F0 ESTIMATION METHOD FOR SINGING VOICE IN POLYPHONIC AUDIO SIGNAL BASED ON STATISTICAL VOCAL MODEL AND VITERBI SEARCH

Hiromasa Fujihara,[†] *Tetsuro Kitahara,*[†] *Masataka Goto,*[‡]
Kazunori Komatani,[†] *Tetsuya Ogata,*[†] *and Hiroshi G. Okuno*[†]

[†]Dept. of Intelligence Science and Technology
Graduate School of Informatics, Kyoto University
Sakyo-ku, Kyoto 606-8501, Japan
{fujihara,kitahara,komatani,ogata,okuno}@kuis.kyoto-u.ac.jp

[‡]National Institute of Advanced Industrial
Science and Technology (AIST)
Tsukuba, Ibaraki 305-8568, Japan
m.goto@aist.go.jp

ABSTRACT

This paper describes a method for estimating F0s of vocal from polyphonic audio signals. Because melody is sung by a singer in many musical pieces, the estimation of F0s of the vocal part is useful for many applications. Based on existing multiple-F0 estimation method, we evaluate the vocal probabilities of the harmonic structure of each F0 candidate. In order to calculate the vocal probabilities of the harmonic structure, we extract and resynthesize the harmonic structure by using a sinusoidal model and extract feature vectors. Then, we evaluate the vocal probability by using vocal and non-vocal Gaussian mixture models (GMMs). Finally, we track F0 trajectories using these probabilities based on Viterbi search. Experimental results show that our method improves estimation accuracy from 78.1% to 84.3%, which is 28.3% reduction of misestimation.

1. INTRODUCTION

Singing voice plays an important role in many musical genres, especially in popular music. The estimation of the fundamental frequency (F0) of the vocal part, therefore, is an important issue. Estimated F0s of vocal are useful in various applications, such as automatic transcription, automatic generation of Karaoke track and music information retrieval (e.g., searching for a song by singing a melody).

Several study have been made on melody extraction from polyphonic audio signals [1, 2, 3, 4]. Their methods usually consisted of two steps. First, they estimated F0 candidates from input audio signals using multiple-pitch estimation method. Then, among these F0 candidates, they constructed a melody trajectory by using some clues such as predominance, timber, meter and F0's continuity. However, the F0s of other predominant instruments performed concurrently with vocals were often detected because they did not assume what a sound source was [1].

In this paper, we focus on the F0 estimation of vocal from polyphonic audio signals. Polyphonic audio signals contain frequency components corresponding to various F0s coming from the instruments. Without vocal/non-vocal discrimination focusing on a frequency component corresponding to a particular F0, it is difficult to estimate the F0s of vocal. In previous vocal/non-vocal discrimination methods [5, 6, 7], since feature vectors were extracted directly from polyphonic audio signals, they did not identify a sound source of a particular F0. In order to overcome this technical issue, we use accompaniment sound reduction we developed [8], which enable us to accurate F0 estimation focusing on vocal.

This research was supported in part by the Ministry of Education, Culture, Sports, Science and Technology (MEXT), Grant-in-Aid for Scientific Research (A), No.15200015, and Informatics Research Center for Development of Knowledge Society Infrastructure (COE program of MEXT, Japan). We thank everyone who has contributed to building and distributing the RWC Music Database [11]. We also thank Hirokazu Kameoka, Kazuyoshi Yoshii and Takuya Yoshioka for their valuable advice.

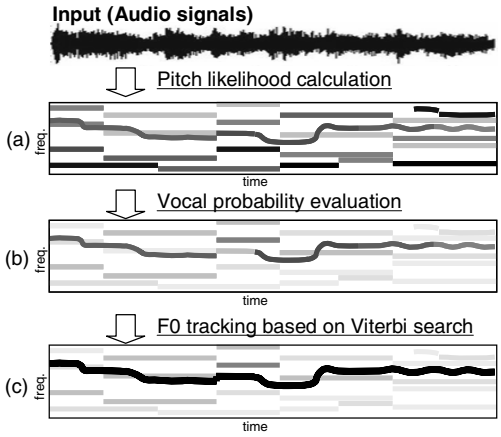

Fig. 1. Overview of our method.

