

Pitch estimation of speech and music sound based on multi-scale product with auditory feature extraction

Mohamed Anouar Ben Messaoud¹ · Aïcha Bouzid¹

Received: 6 July 2015 / Accepted: 20 November 2015 / Published online: 28 November 2015
© Springer Science+Business Media New York 2015

Abstract The pitch is a crucial parameter in speech and music signals. However, due to severe noisy conditions, missing harmonics, unsuitable physical vibration, the determination of pitch presents a great challenge when desiring to get a good accuracy. In this paper, we propose a method for pitch estimation of speech and music sounds. Our method is based on the fast Fourier transform (FFT) of the multi-scale product (MP) provided by a feature auditory model of the sound signals. The auditory model simulates the spectral behaviour of the cochlea by a gammachirp filter-bank, and the out/middle ear filtering by a low-pass filter. For the two output channels, the FFT function of the MP is computed over frames. The MP is based on constituting the product of the speech and music wavelet transform coefficients at three scales. The experimental results show that our method estimates the pitch with high accuracy. Besides, our proposed method outperforms several other pitch detection algorithms in clean and noisy environments.

Keywords Pitch estimation · Speech signal · Music sound · Auditory model · Multi-scale product

1 Introduction

Pitch detection contains important information about speech, and music sounds in the area of speech analysis, speech recognition, prosody analysis, music information

retrieval, chord recognition, automatic music transcription, and onsets detection (Gavat et al. 2002; Klapuri 2004; Bello et al. 2005; Roy et al. 2011). The acoustic music sounds are often quasi-periodic, and they present imperfect periodicity in different ways. Generally, pitch is the subjective perception of a note, and every pitched musical instrument can produce individual notes with well detectable fundamental frequencies (F_0 s) (Muller et al. 2011).

There have been many methods for pitch determination. In most cases, the methods of determination are based on the analysis of spectrum, psychoacoustic model of human hearing or a combination of them. Comparative studies of various methods are presented by Hess (Hess 1992; Klapuri 2000), who compared the methodologies and performance of each method.

In the time domain, many methods of pitch estimation apply the autocorrelation function (ACF) (Brown and Zhang 1991), and average magnitude difference function (AMDF) (Li et al. 2006). Another significant variety, for instance, the approach introduced by De cheveigné (De Cheveigné and Kawahara 2002), implements a modified autocorrelation to analyze the speech signal and applies an operation of subtraction using in AMDF, a parabolic interpolation and a cumulative mean normalization, to lessen error rate. However, the methods based on ACF and AMDF tend to determine the two times of true period, and it produces a sub-harmonic error in the spectral domain.

In the frequency domain, the commonly used transformation is the spectrum, and the cepstrum. The one constraint of these techniques, however, is that it attributes the same weight to all harmonic frequencies, which will be tending to twice octave error (Klapuri 2000). In addition, methods based on spectrum autocorrelation have been proposed, such as logarithmic spectrum ACF (Kunieda et al. 1996). Unfortunately, a major limitation of these

✉ Mohamed Anouar Ben Messaoud
anouar.benmessaoud@yahoo.fr

Aïcha Bouzid
bouzidacha@yahoo.fr

¹ Department of Electrical Engineering, National Engineering School of Tunis, Tunis, Tunisia

methods is that they will result in twice octave error when predominant harmonics exist. As the above declared methods deal with harmonic position estimators. Examples of other more efficient harmonic position estimators are sub-harmonic to harmonic ratio (SHR) by (Sun 2000), and sawtooth waveform method (SWIPE) by (Camacho and Harris 2008).

To recapitulate, existence of salient harmonics, missing harmonics, and other different challenges finding in speech and music signals render the efficient estimation of F_0 very hard. Furthermore, the above presented methods are not able to determinate the F_0 of imperfect harmonic sound signals by the fact that due to the non-stationarity and quasi-periodicity of the speech sound, the physical vibration, the harmonics of the signals produced from the musical instrument cannot be spaced with correct interval, but marginally shift from perfect positions.

Although a large number of fundamental frequency determination algorithms have been described in the state-of-the-art for clean sound, fundamental frequency determination from a noisy sound has been essayed only by a few researchers (Shimamura and Kobayashi 2001; Shahnaz et al. 2007; 2008). However, in practical applications, a fundamental frequency determination task has to be performed using only the given noisy sound. For example, for speech separation in noisy environments, extraction of fundamental frequency of the dominant speech is required as it can be used as a cue for separation of concurrent speech. In (Mahmoodzadeh et al. 2012), experiments have been carried out as an attempt to separate signal from a background noise based on computational auditory scene analysis (CASA). The fundamental frequency determination algorithm employed in (Mahmoodzadeh et al. 2012) for determining the concurrent time–frequency region of the desired sound is based on instantaneous amplitude comb filtering and can determinate the F_0 of vowels in noisy environments but the estimates are not accurate enough.

