Available online at www.sciencedirect.com

SciVerse ScienceDirect

Speech Communication 55 (2013) 841–856

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Multi-band summary correlogram-based pitch detection for noisy speech

Lee Ngee Tan *, Abeer Alwan

Department of Electrical Engineering, University of California, Los Angeles, CA 90095, USA

Received 8 November 2012; received in revised form 19 February 2013; accepted 8 March 2013
Available online 28 March 2013

Abstract

A multi-band summary correlogram (MBSC)-based pitch detection algorithm (PDA) is proposed. The PDA performs pitch estimation and voiced/unvoiced (V/UV) detection via novel signal processing schemes that are designed to enhance the MBSC's peaks at the most likely pitch period. These peak-enhancement schemes include comb-filter channel-weighting to yield each individual subband's summary correlogram (SC) stream, and stream-reliability-weighting to combine these SCs into a single MBSC. V/UV detection is performed by applying a constant threshold on the maximum peak of the enhanced MBSC. Narrowband noisy speech sampled at 8 kHz are generated from Keele (development set) and CSTR – Centre for Speech Technology Research-(evaluation set) corpora. Both 4-kHz full-band speech, and G.712-filtered telephone speech are simulated. When evaluated solely on pitch estimation accuracy, assuming voicing detection is perfect, the proposed algorithm has the lowest gross pitch error for noisy speech in the evaluation set among the algorithms evaluated (RAPT, YIN, etc.). The proposed PDA also achieves the lowest average pitch detection error, when both pitch estimation and voicing detection errors are taken into account.

Keywords: Pitch detection; Multi-band; Correlogram; Comb-filter; Noise-robust

1. Introduction

Pitch or fundamental frequency (F0) detection is important for many speech applications. These applications include speech enhancement, synthesis, coding, source separation, and auditory scene analysis. Due to the increasing use of mobile devices, speech communication seldom takes place in a noise-free environment. Although pitch detection is a well-researched subject, and existing pitch detection algorithms (PDAs) work reasonably well for clean speech, accurate pitch detection for bandpass-filtered speech (e.g., telephone speech with G.712 filter characteristics (ITU, 1996)), and noisy speech still poses a challenge.

A pitch detector performs both pitch estimation and voiced/unvoiced (V/UV) detection. In pitch estimation, the rate of vocal-fold vibration is estimated, while in V/UV detection, voiced or quasi-periodic speech frames are distinguished from the rest of the signal. In general, pitch estimation can be performed using (1) time-domain, (2) frequency-domain, or (3) time-frequency-domain signal processing techniques. Time-domain pitch estimation exploits the signal's temporal periodicity by computing a temporal correlation or difference function directly from the signal samples. Some well-known examples of time-domain pitch estimation algorithms are RAPT (Talkin, 1995), YIN (Cheveigné and Kawahara, 2002), and the average magnitude difference function (AMDF) pitch extractor (Ross et al., 1974), which are known to give accurate pitch estimates for clean speech. Frequency-domain pitch estimation relies on the presence of strong harmonic peaks near integer multiples of F0 in the short-time spectral

* Corresponding author. Address: Electrical Engineering Department, University of California, Los Angeles, 56-125B Engineering IV Building, Box 951594, Los Angeles, CA 90095, USA. Tel.: +1 310 729 1135.
E-mail addresses: ngee@seas.ucla.edu (L.N. Tan), alwan@ee.ucla.edu (A. Alwan).

0167-6393/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.specom.2013.03.001