There are two key ideas behind our method. First is to evaluate the probability that harmonic structure of each F0 candidate is vocal, which we call *vocal probability*, by using vocal and non-vocal GMMs. We enable it by using accompaniment sound reduction. First, we extract and resynthesize the harmonic structure of each F0 candidate by using a sinusoidal model and extract feature vector from the resynthesized audio signal. Then, we evaluate the vocal probability according to the Bayes rule. Second is to track vocal's F0 trajectory based on Viterbi search considering these vocal probabilities. We enable it by modeling and formulating the stochastic dependency among a sequence of F0s, spectra and sound sources using a graphical model.

2. OUR METHOD FOR ESTIMATING F0 OF VOCAL

Figure 1 shows an overview of our method. It consists of three parts: pitch likelihood calculation, vocal probability calculation and F0 tracking based on Viterbi search. We define *pitch likelihood* as likelihood that an F0 is the most predominant F0 in a spectrum.

First, we calculate pitch likelihood (Figure 1 (a)) using PreFEst[1] developed by Goto. Given a spectrum, it calculates a predominance of every possible F0 at each time. We consider the predominance as pitch likelihood. Then, we evaluate vocal probability for each F0 (Figure 1 (b)). Finally, considering the vocal probabilities and the

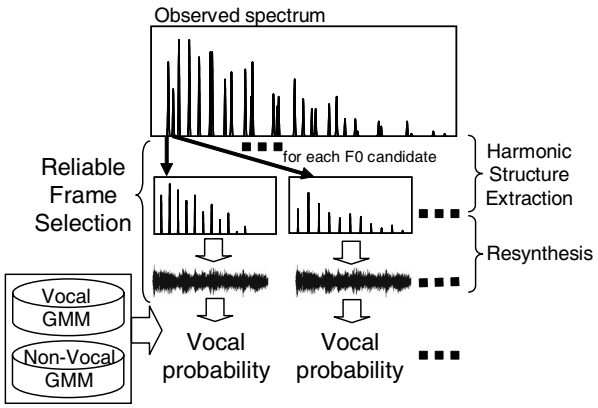

Fig. 2. Vocal probability calculation.

continuity of F0s, we estimate the time series of F0s that gives the highest probability (Figure 1 (c)).

Hereafter, f represents the log-scale frequency denoted in units of cents (a musical-interval measurement).

2.1. Pitch Likelihood

We use Goto's PreFEst [1] for calculating pitch likelihood. Though PreFEst consists of three steps, front-end, core and back-end, we do not use the PreFEst-back-end method here because it selects most predominant pitch trajectory.

We will describe a summary of PreFEst-core below. Given the power spectrum $\psi(f)$, in order to enable the application of statistical methods, we represent each of the frequency components as a probability density function (PDF), called an observed PDF: $p_\psi(f) = \psi(f)/\int_{-\infty}^{\infty}\psi(f)df$. Then, we consider each observed PDF to have been generated from a weighted-mixture model of the tone models of all the possible F0s, which is represented as follows: $p(f|\theta) = \int_{\mathrm{Fl}}^{\mathrm{Fh}} w(F)p(f|F)dF$, $\theta = \{w(f)|\mathrm{Fl} \leq f \leq \mathrm{Fh}\}$, where $p(f|F)$ is the PDF of the tone model for each F0, and Fh and Fl denote lower and upper limits of the possible (allowable) F0 range, and $w(f)$ is the weight of a tone model that satisfies $\int_{\mathrm{Fh}_i}^{\mathrm{Fl}_i} w(f)df = 1$. Tone model represents a typical harmonic structure and indicates where the harmonics of the F0 tend to occur. Then, we estimate $w(f)$ using EM algorithm and regard it as the F0's PDF. We consider F0's PDF estimated by PreFEst-core as pitch likelihood function, that is $p(\psi|f) = w(f)$. Considering the computational cost, we select 10 candidates that have the highest likelihoods.

2.2. Vocal/Non-vocal Probability

We describe the method that enables us to calculate feature vectors corresponding to each F0 candidate and its vocal and non-vocal probabilities. Figure 2 shows an overview of this method.