In the weighted autocorrelation method (Shimamura and Kobayashi 2001), using the same periodicity property of AMDF and ACF, the ACF is weighted by the reciprocal of the AMDF in order to emphasize the true pitch-peak for noisy sound. Since, under a high level of noise, the global maximum of AMDF or the global minimum of ACF may occur at a lag that is a multiple or sub-multiple of true pitch period, in the weighted ACF, the peaks at non-pitch locations may be wrongly emphasized more than those at the true pitch location. This causes inaccurate fundamental frequency determination at a low SNR. It is worth mentioning that most of the fundamental frequency determination algorithms reported determination performance for sound corrupted by white noise only. In general, fundamental frequency determination performance would

deteriorate significantly in a scenario, where sound is corrupted by a realistic noise (Prasanna and Yegnanarayana 2004), but there is a growing demand of many practical applications in which it is important to determinate fundamental frequency accurately from the sound corrupted by the realistic noise, such as multi-talker babble noise with very low levels of SNR.

The objective of this work is to present a robust approach that effectively overcomes the limitations of the existing fundamental frequency determination under severe noisy conditions and simultaneously be efficient in the clean speech and monophonic music.

To this end, this work presents an efficient method for pitch estimation of speech and music signals based on spectrum of multi-scale product of gammachirp auditory filter-bank (GAMMA-MP). The proposed paper is based on the auditory structure characteristics, and depends on both the spectral behaviour of the cochlea and the out/middle ear filtering. The out/middle ear filtering is designed by a low-pass filter. The cochlear filter is designed by a gammachirp filter-bank. The F_0 is next deduced in each channel by calculating its fast Fourier transform of multi-scale product analysis. Then, we extract the best candidate as the F_0 determination. In this paper, the F_0 is determined in all voiced segments of the evaluated databases. So, the study of the voiced decision methods is the objective of future work.

The rest of the paper is organized as follows. Section II provides the details of the approach. Experimental evaluations and discussions are described in Section III. Finally, Section IV concludes the research work and presents future directions.

2 The pitch estimation method of speech and music sounds

The overall approach is decomposed on two stages. A block diagram of the proposed approach is depicted in Fig. 1. In the first stage, a sound signal is analyzed by a gammachirp filter bank. This processing results in a decomposition of the sound signal into a time frequency map. Then, the output sound is split into two channels, the channel below 1 kHz is filtered by a low-pass filter and the high-channel sound is half-wave rectified and low-pass filtered. In the second stage, we compute the multi-scale product in frequency domain in each channel to extract the fundamental frequency F_0 .

2.1 First stage: Auditory feature extraction

The auditory filter modeling corresponds to the mathematical model which tends to simulate the psychophysical aspects and basic perceptual of the human auditory

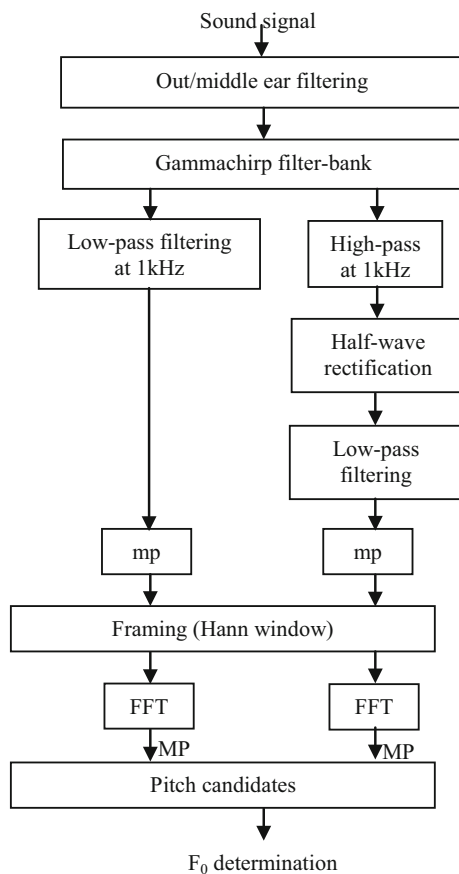


Fig. 1 Block diagram of the proposed method for pitch estimation of sound signal

characteristics (Lyon et al. 2010). The concept of fundamental frequency has a complex relationship to physical properties of the speech and makes reference to auditory perception. Thus, it is natural to apply approaches that assume human perception. The peripheral auditory model using time-domain processing of periodicity properties can simulate many features of pitch perception which are often considered to be more central (Meddis and O’Mard 1997).

First, the sound signal passes through a second-order low-pass filter as an out-middle ear filtering (Van Immerseel and Martens 1992). The main motivation of using an out-middle ear filter is to increase the strength of high frequency harmonics. As a result, the harmonics have relatively the same amplitude. It’s based on the transfer filter below:

$$H(x) = \frac{\omega_0^2}{\omega_0^2 + 0.33\omega_0x + x^2} \tag{1}$$

where $fr = 2\pi/\omega_0$ is the resonance frequency equal to 4 kHz.

