# AUTOMATIC IDENTIFICATION OF SIMULTANEOUS SINGERS IN DUET RECORDINGS

**Wei-Ho Tsai**
Graduate Institute of Computer and Communication Engineering, National Taipei University of Technology, Taipei, Taiwan
whtsai@ntut.edu.tw

**Shih-Jie Liao**
Graduate Institute of Computer and Communication Engineering, National Taipei University of Technology, Taipei, Taiwan
t5418038@ntut.edu.tw

**Catherine Lai**
Open Text Corporation
Ottawa, ON, Canada
clai@opentext.com

## ABSTRACT

The problem of identifying singers in music recordings has received considerable attention with the explosive growth of the Internet and digital media. Although a number of studies on automatic singer identification from acoustic features have been reported, most systems to date, however, reliably establish the identity of singers in solo recordings only. The research presented in this paper attempts to automatically identify singers in music recordings that contain overlapping singing voices. Two approaches to overlapping singer identification are proposed and evaluated. Results obtained demonstrate the feasibility of the systems.

## 1. INTRODUCTION

In music recordings, the singing voice usually catches the listener's attention better than other musical attributes such as instrumentation or tonality. The singer's information, therefore, is essential to people for organizing, browsing, and retrieving music recordings. Most people use singer's voice as a primary cue for identifying songs, and performing such a task is almost effortless. However, building a robust automatic singer identification system is a difficult problem for machine learning. One of the challenges lies in training the system to discriminate among the different sources of sounds in music recordings, which may include background vocal, instrumental accompaniment, background noise, and simultaneous, or overlapping singings.

Although studies on automatic singer identification from acoustic features have been reported, most systems to date, however, reliably establish the identity of singers from recordings of solo performances only [1][2][3]. Tsai *et al*., in [4], investigated automatic detection and tracking of multiple singers in music recordings. However, the study only considered singing by multiple singers who performed in a non-overlapping matter. Other works related to the problem of singer identification include speech overlapping [5][6] in multi-speakers environments and voice separation from music accompaniment [7][8]. The research presented here attempts to automatically identify singers in music recordings that contain both simultaneous and non-simultaneous singings. We refer to this problem as overlapping singer identification (OSID).

## 2. APPLICATIONS

OSID can be applied in a number of areas. For example, a successful OSID system can be used as an automatic tool to locate, identify, and index singers in music recordings, thus reducing, if not replacing, human documentation efforts. OSID, moreover, can be applied in the context of karaoke. A personalized Karaoke system has been developed by Hua et al. [9] to create background visual content using home video and/or photo collections. As an extension to the current Karaoke system, OSID can be used to identify and index singers in their recorded karaoke songs, and using intelligent transformation and transition effects, the singing portions can be aligned with the respective singer's home video and/or photo collections to create a seamless personalize music video. In addition, OSID can be applied in the area of copyright protection and enforcement. Content description such as singer names, among many other metadata, is fundamental to the content and rights managements of digital music. OSID can be applied to generate singer information automatically and be integrated into the existing protection technologies to enhance the current copyright protection solutions.

## 3. PROBLEM FORMULATION

Since it is difficult to consider all application scenarios in an initial development stage, we began this research by identifying the important factors that influence the effectiveness of OSID. We then defined the scope of this preliminary study. The factors we identified include the following:

i) **Multiplicity.** Depending on the number of singers, a music recording may be classified as a pure instrumental, solo, duet, trio, band, or choral performance. In general, the complexity of an OSID

problem grows as the number of singers in a music recording increases. This study focuses on vocal duets, i.e., the OSID system determines the identity of a singer or singers who sang in a given music recording.

ii) **Overlapping duration percentage.** Although two singers perform a duet, they may not always be singing simultaneously. Therefore, an excerpt from a music recording can be a) an instrument-only segment, b) a solo-singing segment, or c) an overlapping-singing segment. To simplify the problem, the test recordings used in this study are either b) or c), with the overlapping duration percentages of 0% or 100%, respectively.

