TOWARD THE DETECTION OF INTERPRETATION EFFECTS AND PLAYING DEFECTS

Antony Schutz, Dirk Slock[*]

EURECOM
Mobile Communication Department
2229 Route des Crêtes BP 193, 06904 Sophia Antipolis Cedex, France
firstname.lastname@eurecom.fr

ABSTRACT

Precise automatic music transcription requires accurate modeling and identification of the spectral content of the audio signal. But music can not be reduced to a succession of notes, and an accurate transcriptor should be able to detect other performance characteristics, such as slow tempo variations or, depending on the instrument detecting some interpretation effects. In a pedagogic way a student could want to improve his level and a good challenge will be to estimate the quality of play of musician. In this paper we present some of the most common playing defects and interpretation effects and we propose a way for detecting them.

Index Terms— Automatic ornemantation, signal processing, Harmonic analysis, Acoustic applications, peaks subtractions

1. INTRODUCTION

Music transcription is the process of creating a musical score (i.e. a symbolic representation, such as a MIDI file, of the music within) from an audio recording. In the traditional sense, automatic transcription implies the estimation of several features such as the pitch and duration of individual notes.

But music can not be reduced to a succession of notes, and an accurate transcriptor should be able to detect other performance characteristics, like interpretation effects. The tools built for automatic transcription can also be used in a pedagogic way such as a student could want to improve his level with the help of a software. This means that the software should be able to detect some defects. For a violin it can be used for improving the use of the bow, for a wind intrument it can be the constancy of the blow.

In this paper we want to review some interpretation effects of some instruments and playing defects to adress a way for detecting and evaluating them. We want to emphasize the utility of the instrumental noise for this kind of study.

The paper is organized as follow first we explain the used harmonic plus noise decomposition for the problem, then we will describe some interpretation effects, playing defects and some associated characteristics. After we present some results about their detections, finally we conclude.

2. HARMONIC PLUS NOISE DECOMPOSITION

2.1. Model

Musical signals are often modeled as the sum of sinusoids [1] and the estimation of their parameters has been dealt with extensively in the literature [2, 3]. We consider the estimation of the parameters of a sinusoidal signal $s(t)$ given by :

$$s(t) = x(t) + n(t), \tag{1}$$

$$x(t) = \sum_{n=0}^{N-1} A_n(t)\cos(2\pi\frac{f_n(t)}{f_s} + \phi_n(t)) \tag{2}$$

Here, $A_n(t)$, $f_n(t)$ and $\phi_n(t)$ are the amplitude, the frequency and the phase of the partial n of the signal at time t, respectively. The sinusoidal part is defined by $x(t)$, the noise part by $n(t)$ and f_s is the sampling frequency.

2.2. Method

The musical signal, which is by nature non-stationary, is piece-wise analyzed [1]. The synthesis method consists of estimating the parameters of each frame, generating each partial signal by using the purely sinusoidal model and then reforming the complete signal by using an overlap and add method. The noise is extracted by subtracting the synthesized signal from the original noisy signal.

The parameters are estimated by peak picking in the spectrum [2], but the estimation is rather bad. For obtaining better performance quadratic interpolation is performed on each peak [4], but due to the interferences from the others peaks it remains a bias. The first task is to find the Nb principle peaks in the spectrum. Taking a fixed value for peaks has some drawbacks: In the case of pure noise, the sinusoidal signal part will be estimated by the noise's dominant peaks and

[*]Institut Eurécom's research is partially supported by its industrial members: BMW Group Company, Bouygues Téécom, Bouygues Telecom, Cisco Systems, ThalesFrance Téécom, Hitachi, SFR, Sharp, ST Microelectronics, Sharp, Swisscom, Texas Instruments, Thales.

the resulting estimate SNR will be bounded at a lower value. Equivalently, for a rich spectrum there will be some harmonics in the noise. After we have found the peaks, we choose one and we calculate the interference of the peaks include into the ± Δ_f interval. For each peak of this interval, we estimates its frequency and amplitude by parabolic interpolation and then we calculate its phase by linear interpolation. When the interference is cleaned from the peak of interest, we interpolate its parameters. For the parabolic interpolation we use:

$$Y_{m'} = S_{dB}(f_m + m'), \qquad m' = -1, 0, 1 \tag{3}$$

Where $S_{dB}(f) = 20\, log_{10}(|X(f)|)$, and $X(f)$ is the Fourier Transform of $x(t)$. The estimate frequency is given by

$$f^{est}_m = f_m + \frac{1}{2}\,\frac{Y_{+1} - Y_{-1}}{Y_{-1} + Y_{+1} - 2Y_0}, \tag{4}$$

and the corresponding amplitude by

$$S^{est}_{dB} = Y_0 - \frac{f^{est}_m}{4}\,(Y_{-1} - Y_{+1}),\ A^{est}_m = 10^{\frac{1}{20}\ S^{est}_{dB}}. \tag{5}$$

