AUTOMATIC DETECTION AND TRACKING OF TARGET SINGER IN MULTI-SINGER MUSIC RECORDINGS

Wei-Ho Tsai and *Hsin-Min Wang*
Institute of Information Science, Academia Sinica, Taiwan, Republic of China
{wesley,whm}@iis.sinica.edu.tw

ABSTRACT

In this paper, we investigate the problem of automatically detecting and tracking a specified person's singing potions within a music recording with multiple simultaneous or non-simultaneous singers. This problem reflects an important issue in multimedia applications which require transcription and indexing of music data in response to the increasing demand nowadays for content-based information retrieval. The major challenges of this study arise from the fact that the singer's voices are inextricably intertwined with the signal of the background accompaniment. To determine whether or when an accompanied voice is present and from a sought singer, methods are presented for separating vocal from non-vocal regions, for extracting and modeling singers' vocal characteristics, and for distinguishing among vocal regions performed by the target singer and other simultaneous or non-simultaneous singers.

1. INTRODUCTION

With the rapid proliferation of popular music on the Internet, the need for effectively and efficiently managing the burgeoning amount of music materials available digitally everywhere is gaining attention. Of particular interest is the problem of automatically extracting information from music in order to lessen or replace human efforts in documenting music data. Much research has been done recently on automatic melody extraction [1], instrument recognition [2], music score transcription [3], and so on. More recently, advances in this research area have made a foray into the extraction of singing information from music, such as lyric recognition [4] – decoding what is sung, or singer identification [5] – determining who is singing. In keeping with this research target, this study addresses two further problems resembling the singer identification, Target Singer Detection (TSD) and Target Singer Tracking (TST).

The TSD aims to decide whether or not a specified target singer is present in a music recording. This task can be viewed as a binary classification, in which one class corresponds to the music data containing the target singer's voices, and another to the music data entirely performed by some singers other than the target one. In our context, only prior information about the target singer's voices is assumed available from his/her solo albums or previous recordings, while no information about the vocal characteristics is available offline from any specific non-target singers. On the other hand, the aim of the TST is to determine where in a music recording, if at all, the target singer is singing. This task can be viewed as a TSD performed as a function of time, and a system built for this task must output a list of regions where singing from the sought person has been located. Notice that the music in question may be instrumental only, solo, duet, or even chorus. However, our efforts in this work are only made to investigate the TSD/TST on solo and duet music data, because they are the most prevalent types in pop music.

There are numerous potential applications that TSD/TST could be able to create. For instance, they can serve as a tool for labeling unlabeled or insufficiently well labeled music collections. Since most of currently documented music data is only labeled by artist or lead singer, a music archive system may require an automatic technique for ascertaining the title and identifying those songs or parts of a song not sung by the lead singer. For unlabeled music data like live concert recordings, TSD/TST could be used to quickly locate a given singer's singing portions or cameo's appearances. In addition, TSD/TST may also enable music companies to rapidly scan suspect websites for piracy. Furthermore, TSD/TST may be of great use for karaoke services to manage their customers' recordings and provide personalization features.

2. METHOD OVERVIEW

In attempts to probe the singers involved in a music recording, it is necessary to extract, analyze, and compare the characteristic features of the singer's voices without interference from non-singer features. As a first step toward this end, we present a method for segmenting a music recording into vocal and non-vocal regions, in which a vocal region consists of concurrent singing and accompaniment, whereas non-vocal regions consist of accompaniment only. Next, a stochastic modeling method is presented for distilling the singers' vocal characteristics from the vocal regions. Then, the decision of whether or where a test music recording contains the target singer's voices is made by examining the extent of how well the target singer's model matches the test recording.

In addition to handling the singers' vocal characteristics, an inevitable problem in dealing with multi-singer music data, particularly in the TST task, is that multiple singers may perform simultaneously, which results in ambiguity of singer attribute. We refer to a music segment with multiple simultaneous singers as *Overlapping Singing* (OS). From the standpoints of information retrieval, an OS segment must be treated as relevant if it contains an inquired singer's voices. However, many OS segments are likely to be discarded, because they are usually corrupted severely in terms of the quality of the target singer's voices and thus tend to be identified as non-target. To alleviate this problem, an automatic OS detection method is developed, which serves as a pre-processor of the TST. After locating the OS, the TST can be done by dynamically modifying its decision with a preference to hypothesize OS segments as target.

