# CONCURRENT ESTIMATION OF SINGING VOICE F0 AND PHONEMES BY USING SPECTRAL ENVELOPES ESTIMATED FROM POLYPHONIC MUSIC

*Hiromasa Fujihara and Masataka Goto*

National Institute of Advanced Industrial Science and Technology (AIST)

**ABSTRACT**

The scarcity of available multi-track recordings constitutes a severe constraint on the training of probabilistic models for voice extraction from polyphonic music. We propose a novel training method to estimate a spectral envelope of a singing voice that makes it possible to train the models from a polyphonic music without segregating a singing voice. We implement this method as an extension to the existing W-PST method, which concurrently estimates singing voice fundamental frequency (F0) and phoneme from polyphonic music. The novel training method is based on random sampling from probabilistic distributions. We conducted experiments on concurrent F0 and phoneme estimation and confirm the effectiveness of our method.

***Index Terms*—** Singing voice, Phoneme recognition, F0 estimation, Spectral envelope estimation.

## 1. INTRODUCTION

We aim to develop a computer that has the ability to distinguish singing voices in the same manner that humans do. When a human hears singing voices that are mixed with other sounds, he or she has an innate ability to distinguish singing voices from such mixed sounds. The current ability of computers to recognize the real world auditory scene is still inadequate when compared to human ability. In this paper, we focus on the fundamental frequency (F0) and phoneme (utterance content) as important elements of the singing voice. Computational recognition of F0 and phonemes is important from an industrial standpoint, since the fruits of this research are applicable in many important areas, including metadata description for music content, content-based music information retrieval, and sophisticated music playback interfaces.

We previously proposed a method for estimating phoneme from singing voice polyphonic music, called W-PST (Weighted-composition of Probabilistic Spectral Template) method [1]. This method stochastically models a mixture of a singing voice and other instrumental sounds without segregating the singing voice. It can also estimate a reliable spectral envelope by estimating it from the harmonic structure of many voices with various F0s. This method can be considered a new framework for recognizing a singing voice in polyphonic music because it is designed to concurrently recognize not only a phoneme but also other elements of a singing voice, including a singer's name and gender.

However, this method had a major technical issue; a monophonic singing voice, which is not always available, is required to train the models. Therefore, the scope of application of the method was limited. Moreover, although we proposed a framework that could concurrently estimate phonemes and F0, we evaluated only phoneme estimation tasks where the correct F0 was given in advance.

This work was supported by CrestMuse, CREST, JST.

We propose a training method that—in contrast to existing approaches—requires no separate singing voice recording, but instead trains a probabilistic model directly from polyphonic mixes of vocal music. Hence, all recordings of vocal music become available as potential training data. The original W-PST method approximated the mixture of sounds from a singing voice and other instruments by using the first-order Taylor expansion. However, the equation obtained by the method was too complicated to train the models. Thus, we developed a new approximative optimization method based on a random sampling from probabilistic distribution and enabled W-PST to estimate the spectral envelope of a clean singing voice from the polyphonic music. Moreover, in this paper, we evaluated our method using concurrent F0 and phoneme estimation tasks.

A great deal of research has been conducted on lyrics or phoneme recognition in the polyphonic singing voice [2, 3, 4, 5], and vocal F0 estimation [6, 7, 8, 9, 10]; however, to our knowledge, no studies have concurrently estimated both F0 and phoneme. The approaches of these studies differ from ours in that they either ignored the influence of accompaniment sound or segregated the singing voice from other instruments.

The rest of this paper is organized as follows. In the next section, the original W-PST method is described. Section 3 explains how the original W-PST method was extended. In Section 4, we describe the concurrent estimation of F0 and phoneme experiments, which was not in our previous paper[1]. In Section 5, we draw conclusions and point out future directions.

## 2. W-PST METHOD

This section gives an overview of the W-PST method [1]. This method consists of the following: (i) modeling of the singing voice accompanied by the other instruments based on templates of spectral envelope of singing voice, and (ii) estimation of the models that represent the templates of the spectral envelope. Section 2.1 and 2.2 explain (i) and (ii), respectively.

### 2.1. Modeling of a Singing Voice in a Polyphonic Music

W-PST method expresses the observed spectrum of a singing voices "as is" without segregation by stochastically modeling the generation process of the spectrogram of a singing voice with accompaniment sounds.

