Automatic Identification of the Sung Language in Popular Music Recordings

Wei-Ho Tsai[1] and *Hsin-Min Wang*[2]

[1]Department of Electronic Engineering & Graduate Institute of Computer and Communication Engineering, National Taipei University of Technology, Taipei 10608, Taiwan
Tel: +886-2-27712171 ext. 2257, Fax: +886-2-27317120
E-mail: whtsai@ntut.edu.tw

[2]Institute of Information Science, Academia Sinica, Taipei 115, Taiwan
Tel: +886-2-27883799 ext. 1714, Fax: +886-2-27824814
E-mail: whm@iis.sinica.edu.tw

Abstract

As part of the research into content-based music information retrieval (MIR), this paper presents a preliminary attempt to automatically identify the language sung in popular music recordings. It is assumed that each language has its own set of constraints that specify the sequence of basic linguistic events when lyrics are sung. Thus, the acoustic structure of individual languages may be characterized by statistically modeling those constraints. To achieve this, the proposed method employs vector clustering to convert a singing signal from its spectrum-based feature representation into a sequence of smaller basic phonological units. The dynamic characteristics of the sequence are then analyzed using bigram language models. As vector clustering is performed in an unsupervised manner, the resulting system does not need sophisticated linguistic knowledge; therefore, it is easily portable to new language sets. In addition, to eliminate interference from background music, we leverage the statistical estimation of the background musical accompaniment of a song so that the vector clustering truly reflects the solo singing voices in the accompanied signals.

1. Introduction

With the ever-increasing capabilities of data storage and transmission applications, and thanks to recent advances in various digital signal processing technologies, the production and distribution of music material has enjoyed unprecedented growth in recent years. Various types of audio formats, coupled with media players and software have revolutionized the way people listen to music. However, the rapid proliferation of musical material has made it increasingly difficult to find a piece of music from the innumerable options available. In recent years, this dilemma has motivated some researchers to develop techniques that extract information from music automatically. Specific topics, such as melody spotting (Akeroyd et al., 2001), instrument recognition (Eronen, 2003), music score transcription (Medina et al., 2003), and genre classification (Tzanetakis and Cook, 2002), have been studied extensively within the overall context of content-based music information retrieval (MIR). Meanwhile, other researchers have focused on the problem of extracting singing information from music, for example, lyric recognition (Wang et al., 2003) – to decode what is sung; and singer identification (Kim and Whitman, 2002) – to determine who is singing. In tandem with the above research, this study presents a preliminary attempt to identify the sung language of a song. Specifically, we try to determine which language among a set of candidate languages is sung in a given popular music recording.

Sung language identification (sung LID) is useful for organizing multilingual music collections that are either unlabeled or insufficiently labeled. For instance, songs with English titles, but sung in different languages, are commonplace in popular music; hence, it is usually difficult to infer the language of a song simply from its title. In such cases, sung LID can be used to categorize music recordings by language, without needing to refer to the lyrics or to additional textual information about the lyrics. This function could support preference-based searches for music and may also be useful in other techniques for classifying music, such as genre

classification. Sung LID can also be used to distinguish between songs that have the same tune, but whose lyrics are in different languages. This is often the case with cover versions of songs, where a singer performs a song written or made famous by a different artist. Since popular songs are often translated into different languages and the titles are changed accordingly, sung LID could help a melody-based MIR system handle multilingual music documents.

To date, only a few researchers have studied sung LID (e.g., Tsai and Wang, 2004; Schwenninger et al., 2006). The closest related research is spoken language identification (spoken LID) (Muthusamy, Barnard, and Cole, 1994; Navratil and Zuhlke, 1998), which tries to identify the language being spoken from a sample of speech by an unknown speaker. Spurred by market trends and the need to provide services to a wider public, spoken LID has gained in importance because it is a key step in developing automatic multilingual systems, such as multilingual speech recognition, information retrieval, and spoken language translation. Various methods (e.g., House and Neuburg, 1977; Hazen and Zue, 1997; Zissman, 1995) have attempted to mimic the ability of humans to distinguish between languages. From a linguistic standpoint, spoken languages can be distinguished from one another by the following traits.