Then, we apply a gammachirp filter-bank to simulate the sound processing in the cochlea. It is an extension of the gammatone filter with a frequency modulation factor. Also,

it allows to determine an approximation of the basilar membrane frequency of the cochlear (Patterson et al. 2003; Irino and Patterson 2006). The impulse response of a gammachirp filter is defined by the equation below:

$$g(t) = At^{k-1}e^{-2\pi mERB(p_0)t}e^{j2\pi p_0t+jr \ln(t)+j\varphi} \tag{2}$$

where $t > 0$, r , p_0 , φ , and A are the chirp rate, the asymptotic frequency, the phase and the amplitude, respectively. The parameters k , m designate the gamma envelope. Based on the work of (Tolonen and Karjalainen 2000), the sound is separated into two channels, below and above 1 kHz. The low-channel sound is low-pass filtered and the high-channel sound is half-wave rectified and low-pass filtered. The high channel is phase-sensitive since it follows the amplitude envelope of the sound in the frequency band above 1000 Hz. Thus, all phase-sensitivity in our approach is inherently caused by the high channel. This is different from the standard system where all channels are phase-sensitive since they follow the envelope of the sound in the corresponding frequency band.

The $ERB(p_0)$ is the equivalent rectangular bandwidth (ERB) of the gammachirp auditory filters centred around p_0 (Wang and Brown 2006). The ERB is presented by the following equation:

$$ERB(p) = 24.7 + 0.108p. \tag{3}$$

The ERB-rate scale describes an logarithmic function which associates the frequency value to the ERBs number. ERB – rate (p) is presented as follow:

$$ERB - rate(p) = 21.4 \log_{10} \left(\frac{4.37}{1000}p + 1 \right). \tag{4}$$

The low and high channels have a different time response. Since high frequency channel has a faster response than low frequency channel, a channel alignment is needed in time domain.

2.2 Second stage: Multi-scale product in frequency domain

In recent years, the wavelet transform (WT) has been successful used in many speech processing applications. WT can analyze time–frequency characteristics of sound, and can track abrupt changes of sound. So it becomes a powerful tool for F_0 determination. The WT shows whether details of a certain scale are introduced in a sound and quantifies their respective participation. Generally, the WT is meant to offer good frequency resolution at low frequencies. They have sets of properties, including: null moments to determine the useful information, uncorrelated coefficients to lessen the temporal correlation, and compact support to ensure local analysis. The quadratic spline wavelet (QSW) is introduced by Mallat (1999). An

appropriately chosen wavelet for discontinuity detection is a wavelet that is the second derivative of a smoothing function corresponding to the QSW. We denote a shifted and dilated version of the QSW by:

$$\Psi_{2^{j+1},i}(x) = \sqrt{1/2^j} \Psi\left(x - i/2^j\right) \quad (5)$$

where i , and 2^{j+1} represent the translation and the dilatation factors respectively.

We have used the Dyadic WT. It is the special case of Continuous WT.

The dyadic WT is described by the equation below:

$$W_{2^{j+1},i}f(x) = f \times \Psi_{2^j,i}(x) \quad (6)$$

where $j = -1, 0$, and 1 .

By using multi-scale products (mp), the extrema due to abrupt transitions are reinforced because of their correlated presence across scales while those due to noise are suppressed because noise is mostly confined to lower scales.

This idea was first applied to magnetic resonance images by Xu (1994). In our paper, we integrate information of some scales to extract feature points.

The MP consists of multiplication of WT coefficients of the function $f(i, s_j)$ at some successive dyadic scales as follows:

$$mp(i) = \prod_{j=1}^i Wf(i, s_j) \quad (7)$$

where $Wf(i, s_j)$ is the WT of the sound signal f at scale 2^j .

The Fig. 2 recapitulates the steps of the mp.

In each output filter channel, the product $mp(i)$ is divided into frames with a sliding analysis window $d(i)$.

For the second step, the product $mp(i)$ is divided into frames by multiplication with a sliding analysis window $d(i)$:

$$mp_d(i, l) = mp(i)d(i - j\Delta i) \quad (8)$$

where Δi is the overlap, and l is the window index.

Then, we compute the Fast Fourier transform (FFT) of multi-scale product for each channel.

The effect of the FFT of mp analysis is shown in Figs. 3 and 4.

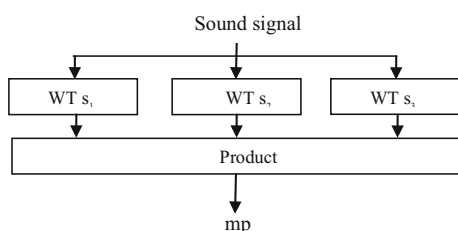


Fig. 2 Block diagram of the sound signal multi-scale product

The FFT function of each weighted block mp_d is given by:

$$MP_d^l(k) = \sum_{i=1}^N mp_d(i, l) e^{-j2\pi ik/N}. \quad (9)$$

After this, we measure the harmonic summation on the FFT of multi-scale product (HSMP) in the l th frame. It consists to summarize the order of dominance for all harmonic elements at each frame.