#### *2.1.1. Probabilistic spectral template*

We assume a spectrum of polyphonic sound mixture, $y(f)$, is generated from the probabilistic variables $Y_f$. We call these variables *the probabilistic spectral template*. Here, $f$ represents a frequency in log scale, and $y$ represents a spectral power in log scale. We then assume that $Y_f$ can be expressed by two different probabilistic spectral templates, $Y'_{v,f}$ and $Y_{n,f}$, and a function $H(f; f_0)$, which depends on a frequency $f$, as the following equation (Figs. 1 and 2),

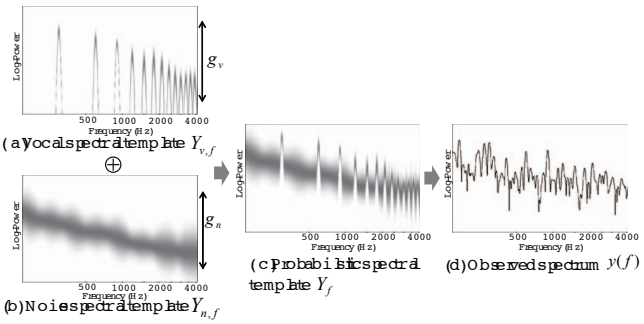



**Fig. 1**. Generation process of the observed spectrum [1]. The probability values are indicated by darkness.

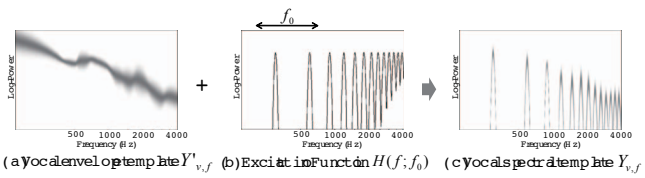



**Fig. 2**. Example of vocal spectral template [1], which is generated from the vocal envelope template and the excitation function.

$$Y_f = \log(\exp(Y_{v,f} + g_v) + \exp(Y_{n,f} + g_n)), \quad (1)$$

$$Y_{v,f} = Y'_{v,f} + H(f; f_0), \quad (2)$$

$$H(f; f_0) = \sum_h \mathcal{N}(f; \log f_0 + \log h, \sigma_H^2). \quad (3)$$

Here, $Y'_{v,f}$ represents a spectrum of a vocal, which is called *the vocal spectral template*, and $Y_{n,f}$ represents that of the other instrumental sounds, which is called *the noise spectral template*. $H(f; f_0)$ represents a spectrum of a vocal cord vibration of pitch $f_0$, which is called *an excitation function*. $g_v$ and $g_n$ represent the gain parameters. By changing them, the S/N ratio of the vocal and noise spectral templates can be controlled. Note that we assume the additivity of the power spectrum is in the linear scale and the source-filter model.

We assume that $Y'_{v,f}$ and $Y_{n,f}$ follow the Gaussian distribution (in log scale) and are represented by $Y'_{v,f} \sim \mathcal{N}(y; \mu'_{v,f}, \sigma^2_{v,f})$, $Y_{n,f} \sim \mathcal{N}(y; \mu_{n,f}, \sigma^2_{n,f})$, where $\mathcal{N}(y; \mu, \sigma^2)$ represents the Gaussian distribution with mean $\mu$ and variance $\sigma^2$.

Since the probabilistic spectral template expressed in (1) is difficult to calculate, we approximate $Y_f$ using a Gaussian distribution based on the first-order Taylor expansion as follows,

$$Y_f \sim \mathcal{N}(y; \mu_f, \sigma_f^2) \quad (4)$$

$$\mu_f = \log(\exp(\mu_{v,f} + g_v) + \exp(\mu_{n,f} + g_n)) \quad (5)$$

$$\mu_{v,f} = \mu'_{v,f} + H(f; f_0) \quad (6)$$

$$\sigma_f^2 = \frac{(\exp(\mu_{v,f} + g_v))^2 \sigma^2_{v,f} + (\exp(\mu_{n,f} + g_n))^2 \sigma^2_{n,f}}{(\exp(\mu_{v,f} + g_v) + \exp(\mu_{n,f} + g_n))^2}. \quad (7)$$