3. VOCAL/NON-VOCAL SEGMENTATION

The basic strategy applied here is to construct a stochastic classifier for distinguishing vocal from non-vocal regions. This classifier consists of a front-end signal processor that converts digital waveforms to cepstrum-based feature vectors, followed by a backend statistical processor that performs modeling and matching. It operates in two phases, training and testing. In the training phase, a music database with manual vocal/non-vocal transcriptions is used to create a set of parametric models for characterizing the vocal and non-vocal classes. The parametric models used here are Gaussian mixture models (GMMs). There are three GMMs created. The first GMM, denoted as λ_T, is formed using the labeled vocal regions sung by a specified target singer. The second GMM, denoted as λ_V, and the third GMM, denoted as λ_N, are, respectively, trained using the labeled vocal regions and non-vocal regions of all the music data available. Parameters of the GMMs are initialized via k-means clustering and iteratively adjusted via expectation-maximization (EM) [7]. During testing, the classifier takes as input the T-length feature vectors $\mathbf{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_T\}$ extracted from an unknown recording, and produces as outputs the frame likelihoods $p(\boldsymbol{x}_t|\lambda_T)$, $p(\boldsymbol{x}_t|\lambda_V)$ and $p(\boldsymbol{x}_t|\lambda_N)$, $1\leq t \leq T$. Since singing tends to be continuous, classification can be made in a segment-by-segment manner. A W-length segment is hypothesized as vocal or not using

$$\max\left(\frac{1}{W}\sum_{i=1}^{W}\log p(\boldsymbol{x}_{kW+i} \mid \lambda_T), \frac{1}{W}\sum_{i=1}^{W}\log p(\boldsymbol{x}_{kW+i} \mid \lambda_V)\right) - \frac{1}{W}\sum_{i=1}^{W}\log p(\boldsymbol{x}_{kW+i} \mid \lambda_N) \underset{\text{non - vocal}}{\overset{\text{vocal}}{\gtrless}} \eta_{\mathrm{V}}, \quad (1)$$

where η_{V} is the threshold, and k the segment index.

4. SINGER CHARACTERISTIC MODELING

Our method for modeling the singers' voice characteristics follows the work of [6,8], in which non-vocal music segments are exploited as a prior knowledge of background signal to assist the estimation of pristine vocal signal. Suppose that an accompanied voice $\mathbf{V} = \{\boldsymbol{v}_1, \boldsymbol{v}_2, ..., \boldsymbol{v}_T\}$ is a mixture of a singing voice $\mathbf{S} = \{\boldsymbol{s}_1, \boldsymbol{s}_2, ..., \boldsymbol{s}_T\}$ and a background music $\mathbf{B} = \{\boldsymbol{b}_1, \boldsymbol{b}_2, ..., \boldsymbol{b}_T\}$. Both $\mathbf{S}$ and $\mathbf{B}$ are unobservable, but $\mathbf{B}$'s stochastic characteristics can be estimated from the non-vocal segments, since in most pop music, substantial similarities exist between the accompaniments of singing regions and instrumental-only regions. Therefore, it is sufficient to build a stochastic model λ_s for the singing voice $\mathbf{S}$, based on the available information from $\mathbf{V}$ and $\mathbf{B}$. Toward this end, we further assume that $\mathbf{S}$ and $\mathbf{B}$ are, respectively, drawn randomly and independently according to GMMs $\lambda_s = \{w_{s,i}, \boldsymbol{\mu}_{s,i}, \boldsymbol{\Sigma}_{s,i} \mid 1 \leq i \leq I\}$, and $\lambda_b = \{w_{b,j}, \boldsymbol{\mu}_{b,j}, \boldsymbol{\Sigma}_{b,j} \mid 1 \leq j \leq J\}$, where $w_{s,i}$ and $w_{b,j}$ are mixture weights, $\boldsymbol{\mu}_{s,i}$ and $\boldsymbol{\mu}_{b,j}$ mean vectors, and $\boldsymbol{\Sigma}_{s,i}$ and $\boldsymbol{\Sigma}_{b,j}$ covariance matrices. If the signal $\mathbf{V}$ is formed from a generative function $\boldsymbol{v}_t = f(\boldsymbol{s}_t, \boldsymbol{b}_t)$, $1 \leq t \leq T$, the probability of $\mathbf{V}$, given λ_s and λ_b can be represented by