iii) **Overlapping energy ratio.** As in many bands, one or more musicians in addition to the lead singer often sing background vocal while they play their instruments. The audio signal energy of the background singer(s), therefore, may be very low compared to that of the lead singer. In such a case, identifying the background singers would be more difficult. For this preliminary study, no test recordings have background singers, and the singers in our test recordings all sing with roughly equal signal energies.

iv) **Tune/lyrics variations.** Multiple singers performing simultaneously may sing in a) exactly the same tune and lyrics, b) exactly the same tune but different lyrics, c) different tunes but exactly the same lyrics, or d) different tunes and different lyrics. We consider only cases a) and d) for this study.

v) **Background accompaniment.** A majority of popular music contains background accompaniment that inextricably intertwines singers' voice signals with a loud, non-stationary background music signal. During this initial stage of the development, we do not attempt to solve the problem of background interference but only deal with vocal music that has no background accompaniments.

vi) **Open-set/close-set.** The OSID problem at hand is a close-set classification problem, which identifies the singer(s) among a set of candidate singers. This study does not discuss the problem of open-set classification, which determines whether the singer(s) identified is/are among the candidate singers performed in a set of test recordings.

vii) **Audio quality.** Although most test recordings are taken from high quality sources such as CDs, many often undergo signal degradation due to audio filtering, encoding/decoding, or noise corruption. A successful OSID system should be robust against various signal distortions. This study, however, places this issue aside because audio quality is an inevitable problem for most music classification research. The music data we use for this study are all of high-quality audio recorded on PC.

## 4. DATA

Since no public corpus of music recordings meets the specific constraints of the OSID problem defined here, a small database of test recordings was created. The database contains vocal recordings by ten male amateur singers, aged between 20 and 35. Every singer was asked to perform 30 passages of Mandarin pop songs with a Karaoke machine. The duration of each passage ranges from 13 to 20 seconds.

The database was divided into two subsets, one for training the OSID system and the other for evaluating the system. The training subset consists of the first 15 passages, while the evaluation subset consists of the remaining 15 passages. Passages of the pop songs were recorded in a quiet room. The Karaoke accompaniments were output to a headset, and thus not recorded. All the passages were recorded at 22.05 kHz, 16 bits, in mono PCM wave.

Test recorindgs of duets were then obtained by mixing the wave files sung by a pair of singers. Two sets of recordings (i.e., for training and evaluation), sung by 45 ($C_2^{10}$=45) different pairs of singers, were created. One set included 675 ($C_2^{10} \times 15$ =675) recordings of duets sung in exactly the same tune and with the same lyrics; the other set included 4,725 ($C_2^{10} \times C_2^{15}$ =4,725) recordings of duets sung in different tunes and with different lyrics.

To facilitate the discussions in the following sections, thereafter, recordings of duets sung in exactly the same tune and with the same lyrics are referred to as "STSL duet recordings." Similarly, recordings of duets sung in different tunes and with different lyrics are referred to as "DTDL duet recordings."

## 5. METHODOLOGY

Following the most popular paradigm of stochastic pattern recognition, we propose two approaches to OSID system design.

### 5.1. Two-Stage OSID System

Figure 1 shows the block diagram of our two-stage OSID system. The system first consists of a "Solo/Duet Recognition" component. If a solo singing is recognized, the problem becomes the conventional single singer identification. If a duet singing is recognized, a "Duet Singer Identification" component handles the case. Each of the components in this system is a Gaussian mixture classifier [9].