The HSMP for the l th peak of the FFT of MP is defined as:

$$HSMP(pc_i, t) = \sum_{m=1}^{mpc_i < p_{\max}} MP(h(mpc_i), t) \quad (10)$$

where pc_i is a pitch candidate, mpc_i is the frequency of its m th harmonic element, and p_{\max} is equal to 1700. The function $h(mpc_i)$ transfers mpc_i to the center frequency of the nearest FFT of MP bin. Then, we find the frequency that maximizes the $HSMP(pc_i, t)$ as the fundamental frequency.

Figure 3 shows a clean voiced speech signal followed by its multi-scale product (mp) and the fft of mp.

Figure 4 shows a monophonic music signal followed by its multi-scale product (mp) and the fft of mp.

The Figs. 3 and 4 show the efficiency of the MP method for pitch estimation. In Fig. 3c and 4c, the obtained signal shows spectral rays. The first element corresponds to the fundamental frequency F_0 . The following rays correspond to the fundamental frequency harmonics.

Figure 5 represents a noisy voiced speech sound corrupted by a White noise at -5 dB followed by its multi-scale product (mp) and the fft of mp.

The mp in Fig. 5b lessens the noise effects leading to an FFT function with clear maxima giving the F_0 determination (see Fig. 5c).

3 Experiments ad evaluation

Performance evaluation of our approach for pitch estimation in the case of Keele database (Meyer et al. 1995) and monophonic music uses the Musical Instruments Samples (University of Iowa 2012). The Keele database contains 10 speakers sampling frequency of 20 kHz. It contains a reference fundamental frequency determination and voiced/unvoiced segmentation of 25.6 ms segments with 10 ms overlapping. The reference fundamental frequency determination of Keele database is based on a simultaneously recorded speech and signal of the laryngograph signal. The F_0 of all compared methods are determined in each reference voicing frame.

The Musical Instruments Samples consists of 4000 notes, and one hundred and 50 min of sound composed by

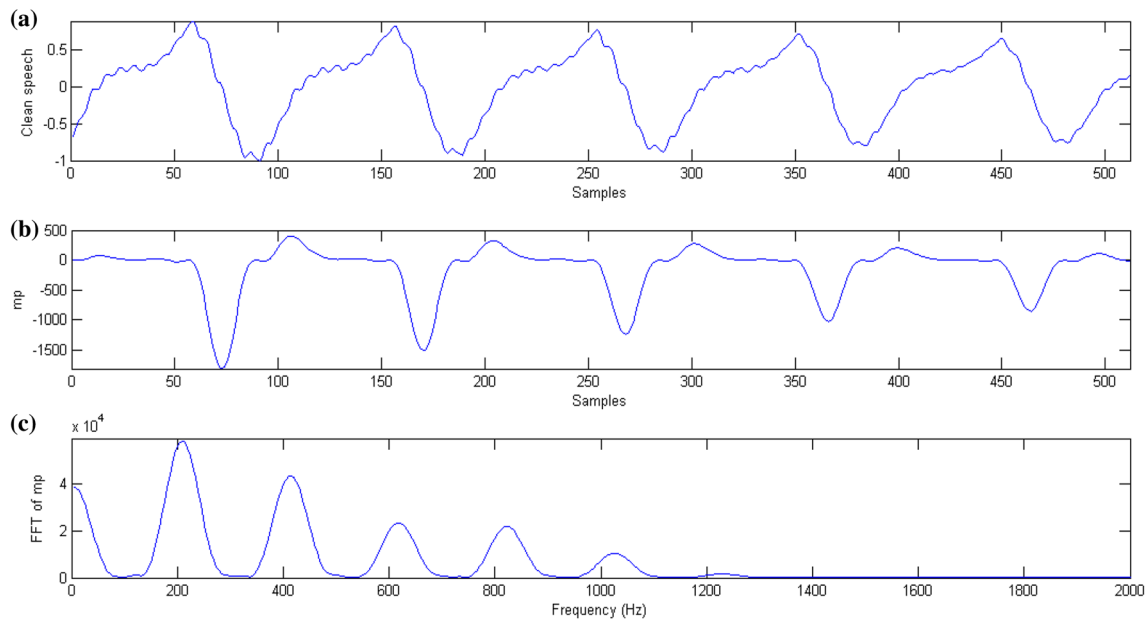


Fig. 3 Pitch estimation of speech sound. **a** Speech sound frame, **b** multi-scale product analysis (mp), **c** FFT of mp