$$p(\mathbf{V} \mid \lambda_s, \lambda_b) = \prod_{t=1}^{T}\left\{\sum_{i=1}^{I}\sum_{j=1}^{J} w_{s,i} w_{b,j} p(\boldsymbol{v}_t \mid \boldsymbol{\mu}_{s,i}, \boldsymbol{\Sigma}_{s,i}, \boldsymbol{\mu}_{b,j}, \boldsymbol{\Sigma}_{b,j})\right\}, \quad (2)$$

where

$$p(\boldsymbol{v}_t \mid \boldsymbol{\mu}_{s,i}, \boldsymbol{\Sigma}_{s,i}, \boldsymbol{\mu}_{b,j}, \boldsymbol{\Sigma}_{b,j}) = \iint_{\boldsymbol{v}_t = f(\boldsymbol{s}_t, \boldsymbol{b}_t)} \mathcal{N}(\boldsymbol{s}; \boldsymbol{\mu}_{s,i}, \boldsymbol{\Sigma}_{s,i}) \mathcal{N}(\boldsymbol{b}; \boldsymbol{\mu}_{b,j}, \boldsymbol{\Sigma}_{b,j}) d\boldsymbol{s} d\boldsymbol{b}. \quad (3)$$

To build λ_s, a maximum-likelihood estimation can be made as

$$\lambda_s^* = \arg\max_{\lambda_s} p(\mathbf{V} \mid \lambda_s, \lambda_b). \quad (4)$$

Using the EM algorithm, a new model $\hat{\lambda}_s$ is iteratively estimated by maximizing the auxiliary function

$$Q(\lambda_s, \hat{\lambda}_s) = \sum_{t=1}^{T}\sum_{i=1}^{I}\sum_{j=1}^{J} p(i, j \mid \boldsymbol{v}_t, \lambda_s, \lambda_b) \log p(i, j, \boldsymbol{v}_t \mid \hat{\lambda}_s, \lambda_b), \quad (5)$$

where

$$p(i, j, \boldsymbol{v}_t \mid \hat{\lambda}_s, \lambda_b) = \hat{w}_{s,i} w_{b,j} p(\boldsymbol{v}_t \mid \hat{\boldsymbol{\mu}}_{s,i}, \hat{\boldsymbol{\Sigma}}_{s,i}, \boldsymbol{\mu}_{b,j}, \boldsymbol{\Sigma}_{b,j}), \quad (6)$$

and

$$p(i, j \mid \boldsymbol{v}_t, \lambda_s, \lambda_b) = \frac{w_{s,i} w_{b,j} p(\boldsymbol{v}_t \mid \boldsymbol{\mu}_{s,i}, \boldsymbol{\Sigma}_{s,i}, \boldsymbol{\mu}_{b,j}, \boldsymbol{\Sigma}_{b,j})}{\sum_{m=1}^{I}\sum_{n=1}^{J} w_{s,m} w_{b,n} p(\boldsymbol{v}_t \mid \boldsymbol{\mu}_{s,m}, \boldsymbol{\Sigma}_{s,m}, \boldsymbol{\mu}_{b,n}, \boldsymbol{\Sigma}_{b,n})}. \quad (7)$$