representation. Some examples of such frequency-domain pitch estimation algorithms are subharmonic-to-harmonic ratio (SHR) (Sun, 2002), dominant harmonics (Nakatani and Irino, 2004), and SWIPE' (SWIPE' is a variant of SWIPE that focuses on harmonics at prime integer multiples of F0) (Camacho and Harris, 2008). In time-frequency-domain pitch estimation algorithms, the input signal is typically decomposed into multiple frequency subbands, and time-domain techniques are applied on each subband signal. A popular time-frequency-domain technique is the auditory-model correlogram-based algorithm inspired by Licklider's duplex theory of pitch perception (Licklider, 1951), in which frequency decomposition is performed using an auditory filterbank (for which gammatone filterbanks (Patterson et al., 1992) are widely used), followed by autocorrelation (ACR) computation on each subband signal. The correlogram is formed by vertically stacking all ACR functions to form a 2-D image (Slaney and Lyon, 1990). Finally, the fundamental period (T0) of the signal is found by locating the ACR delay lag of the maximum peak in the "summary" correlogram (SC), which is typically the averaged ACR function. ACR can be applied either directly on the subband signal or its envelope. The latter is usually performed on mid- and high-frequency subbands only (Rouat et al., 1997; Wu et al., 2003). These subbands have sufficiently wide bandwidths to capture at least two consecutive harmonic peaks, such that the resulting filtered signals have an amplitude modulation frequency equal to F0 (a.k.a. beat frequency) (Delgutte, 1980). It has been shown that correlogram-based techniques can yield estimates close to human's perceived pitch for difficult signals with missing fundamental, inharmonic complexes and noise tones (Meddis and Hewitt, 1991; Cariani and Delgutte, 1996). Being a multi-band approach, correlogram-based techniques also tend to be more noise-robust than time-domain or frequency-domain algorithms whose parameters are fixed regardless of the signal's periodicity in the different subbands. This is because additional subband selection or weighting schemes, such as those in Rouat et al. (1997) and WWB (Wu et al., 2003), can be implemented to give less emphasis to the noise-dominated subbands. Since the filters in a gammatone filterbank are narrower and spaced more closely at lower linear frequencies than at higher frequencies (Patterson et al., 1992), the number of filters at lower frequencies (within the first 1 kHz) can be almost equivalent to the number filters in the mid and high frequencies. When the majority of harmonics at the low frequencies are attenuated due to the transmission channel characteristics or masked by strong low-frequency noise interference, it is challenging to design an effective subband selection and weighting scheme to select the reliable subband ACRs such that the maximum peak of the resulting summary correlogram yields the true pitch value.

As for V/UV detection in a pitch detector, it can be performed by either utilizing the information derived from the pitch estimation module, or using a separate module that is independent of the pitch estimation algorithm. The simplest V/UV detector is one that applies a constant decision threshold on a single degree-of-voicing feature computed by the pitch estimation module, e.g., ACR or cepstral peak amplitudes (Rabiner et al., 1976). To further improve detection accuracy, the initial V/UV decisions are usually smoothed via median filtering (Secrest and Doddington, 1982; Ahmadi and Spanias, 1999). A disadvantage of the constant-threshold scheme is that since the degree-of-voicing feature tends to be very noise-sensitive and dependent on the signal-to-noise ratio (SNR), a threshold level tuned for a particular SNR, generally does not work well at a different SNR. Thus, threshold adaptation techniques have been proposed to improve the noise-robustness of V/UV detectors. Typically, the threshold is adapted based on long-term statistics (min, max, mean, median, etc.) of degree-of-voicing-related features (Medan et al., 1991; Ahmadi and Spanias, 1999). In this case, V/UV detection performance would tend to degrade under a highly non-stationary noise condition. Dynamic programming – a tracking algorithm that integrates V/UV detection with pitch estimation, is another common technique used for pitch detection (Secrest and Doddington, 1983; Talkin, 1995; Luengo et al., 2007). A dynamic programming algorithm finds the least-cost path based on some pre-defined voicing and frequency transition cost functions, leading to performance improvements in both V/UV detection and pitch estimation through utilizing voicing and pitch information from multiple frames. However, when a constant value is used to control the voicing transition cost, such as the voice bias in Talkin (1995), the V/UV detection performance of these pitch detectors is also dependent on SNRs. Since it is generally difficult to perform noise-robust V/UV detection based on the single degree-of-voicing feature from pitch estimation (Atal and Rabiner, 1976), statistically-trained V/UV classifiers have also been proposed, especially for applications that do not require pitch estimates to be computed (e.g., speaker-independent speech recognition). This latter class of V/UV detectors, which can operate independently from pitch estimation, have reported robust V/UV detection performance, especially if their parameters are learned from noisy speech (Shah et al., 2004; Beritelli et al., 2007). In this paper, since pitch estimation is already part of a pitch detector, we are mainly interested in the former class of V/UV detectors. There are also algorithms that perform pitch-tracking using models trained on information extracted during pitch estimation. For example, hidden Markov models (HMMs) are used in WWB (Wu et al., 2003) to form continuous single or dual pitch contours for noisy speech. These data-driven algorithms yield robust voicing/pitch detection performance when the test data has characteristics that are similar to the data used for training the models.

In this paper, a multi-band summary correlogram (MBSC)-based pitch detection algorithm is proposed. This work is an extension of our previous algorithm in Tan and Alwan (2011) that has focused on pitch estimation only.