### 2.1.2. Phoneme recognition and F0 estimation

To recognize a phoneme using this model, first, an individual template, $\theta_v^p$, for each phoneme $p$ has to be prepared. Given the F0 of the singing voice $f_0$ and the observed spectra $y(f)$, we can estimate a pair of the phoneme and the F0, $\{\hat{i}, \hat{f}_0\}$, involved in the spectra by the following equation:

$$\{\hat{i}, \hat{f}_0\} = \underset{i, f_0}{\text{argmax}} \max_{g_v, g_n} \sum_f \log p_f(y(f); \theta_{i,v}, \theta_n, f_0, g_v, g_n) \quad (8)$$

$$\approx \underset{i, f_0}{\text{argmax}} \max_{g_v, g_n} \sum_f \log \mathcal{N}(y(f); u_f(\theta_{i,v}, \theta_n, f_0, g_v, g_n), \sigma_f^2(\theta_{i,v}, \theta_n, f_0, g_v, g_n)) \quad (9)$$

where $u_f$ and $\sigma_f^2$ are defined by (5) and (7), respectively.

To calculate Eq. (9), we need to optimize the parameter $\theta = (g_v, g_n, f_0)$ using the quasi-Newton method based on the BFGS (Broyden-Fletcher-Goldfarb-Shanno) method, which is a class of hill-climbing optimization techniques.

## 2.2. Estimation of Spectral Templates from a Monophonic Singing Voice

We estimate a spectral envelope that is represented by $Y'_{v,f}$ in Eq. (1) from a monophonic singing voice. The spectral envelope of the singing voice cannot be directly observed; what we can observe is a harmonic structure that is considered to be points sampled from the original spectral envelope. Thus, in general, it is difficult to estimate the original spectral envelope from a single harmonic structure. We overcome this difficulty and estimate a reliable spectral envelope using many harmonic structures with various F0s. Moreover, since we estimate the spectral envelope as a set of probabilistic distributions, the estimated envelope is robust against the fluctuation of the singing voice and the difference in conditions between the training and testing data.

Since volumes could differ from frame to frame, we need a scheme to normalize such volume differences when we estimate the envelope from many harmonic structures. We consider the volume of each frame as an unknown parameter, and estimate it concurrently with the parameters of the model for estimating the spectral envelope.

### 2.2.1. Mixture of experts

For a spectral template model, we use the mixture of experts (MoE) model [11] based on the linear regression model. This model represents $\mu_{v,f}$ and $\sigma^2_{v,f}$ of a spectral template as

$$\mu_{v,f} = \sum_i G_m(f|\psi_m, \mu_m, \sigma_m^2)(a_m f + b_m) \quad (10)$$

$$\sigma^2_{v,f} = \sum_i G_i(f|\psi_m, \mu_m, \sigma_m^2)^2 \beta_m^2, \quad (11)$$

where $G_m(f|\psi_m, \mu_m, \sigma_m^2)$ is the output of the gating network, and we use a normalized Gaussian function [12] defined by

$$G_m(f|\psi_m, \mu_m, \sigma_m^2) = \frac{\psi_m \mathcal{N}(f|\mu_m, \sigma_m^2)}{\sum_{m'} \psi_{m'} \mathcal{N}(f|\mu_{m'}, \sigma^2_{m'})}. \quad (12)$$

Here, $\{\psi_m, \mu_m, \sigma_m^2, a_m, b_m, \beta_m^2\}$ is a set of unknown parameters, where $\psi_m$ satisfies $\psi_m \geq 0$ and $\sum_m \psi_m = 1$. These parameters can be estimated using the expectation-maximization (EM) algorithm.

### 2.2.2. Iterative parameter estimation

When we observe harmonic structures $s_i (i = 1, \ldots, I)$, which consist of the log power of the $h$-th harmonic component, $y_{i,h}$, and its frequency, $f_{i,h}$, denoted by

$$s_i = \{(f_{i,1}, y_{i,1}), \ldots (f_{i,h}, y_{i,h}), \ldots (f_{i,H_i}, y_{i,H_i})\}, \quad (13)$$

the target likelihood function to be maximized is defined by

$$L = \sum_{i=1}^{I} \sum_{h=1}^{H_i} \log \mathcal{N}(y_{i,h} + k_i; \mu_{v,f_{i,h}}, \sigma_{v,f_{i,h}}), \quad (14)$$

where $k_i$ represents the offset parameter, which normalizes the volume of the harmonic structure. Since it is difficult to estimate both $k_i$ and the parameters of MoE at the same time, we update them sequentially. As for the noise spectral envelope, we can estimate the parameter in the same manner by considering the spectrum as $s_i (i = 1, \ldots, I)$.