- Phonology. Phonetic inventories differ from one language to another. Even when languages have nearly identical phones, the frequency of the occurrence of phones and the combinations of them differ significantly across languages.
- Prosody. Significant differences exist in the duration of phones, the speech rate, and the intonation across different languages.
- Vocabulary. Each language has its own lexicons; hence, the process of word formation also differs from one language to another.
- Grammar. The syntactic rules that govern the concatenation of words into spoken utterances can vary greatly from language to language.

Although humans identify the language of a speech utterance by using one or several of the

above traits, to date, spoken-LID research has not fully exploited all of these traits. Instead, it has developed methods that are reliable, computationally efficient, and easily portable to new language sets. For example, phonological information and prosodic information are the most prevalent cues exploited for spoken LID, since they can be easily extracted from an acoustic signal without requiring too much language-specific knowledge. More specifically, a feasible way to perform spoken LID is stochastic modeling of the so-called *phonotactics*, i.e., the dependencies between the phonetic elements of utterances (e.g., House and Neuburg, 1977; Nakagawa et al., 1992). A spoken-LID system based on phonotactics usually consists of a phonetic element recognizer and a set of *n-gram*-based language models. A number of approaches have adopted this paradigm (e.g., Harbeck and Ohler, 1999; Nakagawa et al., 1992; Zissman, 1995). Other combinations that use different language-discriminating information (Cummins et al., 1999; DuPreez et al., 1999; Hazen and Zue, 1997), but not complex linguistic knowledge, have also been studied in order to improve spoken-LID performance.

Intuitively, it seems reasonable that spoken LID methods could also be used for sung LID. However, porting a well-developed spoken LID technique to sung LID is not straightforward. First, singing differs from speech in many ways; for example, the various phonetic modifications employed by singers, prosodic shaping to fit the overall melody, and the peculiar syntactic structures used in lyrics. Even state-of-the-art spoken-LID systems may not be able to deal with the huge variations found in singing voices. Second, most popular songs contain background accompaniment, which inevitably causes interference when attempts are made to automatically determine the language underlying a singing signal. In particular, when using phonotactic information for sung LID, it is rather difficult to build a phone recognizer capable of handling accompanied singing signals with satisfactory accuracy and reliability. Third, spoken-LID methods based on prosodic information might fail in the sung-LID task, since the original prosodic structures of the spoken language are largely submerged by the melody in a song.

Recognizing these problems, the goal of this study is to develop a sung-LID method that is data-driven and robust against interference from background music, and does not involve the cumbersome task of phone recognition.

The remainder of this paper is organized as follows. Section 2 gives an overview of the proposed method. The components of sung-LID, namely, vocal/non-vocal segmentation, language characteristic modeling, and stochastic matching, are discussed in Sections 3, 4, and 5, respectively. Section 6 details the experiment results. Then, in Section 7, we present our conclusions.

2. Method Overview

A sung-LID system takes a test music recording as input and produces the identity of the language sung in that recording as output. Since most music is a mixture of assorted sound sources, the key to designing a successful sung-LID system is to extract, model, and compare the characteristic features of language acoustics by eliminating interference from non-language features. The proposed sung-LID process involves two phases: training and testing, as shown in Figure 1.

