A NOVEL FRAMEWORK FOR RECOGNIZING PHONEMES OF SINGING VOICE IN POLYPHONIC MUSIC

Hiromasa Fujihara,[†,‡] *Masataka Goto,*[†] *and Hiroshi G. Okuno*[‡]

[†]National Institute of Advanced Industrial Science and Technology (AIST)
Tsukuba, Ibaraki 305-8568, Japan

[‡]Kyoto University
Sakyo-ku, Kyoto 606-8501, Japan

ABSTRACT

A novel method is described that can be used to recognize the phoneme of a singing voice (vocal) in polyphonic music. Though we focus on the voiced phoneme in this paper, this method is design to concurrently recognize other elements of a singing voice such as fundamental frequency and singer. Thus, this method is considered to be a new framework for recognizing a singing voice in polyphonic music. Our 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 many harmonic structures with various fundamental frequencies (F0s). The results of phoneme recognition experiments with 10 popular-music songs by 6 singers showed that our method improves the recognition accuracy by 8.7 points and achieves a 20.0% decrease in error rate.

Index Terms— Singing voice, Phoneme recognition, Spectral modeling, Mixture of experts

1. INTRODUCTION

Music is an important media content in both industrial and cultural aspects, and a singing voice (vocal) is one of the most important elements of music. Among the many other considerable elements of a singing voice, we deemed the lyrics, voice quality (or name of a singer), and fundamental frequencies (F0s) to be the most fundamental elements of a singing voice. To develop a computer that can recognize a singing voice, we have been working on an automatic synchronization method for singing voice and lyrics [1], an automatic singer identification method [2], and an F0 estimation method for a singing voice [3]. As a stepping stone to further develop these methods, we propose a novel framework for recognizing a singing voice within polyphonic music, which is named the W-PST (Weighted-composition of Probabilistic Spectral Template) method. We also verify the effectiveness of W-PST method by conducting a phoneme recognition experiment on condition that correct F0s are given. Although we mainly focused on lyrics, we designed W-PST method to be applicable in recognizing F0s and names of singers.

Research in automatic lyric recognition from polyphonic music is important because it can be directly applied to a new music information retrieval system. Even if perfect lyric recognition cannot be achieved, the basic techniques for recognizing the phonemes can be used in an automatic synchronization method for lyrics and music. However, automatic recognition of lyrics (or phonemes) of a singing voice is more difficult than that of speech because a singing voice fluctuates wildly due to the vibrato, a wide range of fundamental frequencies (F0s), and the emotional expression of singers. Furthermore, other instrumental sounds, which are usually accompanied with the singing voice, also degrade the performance of lyric (or phoneme) recognition.

We previously developed a method for automatically synchronizing lyrics with the corresponding singing voice [1]. Although this method segregated the singing voice from polyphonic music to reduce the negative influence of the accompaniment sound, it is generally based on conventional speech recognition techniques. That is, it extracts feature vectors that represent the spectral envelopes, such as Mel-frequency cepstral coefficients (MFCCs), and calculates the likelihood of the feature vectors using the Gaussian mixture models (GMMs).

Our previous method, however, exhibits the following two problems:

Problems of segregation The performance of singing voice segregation critically affects that of the recognition because the F0 estimation errors deteriorate segregation quality, and it cumulatively degrades recognition performance. Besides, the segregation process discards the other components of the spectrum, which include important information about the background noise (accompaniment sound) such as the S/N ratios and the degree of vocal distortion it causes.

Inaccurate estimation of spectral envelope Our previous method estimated a spectral envelope from a given (segregated) harmonic structure and calculated the distance between estimated envelopes. However, the harmonic components are considered to be points sampled from their original spectral envelopes at an interval of F0 along the frequency axis, and the perfect reconstruction of the spectral envelope from the harmonic structure is generally impossible. Therefore, it was difficult to calculate the distance of high-pitched sounds, such as a female singing voice.