For estimating the interference of the nearest peak, we have to obtain an expression of the perturbation due to the presence of other peaks in the spectrum on the peak of interest. In our case, we use a Hann window of size L given by :

$$w(n) = 0.5 - 0.5\ \cos(2\pi\frac{n}{L}), \qquad 0 \leq n < L \tag{6}$$

The Hann window is temporally finished, so we express it with the rectangular window :

$$r(n) = \begin{cases} 1 & ,0 \leq n < L \\ 0 & ,otherwise \end{cases} \tag{7}$$

We can rewrite :

$$\begin{aligned} w(n) &= [0.5 - 0.5\ \cos(2\pi\frac{n}{L})]\ r(n) & (8) \\ &= 0.5\, r(n) - 0.25\, e^{2i\pi\frac{n}{L}}\ r(n) - 0.25\, e^{-2i\pi\frac{n}{L}}\ r(n) \end{aligned}$$

The DFT of the rectangular function is :

$$R(f) = \sum_{t=0}^{L-1}(e^{-2i\pi f t}) = e^{-i\pi f(L-1)}\ \frac{sin(\pi f L)}{sin(\pi f)} \tag{9}$$

So for the Hann window we obtain :

$$W(f) = 0.5\ R(f) - 0.25\ R(f - \frac{1}{L}) - 0.25\ R(f + \frac{1}{L}) \tag{10}$$

After the estimation of the parameters of each peak, we subtract their contribution given by :

$$\begin{aligned} W^{est}_m(f) = \sum_{\substack{n=1 \\ n \neq m}}^{Nb_{\in \Delta f}} A^{est}_n W(f - f^{est}_n) e^{i\phi^{est}_n} + ... & \quad (11) \\ + \sum_{n=1}^{Nb_{\in \Delta f}} A^{est}_n W(f + f^{est}_n) e^{i\phi^{est}_n} & \\ f = [f_m - 1, f_m, f_m + 1] & \quad (12) \end{aligned}$$

Here, A^{est}_n,f^{est}_n and ϕ^{est}_n are the estimate's parameters and f_m the frequency corresponding to a maximum of the periodogram. The second term is only used in the case of low frequency and ϕ_n is defined by :

$$\phi^{est}_n \ = \ \phi_{\lfloor f^{est}_n \rfloor} + (f^{est}_n - f_n)\,(\phi_{\lceil f^{est}_n \rceil} - \phi_{\lfloor f^{est}_n \rfloor}) \tag{13}$$

When the contributions are subtracted, we interpolate the value of the parameters on the peak of interest.

2.3. Synthesis, Noise extraction and SNR estimation

For the synthesis signal $\hat{X(t)}$, we use the parameters estimated before and we create a partial signal with the model. The total reconstruction of the signal is made by using the overlap and add method and by using the same window as that for the analysis. Then, we subtract the noise estimates from the original signal and compute the SyNR (Synthesis to Noise Ratio) given by :

$$SyNR \ = \ 10\, log_{10}(\frac{\sum_i \hat{X_i(t)}^2}{\sum_i\ (X_i(t) - \hat{X_i(t)})^2}) \tag{14}$$

3. APPLICATION TO ACOUSTIC INSTRUMENTS

3.1. Introduction

The audio signals are modeled as a sum of sinusoids with time varying parameters. However, in this model the nature of the instruments is completely ignored [5]. The sound of a wind instrument is composed of blow, for a violin the sound is generated by the friction of the bow on the string and more generally for the touch or string instrument the sound is due to striking. When we extract the sinusoidal part of the signal, we obtain a noise composed, in the best case, by the background noise and the instrumental noise. In this situation estimating the SNR can allow us to detect some musical characteristics. The noise part have received a great interested in the field of onset detection and tempo estimation [6] because it emphasize the percussif event (non sinusoidal).