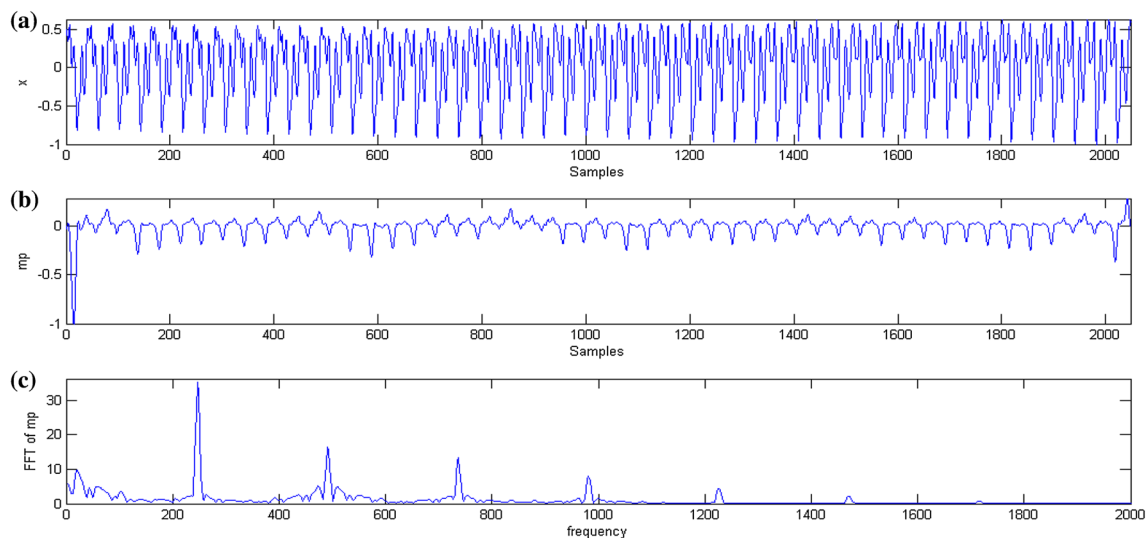


Fig. 4 Pitch estimation of music sound. **a** Music sound frame, **b** multi-scale product analysis (mp), **c** FFT of mp

twenty different musical instruments. All the music sound signals were sampled at a rate of 44.1 kHz and down-sample it to 10 kHz. The notes are given in sequence employing a chromatic scale. Each document usually covers one octave and is identified with the name of the instrument, the initial and final notes. The documents of musical database were separate into files containing a single note without silence. For this purpose, we use an automatic segmentation method, and then testing the quality of the segmentation (Ben Messaoud et al. 2015).

We apply the gross pitch error (GPE) criteria and the root mean square error (RMSE) measures to evaluate the

pitch estimation performance. A GPE is identified when the estimated fundamental frequency value is 20 % higher or lower than the reference one. The RMSE is defined as square root of the average squared estimation error with estimation errors which are smaller than the GPE threshold of 20 Hz. It is used for evaluate the speech sound.

For all compared approaches, we use a default pitch search range is 50–800 Hz (30–1700 Hz) respectively for speech signal (music sound). Each of the methods was asked to give a fundamental frequency determinate every millisecond, using the default settings of the method. In

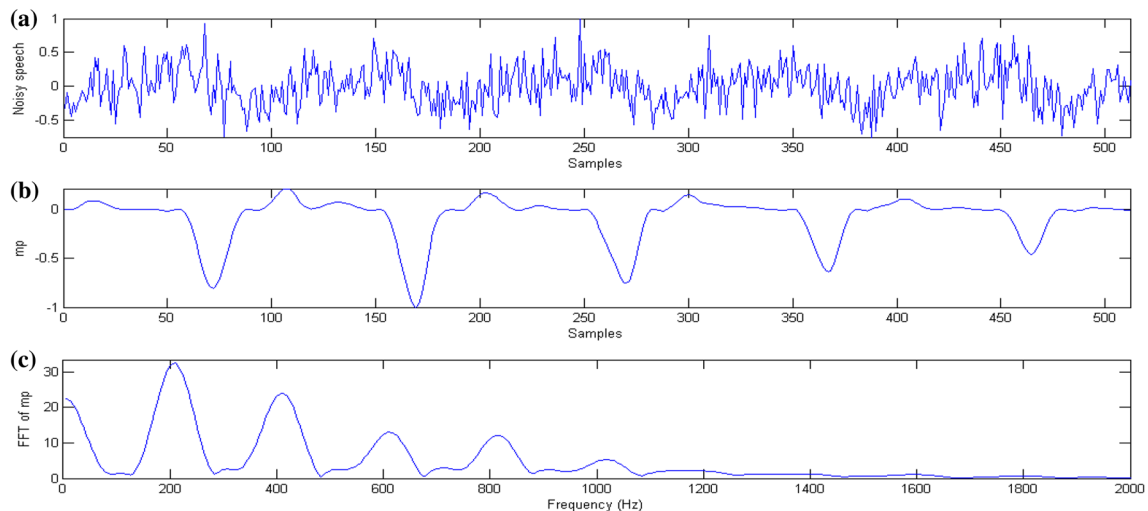


Fig. 5 Pitch estimation of speech sound corrupted by a -5 dB White noise. **a** Speech sound frame, **b** multi-scale product analysis (mp) of the noisy voiced speech, **c** FFT of mp

this work, we follow the recommendations suggested by the authors of the algorithms:

The SWIPE method is based on a sawtooth waveform in frequency domain. $[p,t] = \text{swipe}(x, fs, [50\ 800], 0.001, 1/96, 0.1, -\text{Inf})$;

The TEMPO method applies the instantaneous frequency of the results of a filter-bank. It's tested only with monophonic music. $f_0 \text{ raw} = \text{extraight source}(x, fs)$;

The YIN algorithm is based on computing the normalized difference function and a parabolic interpolation. $p.\text{min } f_0 = 50$; $p.\text{max } f_0 = 800$; $p.\text{hop} = 20$; $p.\text{sr} = fs$; $r = \text{yin}(x, p)$;

The SHR method applies the subharmonic-to-harmonic ratio. $[t, p] = \text{shrp}(x, fs, [50\ 800], 40, 1, 0.4, 1250, 0, 0)$;

3.1 Results in clean speech and monophonic music sound

Table 1 presents the GPE estimation and RMSE measures of the proposed approach (GAMMA-MP), the SWIPE (Camacho and Harris 2008), the YIN (De Cheveigné and Kawahara 2002), and SHR (Sun 2000) for speech database.

For all the compared methods, the fundamental frequency determined in each reference voicing frame of reference Keele database and exactly in the same frames.

The GAMMA-MP shows a reduced GPE rate of 0.64 % and a low RMSE of 1.68 Hz. It's obviously more accurate than the other methods.

Table 2 illustrates the GPE of over estimation and under estimation of the proposed approach (GAMMA-MP), the SWIPE (Camacho and Harris 2008), the TEMPO (Kawahara et al. 1999) the YIN (De Cheveigné and Kawahara 2002), and SHR (Sun 2000) for musical instrument database.

Table 1 Fundamental frequency estimation for speech sound

Methods	GPE (%)	RMSE (Hz)
GAMMA-MP	0.64	1.68
SWIPE	0.62	3.05
YIN	2.35	3.62
SHR	1.41	1.89

Table 2 shows that GAMMA-MP has the lowest GPE in both over estimation, and under estimation. SWIPE and YIN perform better than TEMPO, while SHR produces the largest GPE over the whole database.

In Table 2, the GAMMA-MP appears as the most accurate approach for pitch estimation of musical instrument.

Table 3 presents the GPE results by instrument group. We have classed the musical instruments in five groups. The group bowed contains cello, violin, double bass, and viola. The group brass contains bass, trumpet, trombones, tuba, and French horn. The group plucked contains violin, and double bass. The group woodwinds contain clarinets, saxophones and flutes. The last group contains piano.

In Table 3, our approach performs better than other methods except for the plucked, for which TEMPO gives

Table 2 Fundamental frequency estimation for musical instrument

Methods	Over estimation	Under estimation	Total
GAMMA-MP	0.12	0.87	0.99
SWIPE	0.10	1.00	1.20
TEMPO	0.83	1.60	2.43
YIN	0.29	1.70	1.99
SHR	5.30	15.00	20.30

Table 3 Fundamental frequency estimation for musical instrument by instrument group

Methods	Piano	Bowed	Brass	Plucked	Woodwind
GAMMA-MP	1.05	0.08	0.00	2.63	0.10
SWIPE	2.20	0.19	0.01	8.80	0.14
TEMPO	7.30	2.60	0.00	0.49	1.40
YIN	0.36	1.50	0.03	14.00	1.10
SHR	19.00	25.00	14.00	22.00	8.00

practically no error. On the other hand, SWIPE performance on piano is relatively bad compared to correlation based algorithms. The brass group obtained the fewer GPE errors. However, the bowed, and plucked group have given the most GPE errors, it may be caused by pizzicato sounds.

Table 4 shows the GPE for the musical instrument by octave.

As depicted in Table 4, the GAMMA_MP approach presents the best performance.

3.2 Results in noisy speech

To test the robustness of our algorithm, we add various background noises (white, babble, and vehicle) at three SNR levels to the Keele database speech signals. For this, we use the noisex-92 database (Varga 1993).

Table 5 illustrates the GPE of GAMMA-MP, SWIPE, YIN and SHR methods in a noisy environment.