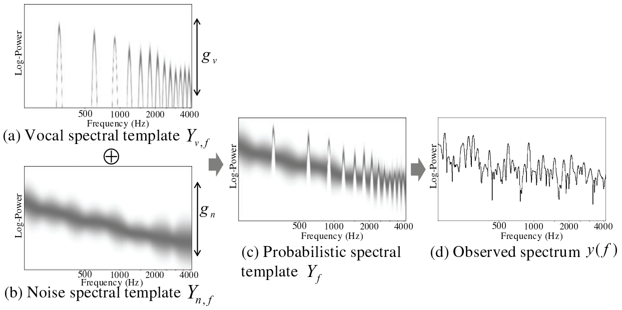
Gruhne *et. al.* [4] have worked on a phoneme recognition problem. They also segregated a vocal using a similar method used in [1]. Several studies have been done on automatic synchronization between music and lyrics [5, 6, 7, 8]. These methods usually use conventional speech recognition techniques or more simplified techniques and try to improve performance by integrating other information such as characteristics of the target language and structure of the music.

We propose a novel method for overcoming these problems. Our method neither segregates a singing voice nor reconstructs a spectral envelope from a single harmonic structure. Instead, it expresses the observed spectrum of a singing voices ”as is“, by stochastically modeling the generation process of the spectrogram of a singing voice with accompaniment sounds. In the training process, we develop a method that can estimate a more accurate spectral envelope by estimating it from many harmonic structures.

2. W-PST METHOD FOR RECOGNIZING SINGING VOICE

As shown in Figs. 1 (c) and (d), we assume that the observed spectrum of polyphonic music is generated from a set of probability distributions, which is called *the probabilistic spectral template*. That is, the power of each spectral bin follows a probabilistic distribution, and forms of the probabilistic distributions differ from bin to bin. Assuming the additivity of the power spectrum, we consider that probabilistic spectral templates are produced by adding two different probabilistic spectral templates in a linear scale; *a*

This research was supported in part by CrestMuse, CREST, JST.



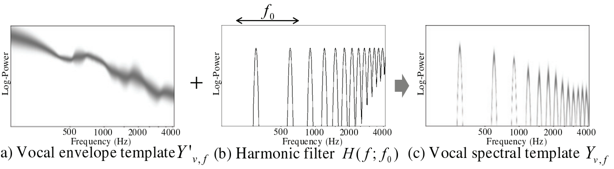
(a) Vocal spectral template $Y_{v,f}$

(b) Noise spectral template $Y_{n,f}$

(c) Probabilistic spectral template Y_f

(d) Observed spectrum $y(f)$

Figure 1. Generation process of the observed spectrum. The probability values are indicated by the darkness. Note that the S/N ratio can be controlled by the gain parameters, g_v and g_n.



(a) Vocal envelope template $Y'_{v,f}$ (b) Harmonic filter $H(f;f_0)$ (c) Vocal spectral template $Y_{v,f}$

Figure 2. Example of vocal spectral template, which generated from the vocal envelope template and the harmonic filter.

vocal spectral template (Fig. 1 (a)), which generates the spectrum of a singing voice, and *a noise spectral template* (Fig. 1 (b)), which generates the spectrum of accompaniment sounds. By introducing gain parameters for both templates and taking their weighted sum, we can represent the spectrum of various S/N ratios and volumes. A vocal spectral template is produced by multiplying a *vocal envelope template* (Fig. 2 (a)), which represents the spectral envelope, by a *harmonic filter* (Fig. 2 (b)), which represents the harmonic structure. The shape of the harmonic filter is controlled by an F0 parameter. The likelihood of observed spectrum for this probabilistic model can be calculated if the parameters of this model — the F0 of the harmonic filter and the gain parameters of the vocal and noise spectral templates — are given. Thus, we can estimate a phoneme involved in observed spectrum by preparing vocal envelope templates of various phonemes and selecting the most likely one (Fig. 3). We can also estimate the F0 of a singing voice by estimating the most likely value.

The novelties of W-PST method are summarized as follows:

- It does not segregate singing voice, but our framework recognizes a singing voice directly from polyphonic sound mixtures. Since humans understand a singing voice without segregating it, this process is natural from the standpoint of human perception.
- It is robust against fluctuations in accompaniment sound because it estimates the S/N ratio of the moment. It becomes more robust if we prepare various types of noise spectral templates and select the most likely one.
- It is robust against high-pitched sounds because it does not estimate an envelope from a single harmonic structure.
- It is easy to integrate unvoiced consonants by only preparing new templates for these phonemes (without using the harmonic filter).