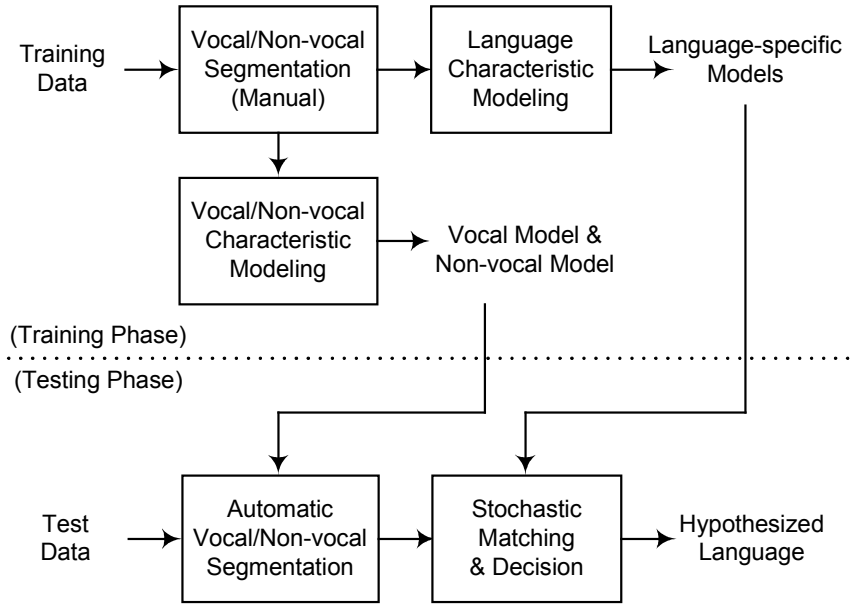


Fig. 1. Illustration of the proposed sung-LID process

In the training phase, we use a database containing a wide variety of music genres to establish the characteristic representation of each sung language of interest. Since information about the sung language is not present in the accompaniment, the training procedure begins by segmenting a music recording into vocal and non-vocal regions. Vocal regions consist of concurrent singing and accompaniment, whereas non-vocal regions consist of accompaniment only. To ensure that the training is reliable, the segmentation should be performed manually. Next, the acoustic characteristics of the vocal and non-vocal regions are stochastically modeled for use in automating the segmentation procedure in the testing phase. Then, a stochastic modeling technique is applied to extract the underlying characteristics of the sung language in the vocal segments by suppressing the background accompaniment. As a result, each language is represented by a language-specific parametric model.

During testing, the vocal and non-vocal segments of an unknown music recording are automatically located and marked accordingly. The vocal segments are then examined using each of the language-specific parametric models. Finally, the language of the model deemed the best match with the observed vocal segments is taken as the language of that test recording.

3. Vocal/Non-vocal Characteristic Modeling and Segmentation

The basic strategy applied in this study follows our previous work (Tsai and Wang, 2006), in which a stochastic classifier is constructed to distinguish between vocal and non-vocal regions. The classifier consists of 1) a front-end signal processor, which converts digital waveforms into spectrum-based feature vectors, e.g., cepstral coefficients; and 2) a backend statistical processor, which performs modeling and matching.

A set of Gaussian mixture models (GMMs) is used to model the acoustic characteristics of the vocal and non-vocal classes. For each language of interest, we construct a GMM using the feature

vectors of the manually-segmented vocal parts of the music data sung in that language. Thus, L vocal GMMs $\Lambda_1, \Lambda_2, \ldots, \Lambda_L$ are formed for L languages. We also construct a non-vocal GMM, Λ_N, using the feature vectors of all the manually-segmented non-vocal parts of the music data. The parameters of the GMMs are initialized via k-means clustering and iteratively adjusted by the expectation-maximization (EM) algorithm (Dempster et al., 1977). When an unknown music recording is tested, the classifier takes as input the T-length feature vectors $\mathbf{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_T\}$ extracted from that recording, and produces as output the frame likelihoods $p(\boldsymbol{x}_t|\Lambda_N)$ and $p(\boldsymbol{x}_t|\Lambda_l)$, $1 \le l \le L$, $1 \le t \le T$. Since singing tends to be continuous, classification can be performed in a segment-by-segment manner. Specifically, a W-length segment is hypothesized as either vocal or non-vocal using

$$\max_{1 \le l \le L} \left(\sum_{i=1}^{W} \log p(\boldsymbol{x}_{kW+i} \mid \Lambda_l) \right) \underset{\text{non - vocal}}{\overset{\text{vocal}}{\gtrless}} \sum_{i=1}^{W} \log p(\boldsymbol{x}_{kW+i} \mid \Lambda_N), \tag{1}$$

where k is the segment index.