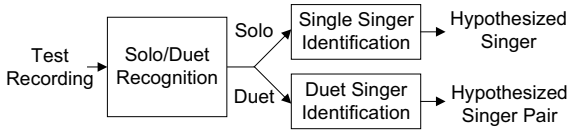


**Figure 1.** Two-stage OSID system.

*5.1.1. Solo/Duet Recognition*

Figure 2 shows the block diagram of the "Solo/Duet Recognition" component. The component is divided into two phases: training and testing. During the training phase, two Gaussian Mixture Models (GMMs), $\lambda^s$ and $\lambda^d$, are created, where $\lambda^s$ represents the acoustic pattern of a solo singing passage while $\lambda^d$ represents the acoustic pattern of a duet singing passage. Combinations of Gaussian densities generate a variety of acoustic classes, which, in turn, reflect certain vocal tract configurations. The GMMs provide good approximations of arbitrarily shaped densities of spectrum over a long span of time [9]. Parameters of a GMM consist of means, covariances, and mixture weights. $\lambda^s$ is generated from all solo singing passages and $\lambda^d$ is generated from all duet singing passages. Then prior to Gaussian mixture modeling, singing waveforms are converted into Mel-scale frequency cepstral coefficients (MFCCs). In the testing phase, an unknown test recording is converted into MFCCs and then tested for $\lambda^s$ and $\lambda^d$. The results are based on likelihood probabilities, $\Pr(\mathbf{X}|\lambda^s)$ and $\Pr(\mathbf{X}|\lambda^d)$, where the recording is hypothesized as a duet singing passage (or a solo singing passage) if $\log\Pr(\mathbf{X}|\lambda^d)$–$\log\Pr(\mathbf{X}|\lambda^s)$ is larger (or smaller) than a pre-set threshold $\eta$.

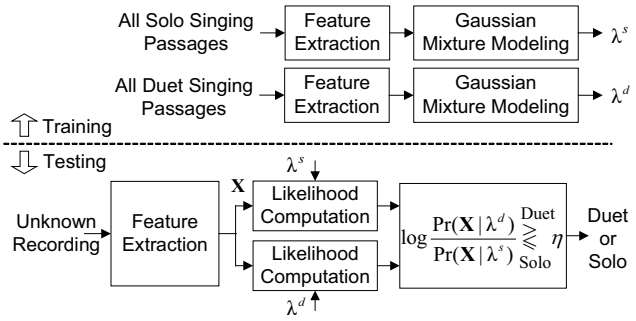


**Figure 2.** Solo/Duet recognition.

*5.1.2. Single Singer Identification*

Figure 3 shows the "Single Singer Identification" component. If there are $N$ different candidates singers, then $N$ GMMs, $\lambda_1, \lambda_2, \ldots, \lambda_N$, are created to represent the acoustic patterns of their singings. When an unknown recording is received at the system input, the component calculates and decides in favor of singer $I^*$ when the condition in Eq. (1) is satisfied:

$$I^* = \arg\max_{1 \le i \le N} \Pr(\mathbf{X}|\lambda_i) \quad (1)$$

*5.1.3. Duet Singer Identification*

The "Duet Singer Identification" component is similar to the "Single Singer Identification" component. The only difference between the two is that the GMMs of solo singers are replaced with the GMMs of pairs of singers. However, generating the GMMs of pairs of singers is not as straightforward as generating the GMMs of solo singers, because it may be impractical to collect singing data from every possible combination of pairs of singers. Hence, two approaches were taken to sidestep the collection of real simultaneous singing data. The first approach uses direct waveform mixing, which is shown in Figure 4.