3. FORMULATION

This section describes a concrete formulation of W-PST method described in Sec. 2. To implement it, we have to design a method for overcoming the following three issues,

- how to represent a probabilistic spectral template,
- how to calculate the sum of two probabilistic spectral templates, and

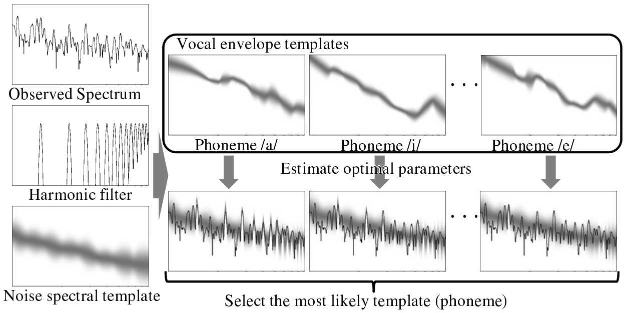


Figure 3. Overview of phoneme estimation process.

- how to optimize gain parameters (volumes).

Our approaches to overcome these issues are as follows,

- We assume the distributions of a probabilistic spectral template as log-Gaussian distributions.
- We approximate that the sum of probability variables that follow the log-Gaussian distributions also follows the log-normal distribution[1].
- We optimize the parameters using the quasi-Newton method.

3.1. 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 divided into two different probabilistic spectral templates, $Y_{v,f}$ and $Y_{n,f}$, by the following equation,

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

where $Y_{v,f}$ represents the spectrum of a vocal, which is called *the vocal spectral template*, and $Y_{n,f}$ represents that of other instrumental sounds, which is called *the noise spectral template*. Here, g_v and g_n represent the gain parameters. By changing g_v and g_n, this model can be used to control the S/N ratio of the vocal and noise spectral templates. Note that we assume the additivity of the power spectrum in linear scale.

We assume that $Y_{v,f}$ and $Y_{n,f}$ follow the Gaussian distribution (in log scale) and are represented by

$$Y_{v,f} \sim \mathcal{N}(y; \mu_{v,f}, \sigma^2_{v,f}), \tag{2}$$

$$Y_{n,f} \sim \mathcal{N}(y; \mu_{n,f}, \sigma^2_{n,f}), \tag{3}$$

where $\mathcal{N}(y; \mu, \sigma^2)$ represents the Gaussian distribution with mean μ and variance σ^2.

Assuming the source-filter model, we consider the harmonic sound $Y_{v,f}$ (singing voice) to be a sum of the probabilistic model of the envelope and a filter function, which represents the harmonic structure in log scale (Fig 2).

$$Y_{v,f} = Y'_{v,f} + \log H(f; f_0) \tag{4}$$

$$\sim \mathcal{N}(y; \mu'_{v,f} + \log H(f; f_0), \sigma^2_{v,f}) \tag{5}$$

$$H(f; f_0) = \sum_h \mathcal{N}(f; \log f_0 + \log h, \sigma^2_H), \tag{6}$$

where $Y'_{v,f} \sim \mathcal{N}(y; \mu'_{v,f}; \sigma^2_{v,f})$ represents probabilistic variables of the vocal envelope, which is called *the vocal envelope template*, and $H(f; f_0)$ represents the filter of F0, f_0, which is called *the harmonic filter*. Note that the harmonic filter $H(f; f_0)$ is not a probabilistic variable.