## 3. ESTIMATION OF SPECTRAL TEMPLATES FROM A SINGING VOICE IN A POLYPHONIC MUSIC

In this section, we describe a new method to expand the W-PST method so that the template can be estimated from polyphonic singing voices. We concurrently estimate the parameters of the voice envelope template and noise spectral template.

In Section 2.1.1, we approximate the sum of log-normal distribution using the first-order Taylor expansion. However, the obtained equations, Eq. (4)!A(7), have become so complex that it is difficult to use these approximated equations to estimate the parameters of the templates. Our approach to estimate them is to strictly calculate the sum of log-normal distribution and estimate the parameters based on an approximative algorithm.

### 3.1. Basic formulation of an objective function

The observed spectra are represented as $y_1(f), \cdots, y_i(f), \cdots, y_I(f)$ and the parameters of the voice envelope template and the noise spectral template to be estimated are written as

$$\theta_v = \{\psi_{v,m}, \mu_{v,m}, \sigma^2_{v,m}, a_{v,m}, b_{v,m}, \beta^2_{v,m}\}, \quad (15)$$

$$\theta_n = \{\psi_{n,m}, \mu_{n,m}, \sigma^2_{n,m}, a_{n,m}, b_{n,m}, \beta^2_{n,m}\}, \quad (16)$$

respectively. The mean parameter of the vocal spectral template is calculated as $\mu_{v,f,i} = \mu'_{v,f} + H(f; f_0(i))$ by adding the excitation function. Note that we assume the F0 sequence of the observed spectrum, $f_0(0), \cdots, f_0(i), \cdots, f_0(I)$, is given.

The objective function to be maximized is written as

$$L = \int \sum_{i=1}^{I} \log p_{i,f}(y; \theta_v, \theta_n, g_{i,v}, g_{i,n}) df \quad (17)$$

$$= \int \sum_{i=1}^{I} \log \Bigg( \int_{-\infty}^{y_i(f)} \mathcal{N}(\log(\exp(y_i(f)) - \exp(U)) ; \mu_{v,f,i} + g_{i,v}, \sigma^2_{v,f}) \mathcal{N}(U; \mu_{n,f} + g_{i,n}, \sigma^2_{n,f}) \frac{\exp(y_i(f))}{\exp(y_i(f)) - \exp(U)} dU \Bigg) df \quad (18)$$

$$= \int \sum_{i=1}^{I} \log( \int_{-\infty}^{y_i(f)} \mathcal{N}(\log(\exp(y_i(f)) - \exp(U)) ; \mu_{n,f} + g_{i,n}, \sigma^2_{n,f}) \mathcal{N}(U; \mu_{v,f,i} + g_{i,v}, \sigma^2_{v,f}) \frac{\exp(y_i(f))}{\exp(y_i(f)) - \exp(U)} dU) df \quad (19)$$

where $p_{i,f}(y; \theta_v, \theta_n, g_{i,v}, g_{i,n})$ represents a probabilistic density function of composed spectral template[1]. Note that $g_{i,v}$ and $g_{i,n}$ are the parameters used to normalize the volume of each frame, which is similar to $k_i$ in Section 2.2.2 . Practically, the integral in the above equation is substituted in sum operations in the discrete frequency scale because continuous wavelet transform is calculate in discrete time. Here, the parameters to be estimated are $\{g_{i,v}, g_{i,n}, \theta_v, \theta_n\}$.

### 3.2. Approximative parameter estimation based on a random sampling

Since it is difficult to estimate all the parameters at the same time, we estimate them successively and iteratively. We first estimate $g_{i,v}$ and $\theta_v$ based on Eq. (18) considering $g_{i,n}$ and $\theta_n$ as constants and we then estimate $g_{i,n}$ and $\theta_n$ based on Eq. (19) considering $g_{i,v}$ and $\theta_v$ as constants. When we consider $g_{i,n}$ and $\theta_n$ as constants, Eq. (18) can be interpreted as an expectation operation. Thus, we approximate the expectation using the sum operation by sampling the probabilistic variable $U$ that follows truncated normal distribution. To be more precise, we obtain $R$ random values $(U_{i,1,f}, \cdots, U_{i,r,f}, \cdots, U_{i,R,f})$, which have the normal distribution $\mathcal{N}(U; \mu_{n,f} + g_{i,n}, \sigma^2_{n,f})$ and lies within the interval $(-\infty, y_i(f))$. The objective function $L$ can be approximated as