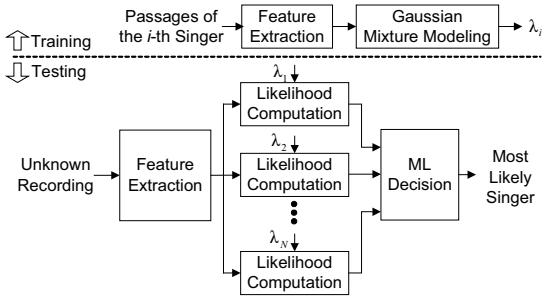


**Figure 3.** The "Single Singer Identification" component.

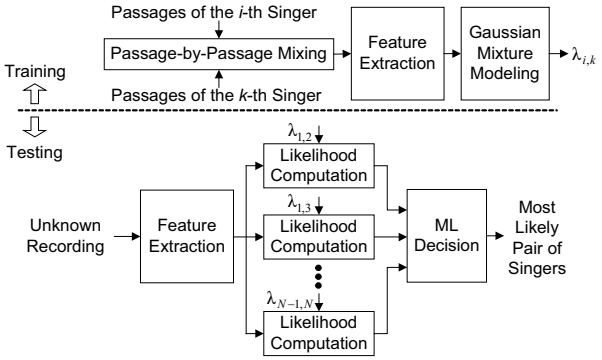


**Figure 4.** The "Duet Singer Identification" component using direct waveform mixing.

In the training phase of this system, audio waveforms from every pairs of singers are mixed, based on roughly equal energies, to simulate real duet singings. The resulting waveforms are then converted into MFCCs. For each pair of singers, a GMM is built using these features. Hence, for a population of $N$ candidate singers, a total of $C_2^N = N! / [2!(N-2)!]$ singer-pair GMMs $\lambda_{i,j}$, $i \neq j$, $1 \le i, j \le N$ are created. In the testing phase, an unknown audio recording is converted into MFCCs and then tested for each of the $C_2^N$ GMMs. The system then determines the most-likely singer pair ($I^*$, $J^*$) performed in the recording based on the maximum likelihood decision rule:

$$(I^*, J^*) = \arg\max_{1 \le i,j \le N, i \neq j} \Pr(\mathbf{X}|\lambda_{i,j}). \quad (2)$$

One shortcoming of the direct waveform mixing approach is that the training process can become very cumbersome if the number of candidate singers is large or

if a new singer needs to be added. As an alternative to this problem, a second approach based on Parallel Model Combination (PMC) technique [10] is used, as shown in Figure 5. Given a set of $N$ solo singer GMMs, each GMM is used to generate $C_2^N$ singer-pair GMMs. Since duet singing signals overlap in the time/frequency domain while the GMMs are in the cepstral/quefrency domain, the parameters of the GMMs need to be converted to the linear spectral/frequency domain before they can be added.

In addition, since two $K$-mixture GMMs would result in a large $K \times K$-mixture GMM, UBM-MAP [11] is used to control the size of the resulting GMM $K$-mixture. The basic strategy of UBM-MAP is to generate a universal GMM using all solo singers' data, and then adapt the universal GMM to each solo singer GMM based on maximum a posterior (MAP) estimation. Since all of the solo singer GMMs are adapted from the universal GMM, the mixtures of the GMMs are aligned. Thus, we do not need to consider the combination of the $k$-th Gaussian of one GMM with the $\ell$-th Gaussian of another GMM, where $k \neq \ell$, but we only need to consider the case when $k = \ell$. For a pair of singers $i$ and $j$, the combined mean and covariance of the $k$-th mixture is computed by

$$\boldsymbol{\mu}_{i,j}^{k} = \mathbf{D}\left\{\log\{\exp(\mathbf{D}^{-1}\boldsymbol{\mu}_i^{k}) + \exp(\mathbf{D}^{-1}\boldsymbol{\mu}_j^{k})\}\right\}, \quad (3)$$

$$\boldsymbol{\Sigma}_{i,j}^{k} = \mathbf{D}\left\{\log\{\exp[\mathbf{D}^{-1}\boldsymbol{\Sigma}_i^{k}(\mathbf{D}^{-1})'] + \exp[\mathbf{D}^{-1}\boldsymbol{\Sigma}_j^{k}(\mathbf{D}^{-1})']\}\right\}, \quad (4)$$

where $\boldsymbol{\mu}_i^{k}$ and $\boldsymbol{\Sigma}_i^{k}$ are the mean vector and covariance matrix of GMMs $\lambda_i$, respectively; $\mathbf{D}$ represents the discrete cosine transform matrix; prime ($'$) denotes the transpose.