4. Language Characteristic Modeling

This section presents a stochastic method for representing the characteristics of sung languages. It does not involve complicated linguistic rules and pre-prepared phonetic transcriptions.

4.1.Vector Tokenization Followed by Grammatical Modeling

Our basic strategy involves exploring the phonotactics-related information of individual languages by examining the statistical dependencies between sound events present in a singing signal. In contrast to the conventional phonotactic modeling approach, which relies on phone recognition as a front-end operation, we use an unsupervised classification method to derive the basic phonological units inherent in any singing process. This allows us to circumvent the cumbersome task of segmenting singing into linguistically meaningful elements.

Given a set of training data consisting of spectrum-based feature vectors computed from the vocal parts of a music recording, language characteristic modeling is performed in two stages, as shown in Figure 2. In the first stage, vector clustering is applied to all feature vectors pertaining to a particular language. This produces a language-specific codebook consisting of several codewords used to characterize the individual clusters. Each feature vector is then assigned the codeword index of its associated cluster. We assume that each cluster represents a specific vocal tract configuration corresponding to a fragment of a broad phonetic class, such as vowels, fricatives, or nasals. The concatenation of different codeword indices in a singing signal may follow some language-specific rules resembling phonotactics. If this is the case, the characteristics of the sung language can be extracted by analyzing the generated codeword index sequences.

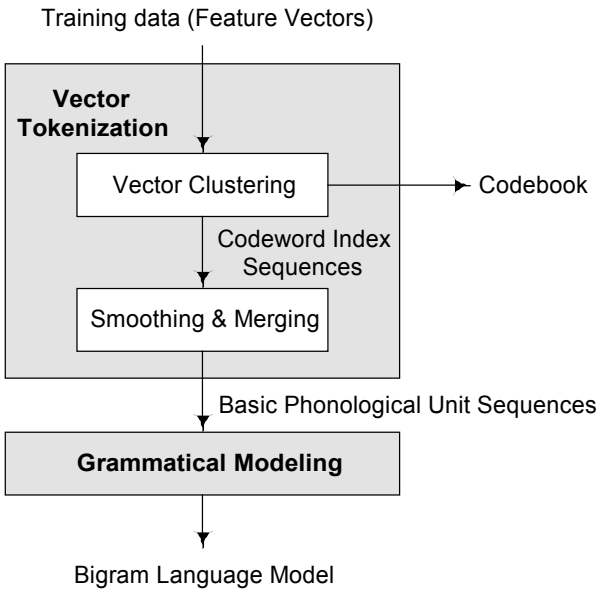


Fig. 2. Language characteristic modeling

To reflect the fact that a vocal tract's configuration does not change suddenly, the generated codeword index sequences are smoothed in the time domain. For smoothing, an index sequence is first divided into a series of consecutive, non-overlapping, fixed-length segments. Each segment is then assigned the majority index of its constituent vectors. After that, adjacent segments are

merged as a homogeneous segment if they have the same codeword index. Since each homogeneous segment is regarded as a basic phonological unit, the vocal parts of a piece of music can be tokenized into a sequence of basic phonological units.