$$L \approx \int \sum_{i=1}^{I} \log \sum_{r=1}^{R} \pi_{i,r,f} \mathcal{N}(\log(\exp(y_i(f)) - \exp(U_{i,r,f})) ; \mu_{v,f,i} + g_{i,v}, \sigma^2_{v,f}) \quad (20)$$

$$\pi_{i,r,f} = \frac{\exp(y_i(f))}{(\exp(y_i(f)) - \exp(U_{i,r,f}))\, C_{y_i(f),i,f} R} \quad (21)$$

$$C_{y,i,f} = \int_{-\infty}^{y} \mathcal{N}(U; \mu_{n,f} + g_{i,n}, \sigma^2_{n,f}) dU \quad (22)$$

It should be noted that $\pi_{i,r,f}$ and $\log(\exp(y_i(f)) - \exp(U_{i,r,f}))$ become constants if $g_{i,n}$ and $\theta_n$ are constants, and we can thus estimate $g_{i,v}$ and $\theta_v$ by using Eq. (20). Similar to Eq. (18), $g_{i,n}$ and $\theta_n$ can be updated in by considering $g_{i,v}$ and $\theta_v$ as consonants.

### 3.3. An EM-like method to optimize a logarithm of a sum

However, Eq. (20) is the shape of logarithm of a sum, which is known to be difficult to maximize directly. Thus, we maximize Eq. (20) by an iterative method resembling the EM algorithm. Hereafter, parameters to be estimated are denoted as $\lambda = \{g_{i,v}, \theta_v\}$ for convenience and $\lambda'$ represents the updated parameters in the previous iteration. First, we introduce new variables as

$$z_{i,r,f} = \frac{\pi_{i,r,f} \psi_{i,r,f}}{\sum_{r'=1}^{R} \pi_{i,r',f} \psi i, r', f} \quad (23)$$

$$\psi_{i,r,f} = \mathcal{N}(\log(\exp(y_i(f)) - \exp(U_{i,r,f})) ; \mu_{v,f,i} + g_{i,v}, \sigma^2_{v,f}) \quad (24)$$

and write $z_{i,r,f}$ calculated by using $\lambda'$ as $z'_{i,r,f}$. $L$ can be maximized by iterating the following two steps: (i) estimate $\lambda$ that maximizes a new objective function $Q_1(\lambda|\lambda')$,

$$Q_1(\lambda|\lambda') = \int \sum_{i=1}^{I} \sum_{r=1}^{R} z'_{i,r,f} \log \pi_{i,r,f} \mathcal{N}(\log(\exp(y_i(f)) - \exp(U_{i,r,f})); \mu_{v,f,i} + g_{i,v}, \sigma^2_{v,f}) df \quad (25)$$

and (ii) calculate $z_{i,r,f}$ using the estimated $\lambda$.

Since the term $\pi_{i,r,f}$ in Eq. (25) is irrelevant to the maximization of $Q_1(\lambda|\lambda')$, we can use new objective function $Q_2(\lambda|\lambda')$,

$$Q_2(\lambda|\lambda') = \int \sum_{i=1}^{I} \sum_{r=1}^{R} z'_{i,r,f} \log \mathcal{N}(\log(\exp(y_i(f)) - \exp(U_{i,r,f})) ; \mu_{v,f,i} + g_{i,v}, \sigma^2_{v,f}) df \quad (26)$$

instead of $Q_1(\lambda|\lambda')$. Finally, since $Q_2$ follows the same form as Eq. (14) in Sec 2.2.2, we can maximize $Q_2$ by using the EM algorithm in the same manner as the template estimation from a monophonic singing voice.