Table 4 Fundamental frequency estimation for musical instrument by octave

Methods	46.2 Hz ± 1/2 octave	92.5 Hz ± 1/2 octave	185 Hz ± 1/2 octave	370 Hz ± 1/2 octave	740 Hz ± 1/2 octave	1480 Hz ± 1/2 octave
GAMMA-MP	0.85	0.74	2.46	0.95	0.03	0.42
SWIPE	1.20	1.00	2.30	0.89	0.13	0.29
TEMPO	15.00	2.80	2.00	1.10	0.52	0.31
YIN	3.20	0.95	5.30	1.80	0.69	0.96
SHR	37.00	0.60	1.80	27.00	70.00	81.00

Table 5 Pitch estimation Performance of GPE in a noisy environment

Type of noise	SNR level (dB)	GAMMA-MP	SWIPE	YIN	SHR
White	5	0.79	1.39	1.27	1.93
	0	0.93	2.08	1.84	2.85
	-5	1.04	5.68	2.65	5.63
Babble	5	1.27	2.74	1.97	2.08
	0	1.69	2.96	3.01	4.73
	-5	2.57	4.05	5.33	6.29
Vehicle	5	3.59	3.94	4.02	4.19
	0	4.50	4.63	5.46	5.80
	-5	5.84	6.03	6.24	7.45

Table 6 Execution time results of our approach

Stage	Time (s)
Out/middle ear filtering	0.35
Gammachirp filter	3.14
Half-wave rectification	4.29
Alignement	0.08
Multi-scale product	1.57
FFT	6.08
HS	4.22
Total	19.73

As depicted in Table 5, when the SNR level decreases, our proposed approach remains robust even at -5 dB in hard situations.

As seen, the GPE of SWIPE method degrades with the Babble and white noises at -5 dB. This can be explained by the fact that the SWIPE method doesn't consider the weak voicing state like in the beginning and the end of any voiced sound. However, our proposed approach has the highest performances in all cases, which proves our Fig. 5.

3.3 Computational complexity of our approach

The proposed approach has only two channels and does not attempt directly to follow human resolvability. The approach produces similar and comparable results to those

of an elaborate multi-channel pitch analysis models. The computational demands of multi-channel F_0 analysis models have prohibited their application in practical cases (Meddis et al. 2010). The computational complexity is mostly determined by the number of channels used in the auditory filter-bank. In this paper, we have presented a suitable model of pitch perception in practical applications. Computational efficiency was shown by testing our approach on a 2.13 GHz Core Duo processor.

Table 6 presented the obtained results. For every file, the total execution time of all stages is equal approximately to 20 s.

4 Conclusions

The proposed method (GAMMA-MP) estimates the fundamental frequency of speech and music sounds. It is based on a new auditory feature extraction technique method combined with a multi-scale product analysis in frequency domain. This auditory model consists of applying the out-middle ear filtering and the cochlea behaviour in frequency domain by a gammachirp filter-bank, where the values of those centre frequencies are selected in accordance to the equivalent rectangular bandwidth. For the two channels, the obtained sound signal is divided into frames, and each frame is weighted by a Hamming window. Next, we calculate the fast Fourier transform of each multi-scale product of weighted frame. Finally, a harmonic summation technique is applied to determine the fundamental frequency F_0 . The experimental results show the efficiency of our proposed method for pitch estimation from a large speech and musical instrument database, and its high accuracy compared with the state-of-the-art methods. Future work may address the extension of the proposed method to the determination of F_0 for multi-talker speech, and polyphonic music sounds.

