrelation peaks in lags below 15 ms) are higher than the threshold function, while the VF example in Fig. 4(a) shows many sub-harmonic components with autocorrelation values smaller than the threshold function.

In practice, null values are set for periodicity measures in frames where the auto-correlation value of any sub-harmonic is lower than the threshold function. Finally, the continuity of IFP measures between frames is also checked, such that null values are also set unless a minimum of three successive frames with periodicity measure above 0.5 is detected.

The white bars in Fig. 5 show the distributions of the IFP measures for VF and NF segments. Fig. 5(a) shows a predominance of large IFP values (concentrated around 0.8 to 0.95), and only few null values in VF segments, confirming our expectations. In NF segments (Fig. 5(b)), a big value is observed in null_2, which corresponds to the null values that were set because of the search range constraint, indicating that no power peak was found in a range of 100 ms. Also, two groups in IPS can be observed in Fig. 5(b); one around low IPS values, as expected, but another around high values. However, most of these high IPS values are possibly caused by periodicity in modal segments (i.e., IPS is also high); note the presence of large IFP values (white bars) in NF segments.

5. Evaluation for automatic detection

The inter-pulse intervals, where high IPS values are detected between successive (or close) pulses, are merged in order to obtain the detected VF segments.

The automatic detection is evaluated in terms of the durations of the segments detected as VF (VFdur), compared with the intervals that are manually labeled as “VF” (VFdur_human). A ratio between VFdur and VFdur_human
is called \(V_{Fratio}\). A segment labeled as “VF” is considered as correctly detected, only if \(V_{Fratio}\) is larger than 2/3. Insertion errors are checked by counting the number of segments detected as VF, in segments non-labeled as “VF” (\(VF_{dur\_ins}\)). Here, the detection and insertion results are visualized in two groups, namely “detected” and “detected?”, according to the performance of detection or the severity of the insertion error. Here, the group “detected?” includes detected VF segments whose \(V_{Fratio}\) values are between 1/3 and 2/3, and insertions whose \(VF_{dur\_ins}\) values are below 30 ms.

Several conditions are tested for different combinations of the proposed parameters, in order to reduce insertion errors without degrading detection. First, power peak thresholds are adjusted by resetting the thresholds of IPS to 0.6 and IFP to 1.0. This corresponds to using only power information. Fig. 6(a) shows the detection results for different power thresholds.

High power thresholds reduce insertion errors (black and gray bars in “NF” group), but also reduces detection rate (black and gray bars in “VF” group). A threshold of 7 dB is found reasonable for reducing part of insertion errors, but still keeping a good detection rate in the “VF” group. Next, IFP thresholds are chosen by fixing thresholds of power to 7 dB and IPS to 0.0. Fig. 6(b) shows the detection results for different IFP values. The detection rate does not change much (“VF” group), but a threshold of 0.6 can reduce some more insertions errors (“NF” group). Finally, several IPS thresholds are tested, by fixing thresholds of power to 7 dB and IFP to 0.6 (Fig. 6(c)). An IPS threshold of 0.6 can reduce some more severe insertion errors (black bars in “NF” group), still keeping a reasonable detection rate. Regarding the group “R” (where VF features are not perceived), most of the samples are not detected as VF, while in the group “VF?” part of samples is detected as VF. These results indicate that the proposed method for automatic detection is reasonably consistent with perceptual features of VF.

6. Conclusion

We proposed a set of acoustic measures \{Power, IFP, IPS\} for automatic detection of vocal fry. The functionality of each measure was analyzed and evaluated for getting high detection rates in vocal fry segments and low insertion errors in non-fry segments. A low insertion error rate of 3.9 % and a relatively high detection rate of 73.3 % were achieved for \{Power = 7 dB; IFP = 0.6; IPS = 0.6\}. For applications where insertion errors are not so problematic, the thresholds of the parameters can be adjusted in order to get higher detection rates. Future research topics are verification of the robustness of the proposed method for environmental noise, automatic detection of other voice qualities (like harshness and breathiness), and application for detection of paralinguistic information using voice quality information.

7. Acknowledgements

This research was supported by Ministry of Internal Affairs and Communications. The speech data used in the present research was supported by JST/CREST. Special thanks for Ken-ichi Sakakibara (NTT), Parham Mokhtari (ATR), and Noriaki Mitsuanga (ATR) for advice and motivating discussions.

8. References