2.2.1. Accompaniment Sound Reduction

To calculate the feature vector of each of the F0 candidate, we use accompaniment sound reduction, which we proposed in [8]. By using this method, we can obtain the audio signal that corresponds only to the F0. This method consists of the following two steps.

1. Harmonic Structure Extraction
   We extract the power and the phase of fundamental frequency component and harmonic components. The extracted power, A_l, and frequency, F_l, of the l-th overtone ($l = 1, \ldots, 20$) can be represented as

$$F_l = \underset{f}{\mathrm{argmax}}\,|\psi(f)|G(f; l\overline{f}, 20), \quad (1)$$

$$A_l = |\psi(F_l)|, \quad (2)$$

   where ψ denotes the spectrum, $\overline{f}$ denotes the target F0, and $G(x; m, \sigma)$ represents the Gaussian distribution.

2. Resynthesis
   We resynthesize the audio signal of the melody from the extracted harmonic structure by using a sinusoidal model [9]. Resynthesized audio signals $s(t)$ are expressed as

$$s(t) = \sum_{l=1}^{L} A_l \cos(\omega_l t). \quad (3)$$

2.2.2. Feature Extraction

From the resynthesized audio signals, we calculate feature vectors consisting of two features.

- LPC-derived mel cepstral coefficients (LPMCCs)
  We use LPMCCs as spectral feature for vocal/non-vocal discrimination because we have reported that, in the context of singer identification, LPMCCs express vocal's characteristics better than mel-frequency cepstral coefficients (MFCCs), which are widely used for music modeling [8].

- ΔF0s
  We use ΔF0s [10], which represent the dynamics of F0's trajectory, because singing voice tends to have temporal variation of F0s in consequence of vibrato and, therefore, ΔF0s are expected to be good cues for vocal/non-vocal discrimination.

2.2.3. Probability Calculation

We introduce two Gaussian mixture models (GMMs): a vocal GMM, s_V, and a non-vocal GMM, s_N. The parameters of the vocal GMM, θ_V, is trained on feature vectors extracted from vocal sections, and the parameters of the non-vocal GMM, θ_N, is trained on those extracted from interlude sections.

When $\boldsymbol{x}(\psi_t, f)$ is the feature vector extracted from a spectrum, ψ_t, with a fundamental frequency f at time t, the likelihoods of the vocal and non-vocal model are represented as

$$p(\psi_t, f|s_\mathrm{V}) = \mathcal{N}_{\mathrm{GMM}}(\boldsymbol{x}(\psi_t, f); \theta_\mathrm{V}), \quad (4)$$

$$p(\psi_t, f|s_\mathrm{N}) = \mathcal{N}_{\mathrm{GMM}}(\boldsymbol{x}(\psi_t, f); \theta_\mathrm{N}), \quad (5)$$

where $\mathcal{N}_{\mathrm{GMM}}(\boldsymbol{x}; \theta)$ denotes the probability density function of the GMM with parameter θ.

The vocal and non-vocal probabilities, according to the Bayes rule, can be represented as

$$p(s_\mathrm{V}|\psi_t, f) = \frac{p(\psi_t, f|s_\mathrm{V})p(s_\mathrm{V})}{p(\psi_t, f|s_\mathrm{V})p(s_\mathrm{V}) + p(\psi_t, f|s_\mathrm{N})p(s_\mathrm{N})}, \quad (6)$$

$$p(s_\mathrm{N}|\psi_t, f) = \frac{p(\psi_t, f|s_\mathrm{N})p(s_\mathrm{N})}{p(\psi_t, f|s_\mathrm{V})p(s_\mathrm{V}) + p(\psi_t, f|s_\mathrm{N})p(s_\mathrm{N})}, \quad (7)$$

where $p(s_V)$ and $p(s_N)$ denote a priori probabilities of vocal and non-vocal. We used 64-mixture GMM and set $p(s_V) = p(s_N)$.

2.3. F0 Tracking Based on Viterbi Search

We formulate the F0 tracking method based on Viterbi search.

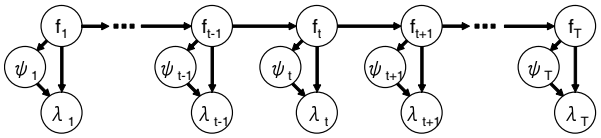

Fig. 3. Stochastic dependency among F, Ψ and Λ.