Letting $\nabla Q(\lambda_s, \hat{\lambda}_s) = 0$ with respect to each parameter to be re-estimated, we have

$$\hat{w}_{s,i} = \frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{J} p(i, j \mid \boldsymbol{v}_t, \lambda_s, \lambda_b), \quad (8)$$

$$\hat{\boldsymbol{\mu}}_{s,i} = \frac{\sum_{t=1}^{T}\sum_{j=1}^{N} p(i, j \mid \boldsymbol{v}_t, \lambda_s, \lambda_b) E\left\{\boldsymbol{s}_t \mid \boldsymbol{v}_t, \boldsymbol{\mu}_{s,i}, \boldsymbol{\Sigma}_{s,i}, \boldsymbol{\mu}_{b,j}, \boldsymbol{\Sigma}_{b,j}\right\}}{\sum_{t=1}^{T}\sum_{j=1}^{N} p(i, j \mid \boldsymbol{v}_t, \lambda_s, \lambda_b)}, \quad (9)$$

$$\hat{\boldsymbol{\Sigma}}_{s,i} = \frac{\sum_{t=1}^{T}\sum_{j=1}^{J} p(i, j \mid \boldsymbol{v}_t, \lambda_s, \lambda_b) E\left\{\boldsymbol{s}_t \boldsymbol{s}_t' \mid \boldsymbol{v}_t, \boldsymbol{\mu}_{s,i}, \boldsymbol{\Sigma}_{s,i}, \boldsymbol{\mu}_{b,j}, \boldsymbol{\Sigma}_{b,j}\right\}}{\sum_{t=1}^{T}\sum_{j=1}^{J} p(i, j \mid \boldsymbol{v}_t, \lambda_s, \lambda_b)} - \hat{\boldsymbol{\mu}}_{s,i} \hat{\boldsymbol{\mu}}_{s,i}', \quad (10)$$

where prime denotes vector transpose, and $E\{\cdot\}$ expectation. The details of Eqs. (8)-(10) can be found in [6,8]. Note that if the number of mixtures in λ_b is zero, then the method above degenerates to directly modeling the accompanied voices as a GMM. This serves as a baseline for performance comparison.

5. TARGET SINGER DETECTION (TSD)

A block diagram of the proposed TSD system is shown in Fig. 1. During training, music data from a training set are segmented into vocal and non-vocal regions. The resulting non-vocal regions are then used to form a GMM which simulates the characteristics of the background accompaniments. The background music GMM together with the segmented vocal regions are then used to create two singing voice models, the target singer model λ_s^T and the universal singer model λ_s^U. The target singer model is trained using the music recordings fully performed by the target singer, while the universal singer model is trained using all the available music data not performed by the target singer. In the testing phase, a background music GMM λ_b is created on-line using the segmented non-vocal regions of a test recording $\mathbf{X}$. The system then hypothesizes whether or not the target singer is present in $\mathbf{X}$ using

$$\log p(\mathbf{X}_V \mid \lambda_s^T, \lambda_b) - \log p(\mathbf{X}_V \mid \lambda_s^U, \lambda_b) \underset{\text{non - target}}{\overset{\text{target singer}}{\gtrless}} \eta_{\mathrm{TSD}}, \quad (11)$$

where $\mathbf{X}_V$ denotes all the segmented vocal regions in $\mathbf{X}$, and η_{TSD} is the threshold.

6. TARGET SINGER TRACKING (TST)

The TST can be intuitively performed in a similar manner to the TSD. Given a test recording, the system hypothesizes whether or not a detected vocal segment is from the target singer by a comparison of the likelihoods for the target singer model and for the universal singer model. However, due to the existence of Overlapping Singing (OS), it is found that considerable amounts of the target singer's voices with background vocals tend to poorly match the target singer model and thus be improperly judged as non-target. To alleviate this problem, we lower down the threshold of hypothesizing a vocal segment as target when this segment is marked as OS. Specifically, a *W*-length vocal segment is hypothesized as target or not using

$$\frac{1}{W}\left(\sum_{i=1}^{W}\log p(\boldsymbol{x}_{kW+i}\,|\,\lambda_s^T)-\sum_{i=1}^{W}\log p(\boldsymbol{x}_{kW+i}\,|\,\lambda_s^U)\right)\underset{\text{non - target}}{\overset{\text{target singer}}{\gtrless}}\eta_{\text{TST}}-\delta_k\theta, \quad (12)$$

and

$$\delta_k=\begin{cases}1, & \text{if the segment } k \text{ is marked as overlapping - singing}\\ 0, & \text{otherwise}\end{cases},$$

where θ is a positive constant, and η_{TST} the global threshold.