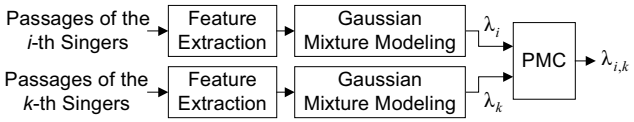


**Figure 5.** The training phase of "Duet Singer Identification" component based on parallel model combination.

### 5.2. Single-Stage OSID System

As an alternative approach, we present a second system that combines the three components in the Two-Stage OSID system. The system unifies the three components into a Single-Stage system, eliminating the stage of first determining if a test recording is a solo or duet singing performance. This is done by using the $N$ single-singer GMMs from the Single-Singer Identification and the $C_2^N$ singer-pair GMMs from the Duet-Singer Identification to build a unified classifier with $(N + C_2^N)$ GMMs. In the testing phase, an unknown recording is converted into MFCCs and then tested for each of the $(N + C_2^N)$ GMMs. Then, if we denote each single-singer GMM $\lambda_i$ as $\lambda_{i,i}$, $1 \leq i \leq N$, the system should decide in favor of singer(s) ($I^*$, $J^*$) if the condition in Eq. (5) is satisfied.

$$(I^*, J^*) = \arg\max_{1 \leq i,j \leq N} \Pr(\mathbf{X} \mid \lambda_{i,j}), \quad (5)$$

Note that if $I^* = J^*$, then the recording is hypothesized to be performed by a single singer $I^*$.

## 6. EXPERIMENTS AND RESULTS

### 6.1. Solo/Duet Recognition Experiments

The first experiment conducted examined the validity of the solo/duet recognition component. There were 150 solo test recordings, 675 STSL duet test recordings, and 4,725 DTDL duet test recordings, with a total of 5,550 test recordings. The recognition accuracy was measured by

$$\frac{\#\text{Correctly} - \text{recognized Recordings}}{\#\text{Testing Recordings}} \times 100\%.$$

Table 1 shows the recognition results with respect to the different numbers of Gaussian mixtures in $\lambda^s$ and $\lambda^d$. We can see that most of the recordings were correctly recognized.

| No. of Mixtures | Accuracy |
|---|---|
| 16 | 96.1% |
| 32 | 94.2% |
| 64 | 95.2% |

(a) Recognition accuracy

| | Actual | |
|---|---|---|
| Classified | Solo | Duet |
| Solo | 99.3% | 4.6% |
| Duet | 0.7% | 95.4% |

(b) Confusion matrix of the 16-mixture case

**Table 1.** Solo/duet recognition results.

### 6.2. Single-Singer Identification Experiments

For the purpose of comparison, experiments of the conventional SID for solo recordings were also conducted. The identification accuracy was measured by

$$\frac{\#\text{Correctly} - \text{identified Recordings}}{\#\text{Testing Recordings}} \times 100\%.$$

Table 2 shows the results of singer identification in 150 recordings sung by 10 different singers. As the singer population was small, the result obtained was almost perfect.

| No. of Mixtures | SID Accuracy |
|---|---|
| 16 | 96.7% |
| 32 | 98.7% |
| 64 | 100.0% |

**Table 2.** Results of singer identification for solo recordings.

### 6.3. Duet-Singer Identification Experiments

Then the feasibility of OSID in duet recordings was examined. In these experiments, test data consisted of 675 + 4,725 duet singing wave files, i.e., no solo recordings were considered. Here the performances of the direct waveform mixing and the PMC methods of Duet-Singer Identification component were evaluated.

Depending on the context of application, the performance of OSID is evaluated differently. This study considers two types of OSID accuracy. The first one takes into account the number of *singer pairs* identified correctly. Specifically,

$$\text{Acc.1 (in \%)} = \frac{\#\text{Correctly} - \text{identified Singer Pairs}}{\#\text{Testing Recordings}} \times 100\%.$$

The second one takes into account the number of *singers* identified correctly. Specifically,

$$\text{Acc. 2 (in \%)} = \frac{\#\text{Correctly} - \text{identified Singers}}{\#\text{Testing Singers}} \times 100\%.$$

For example, if a recording contains simultaneous singings by two performers, $s_1$ and $s_2$, and the identified singers are $s_1$ and $s_4$, then `#Correctly-identified Singer Pairs` = 0 and `#Correctly-identified Singers` = 1. Consequently, Acc. 2 is always higher than Acc. 1.