3.2. Bass : Slap detection

The slap is a very common technique in bass playing. The strike consists of hitting the strings with the thumb, in the beginning of the fretboard, like a hammer. The resulting sound is almost completely percussive upon the strike, and afterwards the sinusoidal regime appears. Note that a note played by slap has a small duration compare to a note played with the finger. Fig. 1 shows the result of the SNR estimation on a bass sequence composed of two single notes. The first note is played with the finger (sweet) and the second is played by slap (percussive). As we expected, the SNR of the slap note is small compare to the note played with the finger.

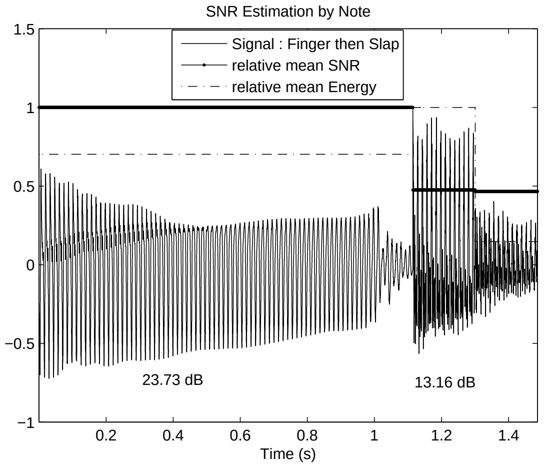

Fig. 1. Estimation of the SNR for bass sequence of two notes, the first is play with the finger and the second is play by slap.

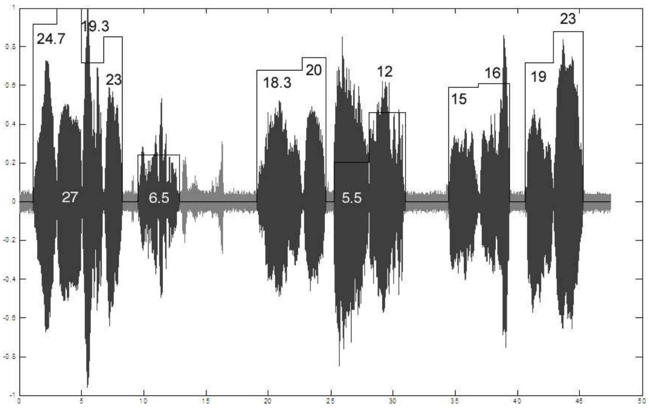

Fig. 2. Estimation of the SNR for a Violin piece played by a student.

3.3. Violin : The practice of the bow

When a violinist plays, he moves the bow upon the strings and the sound is generated by the friction of this movement. A well played note is constrained by at least four parameters:

- The speed of the displacement of the bow.
- The pressure exerted on the string.
- And the two orientations of the bow (on itself and on the string).

A good player exercises constant speed and pressure, keeps the bow completely parallel on the string and the displacement orthogonal to the string. When one or more of these constraints are not respected, the sound becomes more noisy. In the worst case, we only heard the displacement of the bow.

Fig. 2 show a succession of notes played by a student. Here we do not detail the attack detection method used for finding the note. The black part corresponds to the detected note, the line corresponds to the relative SNR and the SNR is given by this value. The SNR is a quantity which is independent of the volume and we can observe the difference in the estimation of each note, the ground truth was realise by 3 people who have classified the notes according to their qualities. We can thus qualify a good note and a bad note. Note that this study can be applicable in a pedagogic way and we do not take into account interpretation effects like hammering or pizzicato. So, here, a good note has an SNR larger than $18\ dB$ and a very bad note has a SNR smaller than $8\ dB$. In practice, all the thresholds have to be adjusted in accordance with the background noise.

3.4. Piano : Sustain pedal detection

In a piano, the sound generation mechanism works as follows: when the musician presses a key, a hammer strikes the string and this interaction triggers the note. When the key is released, a damper comes to stop the vibration of the strings and the note fades out. When the sustain pedal is pressed, all the dampers of the piano are kept raised; this allows the strings to keep vibrating after the key is released, and allows strings associated to other keys to vibrate. If several notes are played with the pedal, they will be mixed with a longer duration. A second effect has yet to be noticed. As a matter of fact, the two higher octaves of the piano do not have any damper, but the use of the pedal still has an influence on the sound. For this range of notes, the note does not last longer with or without the pedal, but a natural reverberation due to the resonance of the sound board appears and this sound leads to an additional floor noise.