2.3.1. Formulation

At each time step, t $(t = 1, \cdots, T)$, we define an F0, spectrum and sound source as f_t, ψ_t and λ_t, respectively. We also define $F = \{f_t | t = 1, \cdots, T\}$, $\Psi = \{\psi_t | t = 1, \cdots, T\}$ and $\Lambda = \{\lambda_t | t = 1, \cdots, T\}$. As a sound source, we consider a singing voice and the other sound, that is $\lambda_t \in \{s_{\mathrm{V}}, s_{\mathrm{N}}\}$.

We assume that stochastic dependency among F_t, Ψ_t, and Λ_t can be expressed as show in Fig. 3. According to Fig. 3, we introduce the following conditional probability distributions: vocal/non-vocal probability $p(\lambda_i | f_i, \psi_i)$, pitch likelihood $p(\psi_i | f_i)$ and transition probability $p(f_i | f_{i-1})$. Pitch likelihood and sound source probability were explained in Sec. 2.1 and Sec. 2.2. We define transition probability, $p(f_i | f_{i-1})$, as

$$p(f_i | f_{i-1}) = G(f_i; f_{i-1}, W_V), \tag{8}$$

where $G(x; m, \sigma)$ represents the Gaussian distribution, and W_V denotes the changeability of F0. We set W_V to 100 cent.

The purpose is, after observing the time series of spectra, $O = \{o_t | t = 1, \cdots, T\}$, and observing that sound source is singing voice, s_V, at every time step, to estimate a sequence of F0s $\hat{F}$ that maximizes the following equation.

$$\hat{F} = \operatorname*{argmax}_{F} \log p(F | \Psi = O, \Lambda = (\text{always } s_V)) \tag{9}$$

$$= \operatorname*{argmax}_{F_T} \left\{ \sum_{t=1}^{T} \log p(\lambda_t = s_V | f_t, \psi_t = o_t) + \sum_{t=1}^{T} \log p(\psi_t = o_t | f_t) + \sum_{t=1}^{T} \log p(f_t | f_{t-1}) \right\} \tag{10}$$

In practice, introduction of connection weights among vocal/non-vocal probability, pitch likelihood and transition probability are effective such as the following equation.

$$\hat{F} = \operatorname*{argmax}_{F_T} \left\{ \alpha \sum_{t=1}^{T} \log p(\lambda_t = s_{\mathrm{V}} | f_t, \psi_t = o_t) + \beta \sum_{t=1}^{T} \log p(\psi_t = o_t | f_t) + \sum_{t=1}^{T} \log p(f_t | f_{t-1}) \right\} \tag{11}$$

We use Eq. (11) instead of Eq. (10), setting $\alpha = 0.2$ and $\beta = 0.8$.

2.3.2. Viterbi Search

Since directly computing Eq. (11) is difficult, we compute it recursively by using the following equations. We introduce an back pointer, $B(t, f)$, and an accumulated probability, $A(t, f)$.

(1) Initialization

$$\forall f \;\; A(1, f) = \alpha \log p(\lambda_1 | f, \psi_1) + \beta \log p(\psi_1 | f) \tag{12}$$

(2) Recursive calculation $(t = 2, \cdots, T)$

$$A(t, f) = \max_{f'} \{ A(t-1, f') + \alpha \log p(\lambda_t | f, \psi_t) + \beta \log p(\psi_t | f) + \log p(f | f') \} \tag{13}$$

$$B(t, f) = \operatorname*{argmax}_{f'} \{ A(t-1, f') + \alpha \log p(\lambda_t | f, \psi_t) + \beta \log p(\psi_t | f) + \log p(f | f') \} \tag{14}$$

Table 1. Training data for Vocal/Non-Vocal model.

