MULTIPLE F0 ESTIMATION IN POLYPHONIC MUSIC (MIREX 2007)

Chuan Cao, Ming Li, Jian Liu and Yonghong Yan
Thinkit Speech Lab., Institute of Acoustics,
Chinese Academy of Sciences,
{ccao,mli,jliu,yyan}@hccl.ioa.ac.cn

ABSTRACT

This paper describes our method for the MIREX 2007 "Multiple Fundamental Frequency Estimation & Tracking" task in which the goal is to identify the active F0s in each time frame and to track notes and timbres continuously in a complex music signal. The introduced method is based on subharmonic-summation pitch estimation method and a spectrum cancelation algorithm. Our algorithm concentrates on the frame-level transcription, so the submission is only for "Task 1".

1 INTRODUCTION

The MIREX (Music Information Retrieval Evaluation eXchange) framework provides a common platform to compare and evaluate a vast variety of MIR systems. And the MIREX 2007 "Multiple Fundamental Frequency Estimation & Tracking" task aims to evaluate state-of-the-art multiple F0 estimation and tracking algorithms.

Musical signals are natural candidates for the problem of multiple-F0 estimation since multiple instruments are played simultaneously often and chords are very familiar in polyphonic music. However, it is not easy to estimate the instruments' fundamental frequencies in the existence of other simultaneous sounds.

The proposed multi-F0 estimation system is based on a spectral cancelation mechanism with the subharmonic-summation(SHS) pitch estimation method. For simplicity, we extract five F0 candidates for every frame, followed by a pitch verification algorithm. F0 candidates that are verified non-valid will be removed from the candidates pool.

2 SYSTEM DESCRIPTION

2.1 Pitch Estimation Method

Subharmonic-summation algorithm used here is generally based on Dik J.Hermes' pitch-determination algorithm [2], which can be basic concluded in the formula below:

$$H(f) = \sum_{n=1}^{N} h_n P(nf) \tag{1}$$

where, $H(f)$ is the subharmonic-summation spectrum of the hypothetic pitch value f, $P(*)$ is the STFT power spectrum and h_n the compression factor (usually $h_n = h^{n-1}$).

Here in our method, we introduce a spectral normalized factor in the form described below:

$$H(f_0) = \sum_{n=1}^{N} h^{n-1} Q_{f_0}(nf_0) \tag{2}$$

where,

$$Q_{f_0}(f) = P(f) \Big/ \Big(\frac{1}{2\rho f_0} \int_{f-\rho f_0}^{f+\rho f_0} P(w)dw \Big) \tag{3}$$

where, ρ is the spectral normalized width factor which can reduce octave errors effectively, compared to the original one in (1). This modified version shows good performance on a test set of about 300s speech data with pitch reference. For convenience, we note SHS shortly for the spectral normalized SHS in the following text.

2.2 Cancelation Mechanism

We use a spectrum cancelation mechanism to remove the predominant harmonic structure and then iteratively estimate the fundamental frequencies of the residual sound. The algorithm can be concluded as procedures:

1. Calculate the SHS spectrum from the STFT spectrum.
2. Choose the F_0 with the following constraints and then add it into the candidates pool:
$$F_0 = \arg\max_f SHS(f) \tag{4}$$
3. Terminate the circulation if enough F0 candidates have been extracted, or else go to step4.
4. Cancelation the STFT spectrum with: $STFT(f) = w(f) * STFT(f)$, for $f = nF_0$ and n=1,2,3..., and then go to step1.

nF_0 in the step4 refers to the n^{th} harmonic frequency of the hypothetic fundamental frequency F_0, and $w(f)$ is a weight function which is proportional to the distance between the center frequency of the STFT bin and the cancelation frequency nF_0. In reality, for the spectral leak of STFT analyze, we not only change the value of