Similar observations can be found in previous work. [7] proposes a polyphonic piano transcription system which detects and takes into account the use of the pedal. The detection of the pedal is based on an estimation of the noise floor. It is estimated as the mean value of the Discrete Fourier Transform (DFT) magnitude over the analysis frame, but only on frequency bins considered as "not active" in the frame (not associated with an actually played note - these frequencies are determined by a varying threshold). Another modelling of the sustain pedal can be found in [8]. Through the analysis of middle-range piano notes, played *legato* with and without the pedal, the authors point out three features that should be able to discriminate between notes played with and without the pedal, and be useful for piano synthesis: noise floor, decay time of the partials and amplitude beating.

Since the sustain pedal is generally used for simulation of a long note which is not easy to play, a way for estimating the pedal will be to track all the note and to decide if it's possible to play them without the pedal (like if we found more than ten notes at the same time). So the only difference between a long note played with or without the use of the pedal is the noise floor. As presented in [9] the analysis is performed after the

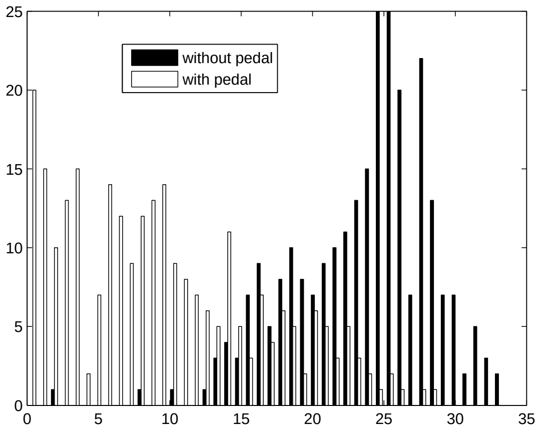

Fig. 3. Estimation of the total power of an AR of the Noise.

attack (after a duration of $250ms$) and on a duration of $250ms$ but the used database is the isolated notes of one of the piano of the RWC Database. The isolated notes represents all the notes of the piano with three kind of playing for each notes ($Piano$, $Mezzo$ and $Forte$).

Figure 3 show the result of the simulation, we have modeled the noise as an AR process of order one which is more robust to variability. In fact, due to the resonance the spectrum of the estimated noise is flatter with the pedal and quite less powerfull at the origine than the non-pedal case. It's for what it's less powerfull. The two classes are not well defined but it can be used as a another hint for detecting the pedal.

3.5. Guitar : Interpretation effects

There exists a lot of interpretation effects for the guitar, here we focus on three of them which have a similar caracteristic: the bend, the hammer and the slide. The bend is the action of deforming the string by pulling it up or down for increasing its length and changing the frequency. For the hammer, after a played note another finger come and strike another frette and become the new note. The slide, also called Glissando, is the action of sliding the finger to anoter frette. The common characteristics of this effects is that they only have one attack for several frequency variation, and except for the bend, the variation is at least a half tone. The bend can have a continuous frequency variation and it's limited to at most two tones. The hammer is limited by the length of the hand but can also be performed with open string so there's no restriction about the frequency variaton range. As for the hammer the slide can also have great variation.

After the detection of an attack the frequencies are tracked and if we found a variation of at least a quarter tone without a new attack the note is judged as one of the three cases explain above. As we want to detect the attack, which is the transient part of the note, the onset detection is done on the derivative of the energy of the signal. For emphasise the attack we use both the signal and the Half Wave Rectified (HWR) signal [10] and we keep the common onset.