Name	Gender	Piece Number
Shingo Katsuta	M	027
Yoshinori Hatae	M	037
Masaki Kuehara	M	032, 078
Hiroshi Sekiya	M	049, 051
Katsuyuki Ozawa	M	015, 041
Masashi Hashimoto	M	056, 057
Satoshi Kumasaka	M	047
Konbu	F	013
Eri Ichikawa	F	020
Tomoko Nitta	F	026
Kaburagi Akiko	F	055
Yuzu Iijima	F	060
Reiko Sato	F	063
Donna Burke	F	081, 091, 093, 097

We can obtain the sequence of F0s, $\hat{F} = \{\hat{f}_1, \cdots, \hat{f}_T\}$, that gives the highest probability by tracking back the back pointer.

$$\hat{f}_T = \operatorname*{argmax}_{f} A(T, f) \tag{15}$$

$$\hat{f}_t = B(\hat{f}_{t+1}) \quad (t = T-1, \cdots, 1) \tag{16}$$

3. EXPERIMENTS

3.1. Construction of Vocal/Non-vocal Models

We describe the construction of vocal/non-vocal models. As the training data, we used 21 songs of the 14 singers listed in Table 1, which were taken from the "RWC Music Database: Popular" [11]. First, we computed boundaries between vocal and non-vocal sections by comparing polyphonic data and vocal-only data. The vocal GMM was trained on feature vectors that were extracted from the vocal section of polyphonic data using the F0s estimated from vocal-only data. The non-vocal GMM was trained on feature vectors that were extracted from the non-vocal section of polyphonic data using the F0s estimated from polyphonic data using PreFEst[1].

3.2. Conditions

We evaluated our method on 10 musical pieces taken from the "RWC Music Database: Popular" [11] (Table 2). The singers of the musical pieces used for training are not included in the 10 musical pieces used for evaluation. To evaluate the effectiveness of our method, we conducted experiments under the following six conditions:

(i) Max density Select the F0s maximizing pitch density, without considering vocal probability nor F0's continuity.

(ii) PreFEst-back-end (baseline) Use the PreFEst-back-end [1], which tracks peak trajectories considering F0s' continuity by introducing a multiple-agent architecture.

(iii) MFCC Use our method with MFCCs as feature vectors.

(iv) MFCC+ΔF0 Use our method with MFCCs and ΔF0 as feature vectors.

(v) LPMCC Use our method with LPMCCs as feature vectors.

(vi) LPMCC+ΔF0 (Proposed) Use our method with LPMCCs and ΔF0 as feature vectors.

Estimation accuracies were evaluated by comparing the estimated F0s with the correct F0s, obtained by estimating the F0s from vocal-only data. The evaluation was made during periods when a vocal was present. We use two estimation accuracy indicators: pitch accuracy and chroma accuracy. The pitch accuracy is the probability of

Table 2. Experimental results of F0 estimation for fifteen musical pieces in RWC-MDB-P-2001.

Piece number	Singer's gender	Accuracy indicator	Method					
			Max density	PreFEst-back-end	MFCC	MFCC+ΔF0	LPMCC	LPMCC+ΔF0
No.007	Female	Pitch acc.	76.5%	87.3%	92.2%	92.9%	92.7%	93.1%
		Chroma acc.	78.5%	88.0%	92.5%	93.2%	93.0%	93.4%
No.012	Male	Pitch acc.	74.0%	78.4%	80.3%	81.8%	80.3%	82.3%
		Chroma acc.	76.3%	80.4%	82.4%	83.9%	82.4%	84.5%
No.019	Male	Pitch acc.	59.8%	62.5%	69.1%	70.0%	68.3%	70.2%
		Chroma acc.	63.9%	66.1%	72.8%	73.8%	72.4%	74.2%
No.021	Female	Pitch acc.	80.1%	83.1%	85.7%	86.5%	86.6%	87.5%
		Chroma acc.	80.6%	83.1%	85.8%	86.5%	86.7%	87.5%
No.039	Male	Pitch acc.	78.0%	81.1%	84.4%	86.3%	85.3%	86.2%
		Chroma acc.	81.3%	83.8%	87.2%	89.2%	88.1%	89.1%
No.065	Female	Pitch acc.	73.8%	79.3%	85.6%	88.9%	87.0%	90.2%
		Chroma acc.	77.0%	80.3%	87.1%	89.0%	87.4%	90.3%
No.075	Female	Pitch acc.	79.9%	83.7%	87.2%	91.0%	86.1%	90.4%
		Chroma acc.	80.9%	84.2%	87.5%	91.3%	86.6%	90.7%
No.083	Male	Pitch acc.	72.5%	72.0%	73.2%	76.4%	74.1%	76.9%
		Chroma acc.	74.0%	72.6%	74.4%	77.7%	75.4%	78.3%
No.088	Male	Pitch acc.	67.5%	76.4%	77.8%	84.9%	84.7%	84.6%
		Chroma acc.	70.3%	77.4%	79.5%	85.8%	85.7%	85.5%
No.092	Female	Pitch acc.	76.6%	77.2%	80.9%	81.3%	80.4%	81.3%
		Chroma acc.	77.6%	77.3%	81.0%	81.4%	80.5%	81.3%
Average		Pitch acc.	73.9%	78.1%	81.6%	84.0%	82.6%	84.3%
		Chroma acc.	76.0%	79.3%	83.0%	85.2%	83.8%	85.5%