Tables 3 shows the OSID result obtained with direct waveform mixing methods. Here, the OSID results for four cases are presented: i) both training and testing data consist of STSL duet recordings; ii) training data consist of STSL duet recordings, while testing data consist of DTDL duet recordings; iii) training data consist of DTDL duet recordings, while testing data consist of STSL duet recordings; iv) both training and testing data consist of DTDL duet recordings.

It can be seen from Table 3 that OSID using STSL duet recordings for training always outperformed than those that using DTDL duet recordings for training. Similarly, the performance of OSID using STSL duet recordings for testing was always better than those that using DTDL duet recordings for testing. When both training and testing data consist of STSL duet recordings, we obtained the best OSID performance, showing that 85.0% of singer pairs or 92.5% of singers in the testing data can be correctly identified.

Tables 4 shows the OSID result obtained with the PMC method. In this experiment, since no duet singing is required in the training process, we considered two cases: i) testing data consist of STSL duet recordings; ii) testing data consist of DTDL duet recordings. It can be observed in Table 4, that similar to the results in Table 3, the performance of OSID was always better when STSL duet recordings were used for testing. Comparing Table 4 with Table 3 (a) and (b), it can also be found that the direct waveform mixing method was superior to the PMC method when STSL duet recordings were used for testing. However, the PMC method performed better than the direct waveform mixing method when DTDL duet recordings were used for testing. This indicates that the PMC method is not only better at scaling up the singer population, but it is also better at generalizing the singer identification problem than the direct waveform mixing method.

| No. of Mixtures | Acc. 1 | Acc. 2 |
|---|---|---|
| 16 | 80.7% | 90.1% |
| 32 | 84.3% | 92.1% |
| 64 | 85.0% | 92.5% |

(a) Both training and testing data consist of STSL duet recordings

| No. of Mixtures | Acc. 1 | Acc. 2 |
|---|---|---|
| 16 | 67.9% | 83.7% |
| 32 | 69.8% | 84.8% |
| 64 | 73.6% | 86. 7% |

(b) Training data consist of STSL duet recordings, while testing data consist of DTDL duet recordings

| No. of Mixtures | Acc. 1 | Acc. 2 |
|---|---|---|
| 16 | 77.3% | 88.7% |
| 32 | 78.4% | 89.3% |
| 64 | 80.7% | 90.4% |

(c) Training data consist of DTDL duet recordings, while testing data are STSL duet recordings

| No. of Mixtures | Acc. 1 | Acc. 2 |
|---|---|---|
| 16 | 52.3% | 75.8% |
| 32 | 47.1% | 73.4% |
| 64 | 43.6% | 71.6% |

(d) Both training and testing data consist of DTDL duet recordings

**Table 3.** Results of identifying duet recordings based on direct waveform mixing method.

| No. of Mixtures | Acc. 1 | Acc. 2 |
|---|---|---|
| 16 | 75.1% | 87.1% |
| 32 | 75.1% | 87.3% |
| 64 | 78.1% | 88.7% |

(a) Testing data consist of STSL duet recordings

| No. of Mixtures | Acc. 1 | Acc. 2 |
|---|---|---|
| 16 | 71.1% | 85.0% |
| 32 | 69.9% | 85.0% |
| 64 | 75.3% | 87.6% |

(b) Testing data consist of DTDL duet recordings

**Table 4.** Results of identifying duet recordings based on PMC method.

### 6.4. Singer Identification Experiments: Solo and Duet Recordings

Lastly, for a more realistic case, a test recording, which may be either a solo singing or duet singing, was tested. There were 150 solo test recordings, 675 STSL duet test recordings, and 4,725 DTDL duet test recordings, with a total of 5,550 test recordings. The identification performance was characterized by Acc. 1 and Acc. 2, as before. However, if a recording contains only single singer $s_1$ but the system hypothesizes two singers $s_1$ and $s_4$, then `#Correctly-identified Singer Pairs` = 0 and `#Correctly-identified Singers` = 1.