## 4. EXPERIMENTS

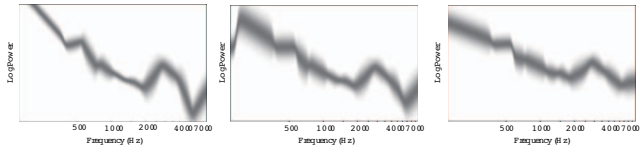
We conducted experiments on phoneme recognition using 10 Japanese songs performed by 6 singers (3 male, 3 female), taken from the "RWC Music Database: Popular Music" (RWC-MDB-P-2001) [13]. The target phonemes were all 5 Japanese vowels; /a/, /i/, /u/, /e/, and /o/. We conducted a 6 fold cross validation; that is, when we evaluated a song of a specific singer, the vocal and noise templates (we call them phoneme model) were trained using songs from the other 5 singers. We used a manually annotated phoneme label and F0 annotation of the songs for both training data and ground-truth. Accuracy for both phoneme estimation and F0 estimation is defined as the ratio of the number of frames that are correctly esti-

[1] We added an index of the observed spectrum $i$ because the shape of the PDFs are different from frame to frame here.

**Table 1**. Experimental results for concurrent estimation of phonemes and F0 (%).

| | Baseline | | W-PST | | Extended W-PST | |
|---|---|---|---|---|---|---|
| Song* | Phoneme | F0 | Phoneme | F0 | Phoneme | F0 |
| No. 4 | 31.1 | 62.6 | 73.5 | 58.9 | 70.2 | 55.4 |
| No. 11 | 56.5 | 65.6 | 57.6 | 71.5 | 57.4 | 71.6 |
| No. 9 | 47.5 | 65.5 | 43.4 | 43.3 | 44.0 | 43.2 |
| No. 12 | 62.8 | 76.8 | 63.9 | 77.6 | 63.5 | 77.3 |
| No. 6 | 51.5 | 69.2 | 60.4 | 80.8 | 58.0 | 81.0 |
| No. 2 | 69.5 | 71.6 | 68.5 | 86.3 | 68.5 | 85.0 |
| No. 16 | 62.7 | 78.2 | 65.4 | 82.6 | 63.0 | 80.1 |
| No. 7 | 60.0 | 73.8 | 67.2 | 82.7 | 65.4 | 79.6 |
| No. 18 | 64.1 | 73.5 | 70.2 | 87.6 | 68.5 | 86.4 |
| No. 14 | 44.1 | 79.1 | 42.3 | 82.0 | 42.0 | 82.5 |
| Average | 55.0 | 71.6 | 61.2 | 75.3 | 60.1 | 74.2 |

*Song number of RWC-MDB-P-2001[13].




(a) A template estimated from a monophonic singing voice using a method described in Sec 2.2. (b) A template estimated from a polyphonic music using our method (c) A template estimate from a polyphonic music using a method described in Sec 2.2.

**Fig. 3**. An example of spectral estimation from a singing voice of a polyphonic music. This example represents an envelope of phoneme /i/ in song No. 7 in the RWC-MDB-P-2001 [13].

mated to the total number of frames. Only frames that involve the target 5 vowels were used for calculating the accuracy.

We tested our method under the following 3 conditions.

**(i) Baseline** Use the F0 estimation method called PreFEst [14] and the feature extraction method used in [15]; segregate the singing voice based on the harmonic structure before extracting MFCCs, ΔMFCCs, and ΔPower and recognize them using the GMMs.

**(ii) W-PST** Use the method described in [1].

**(iii) Extended W-PST** Use the W-PST method with the new template estimation algorithm from polyphonic singing voices proposed in this paper.

In condition (i) , we used the Short Time Fourier Transform for spectrum analysis and we set the number of mixtures of GMMs and the number of dimensions of MFCCs to 12 and 32, respectively. The data used for training GMMs were also segregated. In conditions (ii) and (iii), we used the wavelet transform with the Gabor wavelet for spectrum analysis, and set the number of mixtures of the MoE to 10. In condition (ii), the vocal templates were trained using the vocal-only tracks of the songs, and the noise template was trained using a karaoke (without vocal) track of the songs based on the harmonic structures. In condition (iii), the vocal and noise templates were trained using polyphonic tracks of the songs.

The results are summarized in Table 1. We can see that the original W-PST method (ii) increased the average accuracy by 6.2 points for phoneme estimation and 3.7 points for F0 estimation compared to baseline method (i). We confirmed that the W-PST method can estimate F0 and phonemes better than the conventional method. We can also see that the extended W-PST method (iii), though not as accurate as the original W-PST method, is still more accurate than baseline method (i) by 5.1 points for phoneme estimation and 2.6 points for F0 estimation. These results show that our method worked well even if monophonic singing voices are not available for training data.