The fundamental frequency is found, demodulated and low pass filtered, after we apply a Pisarenko method for tracking the frequency. We construct the correlation matrix for two consecutive sample as $R = A^H_{n,n+1}\, A_{n,n+1}$, where H denote the hermitian and A is the low pass filtered demodulated signal, we use a singular value decomposition of $R = UDU^T$ with T the transpose operator, we apply a Vandermond vector to the eigen vector : $[U1\, U2]^T[1\, z^-1] = U1 + U2\, z^-1 = 0$ we found the pole z and we estimate the instantaneous frequency. The test data set includes some monophonic and mono-instrumental recording, for the rest we define the effects by B for the bend, H for the hammer, S for the slide and P for the played case. The first set is composed by four successions of two notes played alternatively, the first note is always played (P), the second is played or reached by one of the above effects and the last one is played or is the opposite (design by off) of the effect. The data are played on different strings and notes and follow this scheme: PPP-PHH_{off}-PSS_{off}-PBB_{off} (6*12 notes). The sond data set includes other notes (2 notes by set), the first is played and the second is an effect, and represents 24*2 notes.

figure 4 show the results of the onsets detection (the detection function, the adaptive theshold and the onsets), the instantenaous frequency, and amplitude of the fundamental. The adaptive threshold takes as input the first onset (detected when the energy of the signal grows sensibly), if the value of the detection function is upper than ten percents of the last maximum it becomes the new one. For the frequency three line show the position of the previous and next half tone (in Midi). On our dataset, the system find all the played note (60/60), it interprets some effect as played note (8/60) and it detects some artefact note (4/60), which can be post filtered by considering the enveloppe variation. Note that the onset detector is not able to work in other case (Mono-phonic, Mono-Instrumental).

3.6. Guitar : Pizzicato

On bowed string instruments (violin, cello etc.) it's a method of playing which consist on plucking the strings with the fingers, rather than using the bow. The sound produce is very different, short and percussive rather than sustained. On the guitar, it's associated to a kind of plucking, which reach the sound of a pizzicato on a bowed string instrument. For the guitar, pizzicato is often called Palm mute and it's done differently. Palm mutes are executed by placing the side of the picking hand across all of the strings and very close to the bridge before or during the attack. This produces a muted sound. While rare in classical guitar technique, palm muting is a standard technique on an electric guitar, Plam mute is

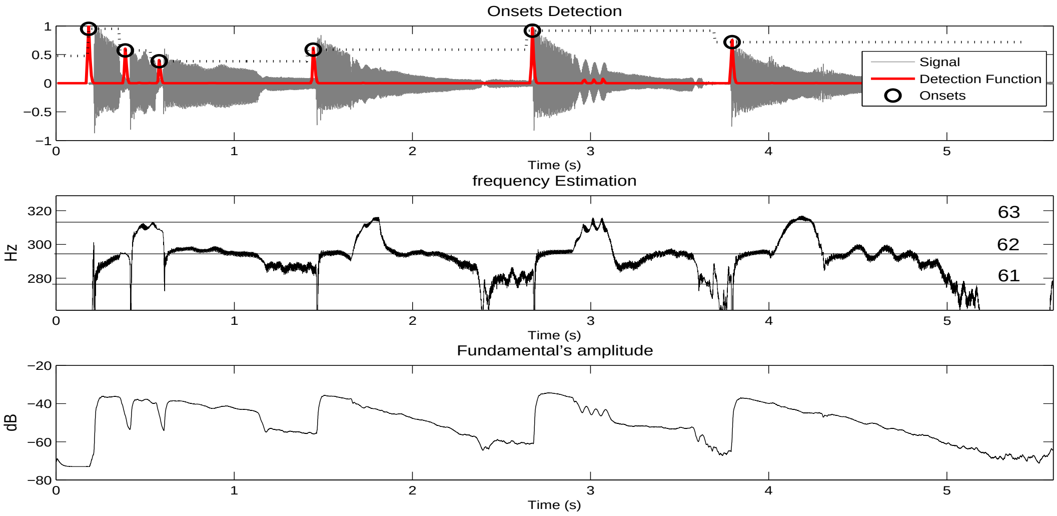

Fig. 4. Example of onsets detection.

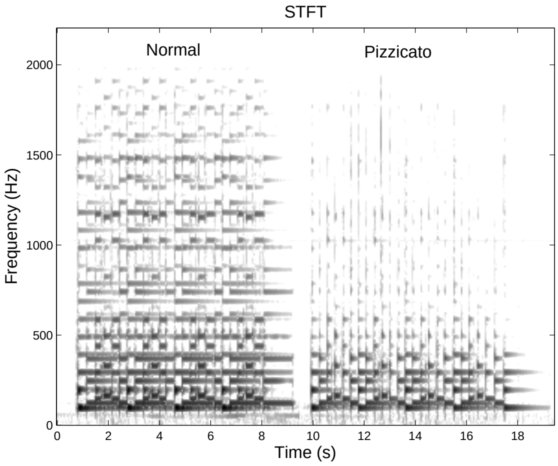

Fig. 5. Short Time Fourier Transform, The song is composed of Non-Pizzicato and Pizzicato notes.