3.2. Approximation of sum of two spectral templates

Since a probabilistic spectral template Y_f, which is expressed in (1), is difficult to calculate, we approximate Y_f using the Gaus-

[1] The sum of probability variables that follow the log-Gaussian distributions does not follow a log-normal distribution in general.

sian distribution. The first-order Taylor series of the function $f(x_1, x_2) = \log(\exp(x_1) + \exp(x_2))$ at $(\mu_{v,f} + g_v, \mu_{n,f} + g_n)$ can be denoted by linear combination of x_1 and x_2. Since linear combination of probabilistic variables that follow the Gaussian distribution also follows the Gaussian distribution, thus, $Y_f = f(Y_{v,f}, Y_{n,f})$ can be approximated as follows.

$$Y_f \sim \mathcal{N}(y; \mu_f, \sigma_f^2) \tag{7}$$

$$\mu_f = \log(\exp(\mu_{v,f} + g_v) + \exp(\mu_{n,f} + g_n)) \tag{8}$$

$$\sigma_f^2 = \frac{(\exp(\mu_{v,f} + g_v))^2 \sigma_{v,f}^2 + (\exp(\mu_{n,f} + g_n))^2 \sigma_{n,f}^2}{(\exp(\mu_{v,f} + g_v) + \exp(\mu_{n,f} + g_n))^2} \tag{9}$$

3.3. Phoneme recognition and F0 estimation

To recognize a phoneme by using this model, first, an indivisual template, θ_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 the phoneme, $\hat{p}$, involved in the spectra by the following equation,

$$\hat{p} = \operatorname*{argmax}_p \max_{g_v, g_n} \int_f \log \mathcal{N}(s(f); u_f, \sigma_f^2), \tag{10}$$

where u_f and σ_f are defined by (8) and (9), respectively. Although we did not evaluate F0 estimation in the experiments in this paper, we can concurrently estimate F0 and a phoneme by taking the argmax of f_0. When determining individual singers, we prepare an individual template for each singer.

3.4. Optimization using quasi-Newton method

We optimize the parameter $\theta = (g_v, g_n)$ using the quasi-Newton method based on the BFGS (Broyden-Fletcher-Goldfarb-Shanno) method, which is a class of hill-climbing optimization techniques. Given the observed spectrum $s(f)$, the target function to be minimized, $Q(\theta)$, becomes as follows;

$$Q(\theta) = -\int_f \log \mathcal{N}(s(f); u_f, \sigma_f^2). \tag{11}$$

4. ESTIMATION OF SPECTRAL ENVELOPE

We estimate a spectral envelope that is represented by $Y'_{v,f}$ in (4). The spectral envelope of the singing voice cannot be directly observed because what we can observe is a harmonic structure which is considered to be points sampled from its original spectral envelope. Thus, it is difficult to estimate its original spectral envelope from a single harmonic structure in general. 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.

4.1. Mixture of Experts

For a spectral template model, we use the mixture of experts (MoE) model [9] based on the linear regression model. This model represents $\mu_{v,f}$ and $\sigma_{v,f}^2$ of a spectral template as

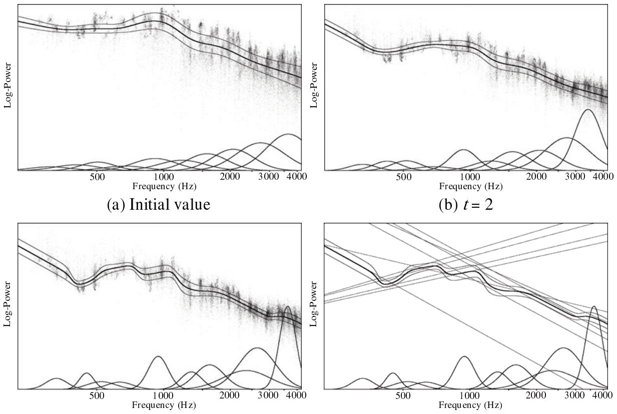
$$\mu_{v,f} = \sum_i G_m(f|\psi_m, \mu_m, \sigma_m^2)(a_m f + b_m) \tag{12}$$

$$\sigma_{v,f}^2 = \sum_i G_i(f|\psi_m, \mu_m, \sigma_m^2)^2 \beta_m^2, \tag{13}$$

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 [10] defined by

$$G_m(f|\psi_m, \mu_m, \sigma_m^2) = \frac{\psi_m \mathcal{N}(f|\mu_m, \sigma_m^2)}{\sum_{m'} \psi_{m'} \mathcal{N}(f|\mu_{m'}, \sigma_{m'}^2)}. \tag{14}$$

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



(a) Initial value (b) $t = 2$

(c) $t = 20$ (d) Each expert ($t = 20$)

Figure 4. Example of parameter estimation of the mixture of experts (MoE), where t represents the number of iteration. The middle line represents the mean of the MoE and the two thin lines around it represent the standard deviation. The background minute dots represent the harmonic components of the training data. The gating functions, $G_i(f|\Phi)$, are drawn at the bottom.