correct pitch value. The chroma accuracy is the probability that the chroma (i.e. the note name) is correct. In other words, octave errors are ignored.

3.3. Results and Discussions

Experimental results, listed in Table 2, show that our method improves the estimation accuracy from 78.1% to 84.3%. By this fact, we can confirm the effectiveness of our method. We can also find that, introduction of ΔF0 improve the accuracy from 82.6% to 84.3%. Though ΔF0 has not been used in conventional vocal/non-vocal discrimination method, we confirmed that this feature is a good cue for vocal/non-vocal discrimination. When we compare MFCCs and LPMCCs, the accuracy for LPMCCs was 0.3% higher than that of MFCCs.

4. CONCLUSION

We have described a method that estimates the F0s of vocal part in polyphonic audio signals. The basic ideas of our method are to calculate the vocal probabilities of each F0 candidate and to track the F0 trajectory using these probability based Viterbi search. Though conventional vocal/non-vocal discrimination method cannot treat the harmonic structure of a particular F0, we made it possible by using our accompaniment sound reduction. Experimental results showed that our system accurately estimated the F0s of vocal and improve estimation accuracy.

In the future, we plan to extend our method to the song in which multiple singers sing together. We also plan to work on the detection of vocal region (i.e. where the singer is singing in a musical piece).

5. REFERENCES

[1] M. Goto, "A real-time music-scene-description system: predominant-F0 estimation for detecting melody and bass lines in real-world audio signals," *Speech Communication*, vol. 43, no. 4, pp. 311–329, 2004.

[2] M. Marolt, "Gaussian mixture models for extraction of melodic lines from audio recordings," in *Proc. ISMIR*, pp.80–83, 2004

[3] J. Eggink and G. J. Brown, "Extracting melody lines from complex audio," in *Proc. ISMIR*, pp.84–91, 2004.

[4] M. P. Ryynanen and A. Klapuri, "Note event modeling for audio melody extraction," in *Online Proc. MIREX2005*, 2005.

[5] A. L. Berenzweig and D. P. W. Ellis, "Locating singing voice segments within music signals," in *Proc. WASPAA*, 2001.

[6] W. Tsai and H. Wang, "Automatic detection and tracking of target singer in multi-singer music recordings," in *Proc. ICASSP*, pp. 221–224, 2004.

[7] T. L. Nwe, "Automatic detection of vocal segments in popular songs," in *Proc. ISMIR*, pp. 138–145, 2004.

[8] H. Fujihara, T. Kitahara, M. Goto, K. Komatani, T. Ogata, and H. G. Okuno, "Singer identification based on accompaniment sound reduction and reliable frame selection," in *Proc. ISMIR*, pp. 329–336, 2005.

[9] J. A. Moorer, "Signal processing aspects of computer music: A survey," *Proceedings of the IEEE*, vol. 65, no. 8, pp. 1108–1137, 1977.

[10] Y. Ohishi, M. Goto, K. Itou, and K. Takeda, "Discrimination between singing and speaking voices," in *Proc. Eurospeech*, pp. 1141–1144, 2005.

[11] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, "RWC Music Database: Popular, classical, and jazz music databases," in *Proc. ISMIR*, pp. 287–288, 2002.