References

- Bello, J. P., Daudet, L., Abdallah, S., & Duxbury, C. (2005). A tutorial on onset detection in music signals. *IEEE Transactions on Speech, Audio Processing*, 13, 1035–1048.
- Ben Messaoud, M. A., Bouzid, A., & Ellouze, N. (2015). Automatic segmentation of the clean speech signal. *World Academy of Science, Engineering and Technology International Journal of Electrical, Computer, Electronics and Communication Engineering*, 9, 114–117.
- Brown, J., & Zhang, B. (1991). Musical frequency tracking using the methods of conventional and 'narrowed' autocorrelation. *Journal of the Acoustic Society of America*, 89, 2346–2354.
- Camacho, A., & Harris, J. (2008). A sawtooth waveform inspired pitch estimator for speech and music. *Journal of the Acoustic Society of America*, 124, 1638–1652.
- De Cheveigné, A., & Kawahara, H. (2002). YIN, a fundamental frequency estimator for speech and music. *Journal of the Acoustic Society of America*, 111, 1917–1930.
- Gavat, I., Zira, M., & Sabac, B. (2002). Pitch estimation by block and instantaneous methods. *International Journal of Speech Technology*, 5, 269–279.
- Hess, W. J. (1992). Pitch and voicing determination. In S. Furni, M. Sondhi, & M. Dekker (Eds.), *Advances in speech signal processing*. New York: Marcel Dekker, Inc.,
- Irino, T., & Patterson, R. D. (2006). A dynamic compressive gammachirp auditory filterbank. *IEEE Transactions on Audio, Speech and Language Processing*, 14, 2222–2253.
- Kawahara, H., Katayose, H., De Cheveigné, A., & Patterson, R. D. (1999). Fixed point analysis of frequency to instantaneous frequency mapping for accurate estimation of F_0 and periodicity. *Proceedings 6th EUROSPEECH* (pp. 2781–2784).
- Klapuri, A. (2000). Qualitative and quantitative aspects in the design of periodicity estimation algorithms. *European signal processing conference proceedings* (pp. 2069–2072).
- Klapuri, A. (2004). Automatic music transcription as we know it today. *Journal of New Music Research*, 33, 269–282.
- Kunieda, N., Shimamura, T., & Suzuki, J. (1996). Robust method of measurement of fundamental frequency by aclos: autocorrelation of log spectrum. *International conference on acoustics, speech, and signal processing proceedings* (pp. 232–235). Atlanta, GA.
- Li, H., Dai, B., & Lu, W. (2006). A pitch detection algorithm based on AMDF and ACF. *International conference on acoustics, speech and signal processing proceedings*. Toulouse (pp. 377–380).
- Lyon, R. F., Katsiamis, A. G., & Drakakis, E. M. (2010). History and future of auditory filter models. *Proceedings of 2010 IEEE international symposium on circuits and systems (ISCAS)* (pp. 3809–3820).
- Mahmoodzadeh, A., Abutalebi, H. R., Soltanian-Zadeh, H., Sheikhzadeh, H. (2012). Single channel speech separation with a frame-based pitch range estimation method in modulation frequency. *International symposium on telecommunications* (pp. 609–613).
- Mallat, S. (1999). *A wavelet tour of signal processing*. San Diego: Academic Press.
- Meddis, R., Lopez-Poveda, E. A., Fay, R. R., & Popper, A. N. (2010). *Computational models of the auditory system*, Springer Handbook of Auditory Research New York: Springer.
- Meddis, R., & O'Mard, L. (1997). A unitary model for pitch perception. *Journal of the Acoustic Society of America*, 102, 1811–1820.
- Meyer, G., Plante, F., & Ainsworth, W. A. (1995). A pitch extraction reference database. 4th European Conference on Speech Communication and Technology. EUROSPEECH'95, Madrid, pp. 837–840.
- Muller, M., Ellis, D., Klapuri, A., & Richard, G. (2011). Signal processing for music analysis. *IEEE Journal of Selected Topics in Signal Processing*, 5, 1088–1110.
- Patterson, R. D., Unoki, M., & Irino, T. (2003). Extending the domain of centre frequencies for the compressive gammachirp auditory filter. *Journal of the Acoustic Society of America*, 114, 1529–1570.
- Prasanna, S. R. M., & Yegnanarayana, B. (2004). Extraction of pitch in adverse conditions. *International conference on acoustics, speech and signal processing proceedings* (pp. 109–112).
- Roy, S. J., Molla, M. K. I., Hirose, K., & Hasan, M. K. (2011). Harmonic modification and data adaptive filtering based approach to robust pitch estimation. *International Journal of Speech Technology*, 14, 339–349.
- Shahnaz, C., Zhu, W. P., & Ahmad, M. O. (2007). A robust pitch estimation algorithm in noise. *International conference on*

- acoustics, speech and signal processing proceedings* (pp. 1037–1076).
- Shahnaz, C. Zhu, W. P., & Ahmad, M. O. (2008). A pitch extraction algorithm in noise based on temporal and spectral representations. *International conference on acoustics, speech and signal processing proceedings* (pp. 4477–4480).
- Shimamura, T., & Kobayashi, H. (2001). Weighted autocorrelation for pitch extraction of noisy speech. *IEEE Transactions on Speech and Audio Processing*, 9, 727–730.
- Sun, X. (2000). A pitch determination algorithm based on subharmonic-to-harmonic ratio. *International conference on spoken language processing proceedings* (pp. 676–679). Beijing.
- Tolonen, M., & Karjalainen, M. (2000). A computationally efficient multipitch analysis model. *IEEE Transactions on Speech and Audio Process*, 8, 708–716.
- University of Iowa. (2012). Electronic music studios. <http://theremin.music.uiowa.edu>.
- Van Immerseel, L. M., & Martens, J. P. (1992). Pitch and voiced/unvoiced determination with an auditory model. *Journal of the Acoustic Society of America*, 91, 3511–3526.
- Varga, A. (1993). Assessment for automatic speech recognition: II. Noisex-92: A database and an experiment to study the effect of additive noise on speech recognition systems. *Elsevier Speech Communication*, 12, 247–251.
- Wang, D. L., & Brown, G. J. (2006). *Principles, computational auditory scene analysis: Algorithms, and applications*. Hoboken, NJ: Wiley/IEEE Press.
- Xu, Y., Weaver, J., Healy, D., & Lu, J. (1994). Wavelet transform domain filters: A spatially selective noise filtration technique. *IEEE Transactions on Image Processing*, 3, 747–758.