4.2. 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

$$s_i = \{(f_{i,1}, y_{i,1}), \ldots (f_{i,h}, y_{i,h}), \ldots (f_{i,H_i}, y_{i,H_i})\}, \tag{15}$$

the target likelihood function to be maximized is defined by

$$L = \sum_i^N \sum_h^{H_i} \mathcal{N}(y_{i,h} + k_i; \mu_{v,f_{i,h}}, \sigma_{v,f_{i,h}}), \tag{16}$$

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.

The procedure of parameter estimation is summarized as follows.

Step 0 Set $k_i = 0$ and initialize the other parameters.

Step 1 Update parameters of the MoE using the EM algorithm.

Step 2 Update k_i using the following equation,

$$k_i = \frac{\sum_{h=1}^{H_i} \frac{\mu_{v,f_{i,h}} - y_{i,n}}{\sigma_{v,f_{i,h}}^2}}{\sum_{h=1}^{H_i} \frac{1}{\sigma_{v,f_{i,h}}^2}} \tag{17}$$

Step 3 Go back to step 1.

Figure 4 shows an example of the parameter estimation process. As for the noise spectral envelope, we can estimate the parameter in the same manner by considering the spectrum as $s_i (i = 1, \ldots, I)$.

5. EXPERIMENTS

We conducted experiments on phoneme recognition by using 10 Japanese songs performed by 6 singers (3 male, 3 female) taken from the "RWC Music Database: Popular Music" (RWC-MDB-P-2001)[11]. The target phonemes were 5 Japanese vowels; "a", "i",

Table 1. Experimental results (%).

			(i)	(ii)	(iii)
Song #*	Gender***	Singer	Baseline 1	Baseline 2	Proposed
No. 4	M	A	31.1**	33.0**	64.3
No. 11	M	A	52.0	57.1	63.0
No. 9	M	B	30.0**	48.4	52.6
No. 12	M	B	33.8**	67.5	69.3
No. 15	M	C	42.6**	50.8	61.7
No. 2	F	D	59.1	70.7	70.7
No. 16	F	D	57.2	63.1	69.9
No. 7	F	E	54.4	62.3	70.2
No. 18	F	E	59.0	66.9	71.6
No. 14	F	F	40.4	43.9	46.2
Average			46.0	56.4	65.1

*Song number of RWC-MDB-P-2001[11].
**Songs of which the model of the difference gender was selected.
***M and F denotes male and female, respectively.

"u", "e", and "o". We conducted 6 fold cross validation, that is, when we evaluated a song of a singer, the vocal and noise templates (we call them phoneme model) were trained using songs of the other 5 singers. We used a manually annotated phoneme label of the songs for both training data and ground-truth. Accuracy is defined as the ratio of the number of frames that are correctly estimated to the total number of frames. Only frames that involve the target 5 vowels are used for calculating the accuracy. We used the gender-dependent phoneme models here, that is, we trained the phoneme models for male and female separately. When testing, we calculate a likelihood for each phoneme model and select a result of the most likely model.

We tested our method under the following 3 conditions.
(i) Baseline 1 Extract MFCCs, ΔMFCCs, and ΔPower from polyphonic audio signals and recognize them using the GMMs.
(ii) Baseline 2 Use the feature extraction method used in [1]; segregate the singing voice based on the harmonic structure before extracting MFCCs, ΔMFCCs, and ΔPower and recognize them using the GMMs.
(iii) Proposed Use the W-PST method proposed in this paper.

Note that manually-annotated F0s of the singing voice were used for conditions (ii) and (iii) for both training and testing.

In condition (iii), we used the wavelet transform with the Gabor wavelet for spectrum analysis, and set the number of mixture of the MoE to 10. The vocal and noise templates were created by combining the templates that are trained using each of the songs used for training. Therefore, there were a number of templates for each phoneme in the templates. 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. In condition (i) and (ii), we used the Short Time Fourier Transform for spectrum analysis, and set the number of mixtures of GMMs and the number of dimensions of MFCCs to 12 and 32, respectively. As for condition (ii), the data used for training GMMs were also segregated based on the harmonic structures.