Participant	Recall	Precision	E_{tot}	E_{subs}	E_{miss}	E_{fa}	Overall Accuracy
Ryynänen & Klapuri(1)	70.9%	69.0%	0.474	0.158	0.133	0.183	60.5%
Yeh, C.	65.5%	76.5%	0.460	0.108	0.238	0.115	58.9%
Zhou & Reiss	66.1%	71.0%	0.498	0.141	0.197	0.16	58.2%
Pertusa & Iñesta	60.8%	82.7%	0.445	0.094	0.298	0.053	58.0%
Vincent, Bertin & Badeau(2)	51.3%	65.9%	0.594	0.171	0.317	0.107	46.6%
Cao, Li, Liu & Yan (1)	67.1%	56.7%	0.685	0.20	0.128	0.356	51.0%
Raczyński, Ono & Sagayama	59.5%	61.4%	0.670	18.5	0.219	0.265	48.4%
Vincent, Bertin & Badeau (1)	51.3%	65.9%	0.594	17.1	0.317	0.107	46.6%
Poliner & Ellis (1)	50.5%	73.4%	0.639	0.12	0.375	0.144	44.4%
Leveau, P.	41.7%	68.9%	0.639	0.151	0.432	0.055	39.4%
Cao, Li, Liu & Yan (2)	76.7%	35.9%	1.678	0.232	0.001	1.445	35.9%
Egashira, Kameoka & Sagayama (2)	54.6%	34.8%	1.188	0.401	0.052	0.734	33.6%
Egashira, Kameoka & Sagayama (1)	61.8%	33.5%	1.427	0.339	0.046	1.042	32.7%
Cont, A. (2)	43.1%	37.3%	0.99	0.348	0.221	0.421	31.1%
Cont, A. (1)	53.0%	29.8%	1.444	0.332	0.138	0.974	27.7%
Emiya, Badeau & David (1)	15.7%	53.0%	0.957	0.070	0.767	0.120	14.5%

Table 1. Results of all participants of task1.

$STFT(nF_0)$, but also that of the $STFT(f)$ with the f near nF_0.

As referred above, we iteratively extract five F0 candidates for every frame, and add them into the candidates pool. Since the candidates simply stem from the *SHS spectrum* maxima, a F0 verification procedure is needed.

2.3 Verification of F0 Candidates

We use a spectral-based method here for F0 verification. Generally speaking, for a specific F0 candidate, we check the existence of the spectral peak for all of its harmonics and then assign a saliency score for this F0 candidate, which can be concluded as follows:

$$S_{F_0} = \sum_n W_n, \tag{5}$$

$$W_n = \begin{cases} 0.9^{n-1}, & \text{peak exists around } nF_0 \\ 0, & \text{else} \end{cases} \tag{6}$$

Then candidates with saliency small than θ_s is removed from the pool and the final output multi-pitch stream is formed with the surviving F0 values at every frame.

3 IMPLEMENTATION

The algorithm is implemented in C++ and is for Windows platform. The execution time on a P4 3.2G CPU with 1GB RAM is about 4 times of the real-time without any particular optimizations.

4 EVALUATION RESULTS

The task1 evaluation results for all participants are listed in the table1. Our system 'Cao, Li, Liu & Yan (1)' with performance of 51.0% overall accuracy and 'Cao, Li, Liu & Yan (2)' with overall accuracy of 35.9%.

5 CONCLUSION AND FUTURE WORK

As seen in the table1, our system outperformed other systems (except for 'Ryynänen & Klapuri(1)'s system) for the pitch recall rate. But they are not so well on the precision criterion. For system 'Cao, Li, Liu & Yan (2)', the poor precision performance was not suprising some what, for the reason that we did not do the verification procedure for this submission and reported five pitches for every frame. So a lot of output pitches are not valid. However, even the system 'Cao, Li, Liu & Yan (1)' was with the verification procedure, it also performed poor for precision. Maybe our verification procedure is not efficient enough and our saliency threshold θ_s needs to be tuned more carefully and on more test files. So in the future, we are expecting to find more efficient ways to verify the validation of a specific hypothetic F0 candidate.

6 ACKNOWLEDGEMENTS

Many thanks to the IMIRSEL team at the University of Illinois at Urbana-Champaign for organizing and running the MIREX evaluations.

7 REFERENCES

[1] A.Klapuri. "Multiple fundamental frequency estimation by summing harmonic amplitudes," In *Proc.7th International Conference on Music Information Retrieval*, 2006.

[2] Dik Hermes. "Measurement of pitch by subharmonic summation," *Journal of Acoustic of Society of America*, vol.83, pp.257-264,1988.