To Realize the TST method above, an automatic technique for detecting the OS is proposed. Our basic strategy is to treat the OS as the third class other than the target singer and universal singer classes. It is assumed that the generic acoustic characteristics of the OS can be statistically modeled using large amounts of music data with simultaneous singers. During training, an OS model λ_s^O is created through the same training method as that of the target and universal singer models. During a test, the system hypothesizes each of the segmented vocal regions as OS or not using

$$\max\left(\frac{1}{W}\sum_{i=1}^{W}\log p(\boldsymbol{x}_{kW+i}\,|\,\lambda_s^T,\lambda_b),\frac{1}{W}\sum_{i=1}^{W}\log p(\boldsymbol{x}_{kW+i}\,|\,\lambda_s^U,\lambda_b)\right) \quad (13)$$

$$-\frac{1}{W}\sum_{i=1}^{W}\log p(\boldsymbol{x}_{kW+i}\,|\,\lambda_s^O,\lambda_b)\underset{\text{non - OS}}{\overset{\text{OS}}{\lesseqgtr}}\eta_{\text{OSD}},$$

where η_{OSD} is the threshold. Implicit in Eq. (13) is the preference that a detected OS segment contains the target singer's voices.

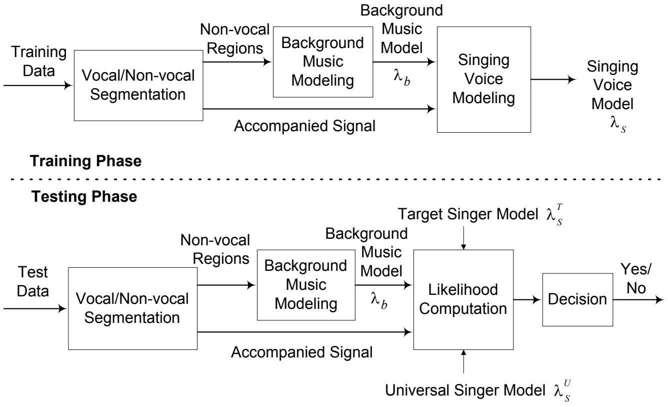

Figure 1: The proposed target singer detection system.

7. EXPERIMENTAL RESULTS

Music data used in this study consisted of 242 solo tracks, 34 duet tracks and 174 instrumental-only tracks from Mandarin pop music CDs. All the tracks were manually labeled with singer identity and the vocal/non-vocal boundaries. The 242 solo tracks were grouped into two subsets by singer, respectively, denoted as DB-S-1 and DB-S-2. The DB-S-1 comprised 200 tracks performed by 10 male and 10 female singers, with 10 distinct songs per singer. The DB-S-2 comprised the remaining 42 tracks, involving 13 female and 8 male singers, none of whom appeared in DB-S-1, and each of the singers performed two distinct songs. Furthermore, we divided DB-S-1 into two sub-subsets, one for training the singer-specific models, and another for evaluation. The sub-subset for training, denoted as DB-S-1-T, contained eight tracks per singer, while the sub-subset for evaluation, denoted as DB-S-1-E, contained two tracks per singer. The music data in DB-S-2 were used for creating the universal singer model. On the other hand, 22 among the 34 duet tracks encompassed the vocals sung by the singers in DB-S-1, and each of the singers in DB-S-1 at least involved in one of these 22 duet tracks. We denoted these 22 duet tracks as DB-D-1, and the remaining 12 duet tracks as DB-D-2. The DB-D-1 was used for evaluation, while DB-D-2 was used for training the OS model. Additionally, the 174 instrumental-only tracks were used for training the non-vocal model. All these data were down-sampled from the CD sampling rate of 44.1 kHz to 22.05 kHz, to exclude the high frequency components beyond the range of normal singing voices. Feature vectors, each consisting of 20 Mel-scale frequency cepstral coefficients, were extracted from these data using a 32-ms Hamming-windowed frame with 10-ms shifts.