Fig. 3 shows examples of a spectrum estimated from a monophonic singing voice using our previous method described in 2.2.2 (Fig. 3 (a)), and a spectrum estimated from a singing voice in a polyphonic music using a method proposed in this paper (Fig. 3 (b)), and a spectrum estimated from a singing voice segregated from a polyphonic music in using the accompaniment sound reduction method proposed in [15] (Fig. 3 (c)). We can see the spectrum proposed by the proposed method (b) is closer to the spectrum of optimal case (a) than the spectrum directly estimated from a polyphonic music without dealing with the accompaniment sound (c).

## 5. CONCLUSION

We described how to expand a method called W-PST [1] that concurrently estimates the F0 and phonemes of a polyphonic singing voice. Since the original W-PST method could train models only from monophonic singing voices, we made it possible to train models from polyphonic singing voices by developing a method for estimating spectral envelopes from polyphonic singing voices. We conducted concurrent F0 and phoneme estimation experiments and found that our method indicated improved performance compared to the conventional method. In the future, we plan to extend this method to deal with consonant phonemes and vocal activity detection. Future plan also includes integration with temporal modeling methods such as the hidden Markov models (HMMs).

## 6. REFERENCES


[1] Hiromasa Fujihara, Masataka Goto, and Okuno G. Hiroshi, "A novel framework for recognizing phonemes of singing voice in polyphonic music," in *Proc. WASPAA*, 2009, pp. 17–20.

[2] Matthias Gruhne, Konstantin Schmidt, and Christian Dittmar, "Phoneme recognition in popular music," in *Proc. ISMIR*, 2007, pp. 369–370.

[3] Kai Chen, Sheng Gao, Yongwei Zhu, and Qibin Sun, "Popular song and lyrics synchronization and its application to music information retrieval," in *Proc. MMCN'06*, 2006.

[4] Denny Iskandar, Ye Wang, Min-Yen Kan, and Haizhou Li, "Syllabic level automatic synchronization of music signals and text lyrics," in *Proc. ACM Multimedia*, 2006, pp. 659–662.

[5] Annamaria Mesaros and Tuomas Virtanen, "Automatic recognition of lyrics in singing," *EURASIP J. Audio, Speech, Music Process.*, vol. 2010, 2010.

[6] Yipeng Li and DeLiang Wang, "Detecting pitch of singing voice in polyphonic audio," in *Proc. ICASSP*, 2005, pp. III–17–20.

[7] Matti Ryynänen and Anssi Klapuri, "Transcription of the singing melody in polyphonic music," in *Proc. ISMIR*, 2006, pp. 222–227.

[8] Christopher Sutton, Emmanuel Vincent, Mark D. Plumbley, and Juan P. Bello, "Transcription of vocal melodies using voice characteristics and algorithm fusion," in *Proc. MIREX*, 2006.

[9] Jean-Louis Durrieu, Gaël Richard, and Bertrand David, "Singer melody extraction in polyphonic signals using source separation methods," in *Proc. ICASSP*, 2008, pp. 169–172.

[10] Jean-Louis Durrieu, Gaël Richard, and Bertrand David, "An iterative approach to monaural musical mixture de-soloing," in *Proc. ICASSP*, 2009, pp. 105–108.

[11] R. J. Jacobs, M.I Jordan, S. J. Nowlan, and G. E Hinton, "Adaptive mixtures of local experts," *Neural Computation*, vol. 3, pp. 79–87, 1991.

[12] L Xu, M. I. Jordan, and G. E. Hinton, "An alternative model for mixtures of experts," *Advances in Neural Information Processing Systems 7*, pp. 633–640, 1994.

[13] Masataka Goto, Hiroki Hashiguchi, Takuichi Nishimura, and Ryuichi Oka, "RWC Music Database: Popular, classical, and jazz music databases," in *Proc. ISMIR*, 2002, pp. 287–288.

[14] Masataka Goto, "A real-time music-scene-description system: Predominant-f0 estimation for detecting melody and bass lines in real-world audio signals," *Spe. Comm.*, vol. 43, no. 4, pp. 311–329, 2004.

[15] Hiromasa Fujihara, Masataka Goto, Tetsuro Kitahara, and Hiroshi G. Okuno, "A modeling of singing voice robust to accompaniment sounds and its application to singer identification and vocal-timbre-similarity based music information retrieval," *IEEE Trans. Audio, Speech, Language Process.*, vol. 18, no. 3, pp. 638–648, 2010.