The results are summarized in Table 1. We can see that our method increased the average accuracy by 8.7 points compared to the result of condition (ii). We can also see that there was no song of which the accuracy was decreased. As shown in Table 1, our method was able to select the model of the proper gender for all songs, while baseline methods failed to select the model of the proper gender for some songs. Inspection of the incorrect frames with our method (condition (iii)) and the correct frames with the baseline method (condition (ii)) showed that 52.6% of incorrect frames of our method was correctly recognized using the baseline method. By this fact, we can say that there is a possibility to further improve performance by combining our method with the baseline method.

6. DISCUSSION AND CONCLUSIONS

We have developed the W-PST method for recognizing the phonemes of a singing voice in polyphonic music. Although we evaluated this method through only phoneme recognition experiments, it is directly applicable to F0 estimation and singer identification tasks. Thus, W-PST method can be considered as a new framework for recognizing singing voice in polyphonic music. The underlying idea of this framework is that humans do not segregate the target sound when listening to it. Unlike the typical framework for polyphonic sound recognition, which first segregates the target and then recognizes it, our framework recognizes a singing voice directly from a polyphonic sound mixture. This policy takes advantage of the characteristics of noise to improve the performance, which is generally discarded in conventional methods.

The W-PST method has a commonality with the parallel model combination (PMC) technique [12] used in research of automatic speech recognition, which decompose noisy speech hidden Markov models (HMMs) from clean speech and noise HMMs. The advantage of our method is that it estimates the S/N ratio of a vocal and noise at every frame, while the PMC technique is used to decompose HMMs on a fixed S/N ratio beforehand.

Our goal is to develop an automatic lyrics recognition system. To achieve this goal, we will expand our method to achieve higher performance. For example, we plan to introduce a 3-D template that consists of not one frame but a number of consecutive frames to represent the dynamic features of a singing voice. This idea is promising because many researchers have proven the importance of the delta features for speech recognition.

7. REFERENCES

[1] H. Fujihara, M. Goto, J. Ogata, K. Komatani, T. Ogata, and H. G. Okuno, "Automatic synchronization between lyrics and music CD recordings based on viterbi alignment of segregated vocal signals," in *Proc. ISM*, 2006, pp. 257–264.

[2] 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*, 2005, pp. 329–336.

[3] ——, "F0 estimation method for singing voice in polyphonic audio signal based on statistical vocal model and Viterbi search," in *Proc. ICASSP*, 2006, pp. 253–256.

[4] M. Gruhne, K. Schmidt, and C. Dittmar, "Phoneme recognition in popular music," in *Proc. ISMIR*, 2007, pp. 369–370.

[5] K. Chen, S. Gao, Y. Zhu, and Q. Sun, "Popular song and lyrics synchronization and its application to music information retrieval," in *Proc. MMCN'06*, 2006.

[6] D. Iskandar, Y. Wang, M.-Y. Kan, and H. Li, "Syllabic level automatic synchronization of music signals and text lyrics," in *Proc. ACM Multimedia*, 2006, pp. 659–662.

[7] C. H. Wong, W. M. Szeto, and K. H. Wong, "Automatic lyrics alignment for cantonese popular music," *Multimedia Syst.*, vol. 4-5, no. 12, pp. 307–323, 2007.

[8] M.-Y. Kan, Y. Wang, D. Iskandar, T. L. Nwe, and A. Shenoy, "Lyrically: Automatic synchronization of textual lyrics to acoustic music signals," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 16, no. 2, pp. 338–349, 2008.

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

[10] L. Xu, M. I. Jordan, and G. E. Hinton, "An alternative model for mixtures of experts," *Adv. Neural Inf. Process. Syst. 7*, pp. 633–640, 1994.

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

[12] M. J. F. Gales and S. Yound, "An improved approach to the hidden Markov model decomposition of speech and noise," in *Proc. ICASSP*, 1997, pp. 835–838.