Our first experiment was conducted to test the performance of the vocal/non-vocal segmentation. The test data used here were DB-S-1-E and DB-D-1. As with binary decision, performance assessment was characterized by two error measures, Miss Error Rate (MER) and False Alarm Rate (FAR). However, in view of the limited precision with which the human ear detects vocal/non-vocal changes, all frames that occurred within 0.5 seconds of a perceived switch-point were ignored in the computation. Fig. 2 shows the vocal/non-vocal segmentation results reported using the detection error trade-off (DET) plot. Here, the number of mixtures in GMM λ_T, λ_V, and λ_N, were, respectively, 32, 32, and 64 (empirically the most accurate configuration). We found that an adequate length of analysis segment was 1.5 sec, which yielded an equal error rate (MER = FAR) of 14.6%. This served as a front-end processing result for the subsequent experiments.

Next, we examined the validity of the TSD. The test set used here included DB-S-1-E and DB-D-1. The evaluation was conducted in a leave-one-out manner, which uses each of the singers in DB-S-1 as a target one once at a time and rotating through all the singers. In addition, each of the tracks in the test set was uniformly segmented into three music clips, and the TSD was performed on a clip-by-clip basis. In DB-S-1-E there were a total of 120 test samples treated as target singer trials and 2280 test samples treated as non-target singer trials, and in DB-D-1 there were a total of 72 test samples treated as target singer trails and 1153 test samples treated as non-target singer trials. Fig. 3 shows the TSD results in terms of MER and FAR. Here, the number of mixtures used in the target singer, universal singer, and background music model were empirically determined to be 32, 32, and 8, respectively. We can see that TSD in solo music was much easier than it was in duet music. It is also clear that a better TSD performance can be obtained by explicitly exploiting prior knowledge of background music.

Lastly, performance of the TST was evaluated on DB-D-1. The target singer set and the model configurations used here were the same as those in the TSD experiments. After discarding the frames that occurred within 0.5 seconds of a labeled switch-point, there were 232746 test frames treated as target trials, 120576 test frames treated as non-target trials, and 160710 test frames treated as non-vocal trials. Among the 232746 target trials, 112198 were from the frames purely labeled as target, while 120548 were from the OS frames. Fig. 4 shows the TST results. The solid line represents the TST performance obtained without taking into account the 120548 trials from the OS frames. Compared to the dashdot line, we can see that the TST performance deteriorated significantly after the inclusion of unmarked OS frames. Further analysis of our results showed that when MER = FAR, 80483 among the 120548 trials from the OS frames were hypothesized as non-target. This result reveals that a better method for handling the OS is highly desirable. The dashed line and dotted line, respectively, represent the TST results obtained with the pre-processing of the manual OS marking and that of the automatic OS detection. The value of θ in Eq. (13) was empirically set to be 0.9. The automatic OS detection used a 32-mixture OS model and achieved an empirically best equal error rate of 34.4%. We can see that a better TST performance can be obtained by modifying the decision rule with respect to the vocals from multiple simultaneous singers, though the OS detection is far from perfect.

8. CONCLUSIONS

This study has examined the feasibility of automatic detection and location of target singer in a multi-singer music recording. We have shown that the characteristics of a singer's voices can be extracted from music via vocal segment detection followed by singing signal modeling. The determination of when and whether an accompanied voice is present and from a sought singer has been formulated and solved using maximum likelihood classification and hypothesis testing rules. Furthermore, an overlapping singing detection technique has been proposed to better handle the music with multiple simultaneous singers.

9. REFERENCES

[1] M. A. Akeroyd, B. C. J. Moore, and G. A. Moore, "Melody recognition using three types of dichotic-pitch stimulus," *J. Acoust. Soc. Am.*, vol. 110, no. 3, pp. 1498–1504, 2001.

[2] A. Eronen, "Musical instrument recognition using ICA-based transform of features and discriminatively trained HMMS," *Proc. Int. Sym. Signal Processing and Its Applications*, 2003.