Table 5 shows the results obtained by the proposed two OSID systems. The singer-pair GMMs used in this experiment were generated using the PMC method. The number of Gaussian mixtures was set to 64 for both solo singer and singer pair GMMs. Compared to the results in Table 4, it is observed that while more uncertainties are added in the testing data, the resulting accuracies only decrease slightly. In addition, it is also found that the Two-Stage OSID system performed better than the Single-Stage OSID system. This indicates that although the Single-Stage OSID system takes advantage of the simplicity in design, it pays the loss of accuracy that can be achieved with the Two-Stage OSID system.

| System | Acc. 1 | Acc. 2 |
|---|---|---|
| Two-Stage OSID System | 76.2% | 88.1% |
| Single-Stage OSID System | 73.0% | 87.9% |

**Table 5.** Results of identifying both solo and duet recordings.

## 7. CONCLUSION

This paper examined the feasibility of overlapping singer identification and compared two approaches to the problem of detecting and identifying singers in duet recordings. The research presented here extends previous works on singer identification. The systems proposed up-to-date only focus on the identification of singers in solo recordings. In reality, many music recordings involve multiple singers such as duet love songs, gospel music, or folk and country music. Encouraging results arrived at this initial stage of investigation laid a good foundation for the future development of a robust automatic singer identification system. Regarding future work, a wider variety of music will be acquired to scale up and further test the system.

## 8. REFERENCES


[1] Kim, Y. and Whitman, B. "Singer identification in popular music recordings using voice coding features", *Proc. ISMIR*, 2002.

[2] Fujihara, H., Kitahara, T., Goto, M., Komatani, K., Ogata, T., and Okuno, H. "Singer identi-fication based on accompaniment sound reduction and reliable frame selection", *Proc. ISMIR*, 2005.

[3] Mesaros, A., Virtanen, T., and Klapuri, A. "Singer identification in polyphonic music using vocal separation and pattern recognition methods", *Proc. ISMIR*, 2007.

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

[5] Shriberg, E., Stolcke, A., and Baron, D. "Observations on overlap: findings and implications for automatic processing of multi-party conversation", *Proc. Eurospeech*, 2001.

[6] Yamamoto, K., Asano, F., Yamada, T., and Kitawaki, N. "Detection of overlapping speech in meetings using support vector machines and support vector regression", *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, 89(8): 2158-2165, 2006.

[7] Ozerov, A., Philippe, P., Gribonval, R., and Bimbot, F. "One microphone singing voice separating using source-adapted models", *Proc. WASPAA (IEEE Workshop on Applications of Signal Processing to Audio and Acoustics)*, 2005.

[8] Lee, Y. and Wang, D. "Singing voice separation form monaural recordings", *Proc. ISMIR*, 2006.

[9] Hua, X., Lu, L., and Zhang, H. "Personalized karaoke", *Proc. ACM Multimedia*, 2004.

[10] Reynolds, D. and Rose, R. "Robust text-independent speaker identification using Gaussian mixture speaker models", *IEEE Transactions on Speech and Audio Processing*, 3(1): 72-83, 1995.

[11] Gales, M. and Young, S. "Robust continuous speech recognition using parallel model combination", *IEEE Transactions on Speech and Audio Processing*, 4(5): 352 - 359, 1996.

[12] Reynolds, D., Quatieri, T., and Dunn, R. "Speaker verification using adapted Gaussian mixture models", *Digital Signal Processing*, 10: 19-41, 2000.