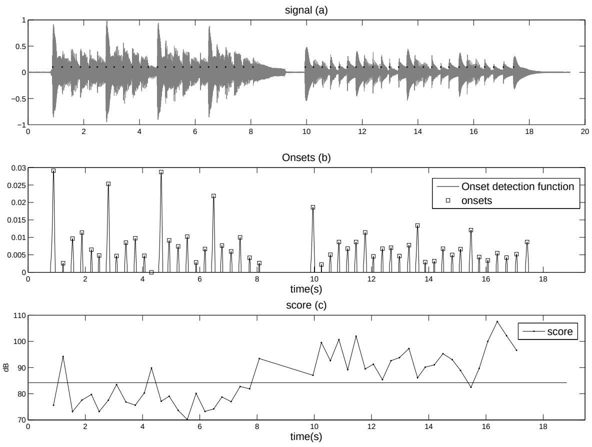

Fig. 6. Signal and onsets (a), onsets detection function and onsets (b), result of the detection (s).

more used when the musician play with a pick. For more details, the hand operate a low pass filtering (as for the damper pedal of the piano), figure 5 illustrates this effect. This results an attenuation of the power of the harmonics. Here we present some results for the guitar but the method can also be applied to instruments which can use a mute style like piano, bass guitar (rarely used) and obviously violin. Since previous remark, the detection of a Mute note is done as follow: First the notes are found using the same onset detector as previously, the pitch is determined by using the subharmonic-to-harmonic ratio method [11]. When the fundamental frequency and its harmonics are founded we subtract the contribution of the sub-harmonics to the harmonics. The sub-harmonic represents the valley between two succesive harmonics, the subtraction give a relative harmonic to noise ratio which emphasise the difference between the mute and the non-mute case. The score is given by the summation of the ratio of the fundamentale with the harmonic to noise ratio. Note that we use this criterium because it can be more robust in a polyphonic case, if we adapt the position of the sub-harmonics. We have trained a threshold using ten notes played in two cases, and we have applied the detector to a piece containing 49 notes, the system gave us a total score of 90 percents of good recognition. One mute note was interpretated as a normal note, and three for the opposite case.

4. CONCLUSION AND FUTURE WORK

In this paper we have presented some characteristics of some interpretation effects for differents intruments. We have focused on the importance of the noise for judge the quality of play and the goal was to correcting some playing defects. We have proposed some way for detecting some interpretation effects via some simple criterium. A lot of work is again needed, only one criterium is not enough robust to variability, and all the detectors have to be implemented in an automatic transcription system.

5. REFERENCES

[1] X. Serra, "Musical sound modeling with sinusoids plus noise," *Musical Signal Processing*, 1997.

[2] Marchand S. Keiler F., "Survey on extraction of sinusoids in stationary sounds," *DAFX*, 2002.

[3] Keiler R. Zolzer U. Althoff, R., "Extracting sinusoids from harmonic signals," *DAFX*, 1999.

[4] M. Abe and J. O. Smith, "Design criteria for the quadratically interpolated fft method," *ICASSP*, 2005.

[5] Rossing T.D. Fletcher, N.H., "The physics of musical instruments," *Springer Verlag*, 1991.

[6] David B. Richard G. Alonso M., Badeau R., "Musical tempo estimation using noise subspace projection," *WASPAA*, 2003.

[7] Jurado A Tardon L.J Barbancho I, Barbancho A.M, "Transcription of piano recordings," *Applied Acoustics*, 2004.

[8] Rauhala J Lehtonen H.M, Penttinen H and Valimaki V, "Analysis and modeling of piano sustain-pedal effects," *JASA*, 2007.

[9] Slock D. David B. Badeau R. Schutz A., Bertin N., "Piano forte pedal analysis and detection," *AES124*, 2008.

[10] A. Klapuri, "A perceptually motivated multiple-f0 estimation method," *IEEE Workshop on Applications od Signal Processing to Audio and Acoustics*, 2005.

[11] X. Sun, "A pitch determination algorithm based on subharmonic-to-harmonic ratio," *the 6th International Conference of Spoken Language Processing*, 2000.