[3] R. A. Medina, L. A. Smith, and D. R. Wagner, "Content-based indexing of musical scores," *Proc. Joint Conf. Digital Libraries* (*JCDL*), 2003.

[4] C. K. Wang, R. Y. Lyu, and Y. C. Chiang, "An automatic singing transcription system with multilingual singing lyric recognizer and robust melody tracker," *Proc. Euro. Conf. Speech Communication and Technology* (*Eurospeech*), 2003.

[5] Y. E. Kim and B. Whitman, "Singer identification in popular music recordings using voice coding features," *Proc. Int. Conf. Music Information Retrieval* (*ISMIR*), 2002.

[6] R. C. Rose, E. M. Hofstetter, and D. A. Reynolds, "Integrated models of signal and background with application to speaker identification in noise," *IEEE Trans. Speech Audio Process.*, 2(2): 245-257, 1994.

[7] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Royal Statist. Soc.*, 39: 1-38, 1977.

[8] W. H. Tsai, H. M. Wang, D. Rodgers, S. S. Cheng, and H. M. Yu, "Blind clustering of popular music recordings based on singer voice characteristics," *Proc. Int. Conf. Music Information Retrieval* (*ISMIR*), 2003.

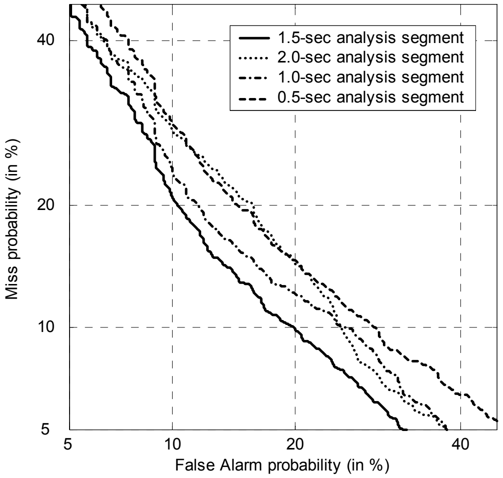

Figure 2: Performance of the vocal/non-vocal segmentation.

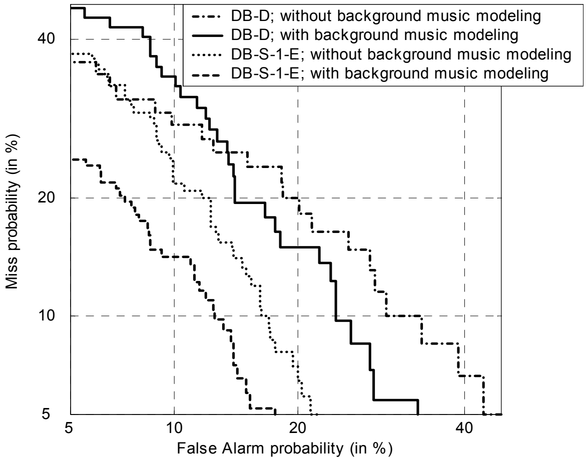

Figure 3: Performance of the target singer detection.

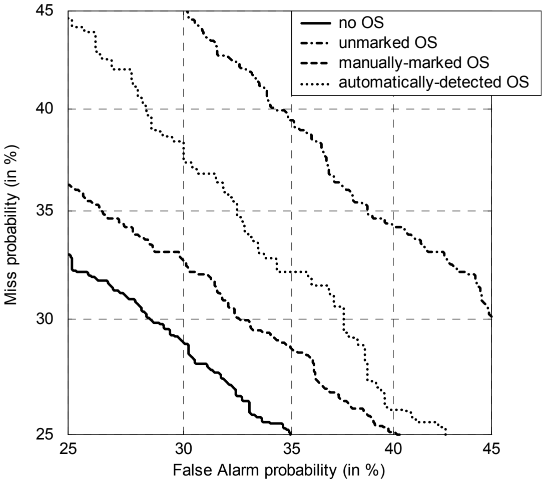

Figure 4: Performance of the target singer tracking.