In the second stage, a grammatical model is used to characterize the dynamics of the generated basic phonological unit sequences. There are many ways to do this. In our implementation, bigram language models (Jelinek, 1990) are used. The parameters of a bigram language model, which consist of interpolated bigram probabilities, are estimated using the following relative frequency method:

$$p(w_t = j \mid w_{t-1} = i) = \alpha \frac{n_{ij}}{\sum_{k=1}^{K} n_{ik}} + (1-\alpha) \frac{n_j}{\sum_{k=1}^{K} n_k}, \quad (2)$$

where w_t and $w_{t\text{-}1}$ denote two successive basic phonological units, α is an interpolating factor subject to $0 \le \alpha \le 1$, K is the codebook size, n_j is the number of basic phonological units assigned as codeword j, and n_{ij} denotes the number of two successive basic phonological units assigned as codewords i and j, respectively. Note that the transition between two separate vocal regions in a music recording is not considered in the computation of bigram probabilities. In summary, a language-specific model consists of a codebook and a bigram language model.

4.2.Solo Voice Codebook Generation

The effectiveness of the above language characteristic modeling technique depends on whether the vector tokenization truly reflects the phonology. Since the majority of popular songs contain background musical accompaniment during most or all vocal passages, applying conventional vector clustering methods, such as a *k*-means algorithm, directly to the accompanied singing signals may generate clusters related to both the vocal tract's configuration and the instrumental accompaniment. To address this problem, we develop a codebook generation method for vector clustering, which estimates the stochastic characteristics of the underlying solo voices from accompanied singing signals.

Let $\mathbf{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_T\}$ denote all the feature vectors computed from the vocal regions of a music recording. Because of the accompaniment, **X** can be considered as a mixture of a solo voice $\mathbf{S} = \{\boldsymbol{s}_1, \boldsymbol{s}_2, ..., \boldsymbol{s}_T\}$ and background music $\mathbf{B} = \{\boldsymbol{b}_1, \boldsymbol{b}_2, ..., \boldsymbol{b}_T\}$. More specifically, **S** and **B** are added in the time domain or linear spectrum domain, but neither of them is independently observable. Our goal is to create a codebook to represent the generic characteristics of the solo voice signal **S**, such that vector tokenization can be performed on the basis of this codebook. Under the vector clustering framework, we assume that the solo voice signal and background music are characterized, respectively, by two independent codebooks $\mathbf{C}_s = \{\boldsymbol{c}_{s,1}, \boldsymbol{c}_{s,2}, \ldots, \boldsymbol{c}_{s,K_s}\}$ and $\mathbf{C}_b = \{\boldsymbol{c}_{b,1}, \boldsymbol{c}_{b,2}, \ldots, \boldsymbol{c}_{b,K_b}\}$, where $\boldsymbol{c}_{s,i}$, $1 \le i \le K_s$ and $\boldsymbol{c}_{b,j}$, $1 \le j \le K_b$ are the codewords. To better represent the acoustic feature space, each cluster is modeled by a Gaussian density function. Therefore, a codeword consists of a mean vector and a covariance matrix, i.e., $\boldsymbol{c}_{s,i} = \{\boldsymbol{\mu}_{s,i}, \boldsymbol{\Sigma}_{s,i}\}$ and $\boldsymbol{c}_{b,j} = \{\boldsymbol{\mu}_{b,j}, \boldsymbol{\Sigma}_{b,j}\}$, where $\boldsymbol{\mu}_{s,i}$ and $\boldsymbol{\mu}_{b,j}$ are mean vectors, and $\boldsymbol{\Sigma}_{s,i}$ and $\boldsymbol{\Sigma}_{b,j}$ are covariance matrices. Vector clustering can be formulated as a problem of how to best represent **X** by choosing and combining the codewords from $\mathbf{C}_s$ and $\mathbf{C}_b$. To measure the accuracy of vector clustering, we compute the following conditional probability:

$$p(\mathbf{X} \mid \mathbf{C}_s, \mathbf{C}_b) = \prod_{t=1}^{T} \left\{ \max_{i,j} p(\boldsymbol{x}_t \mid \boldsymbol{c}_{s,i}, \boldsymbol{c}_{b,j}) \right\}, \tag{3}$$

where $p(\boldsymbol{x}_t \mid \boldsymbol{c}_{s,i}, \boldsymbol{c}_{b,j})$ denotes one possible combination of the solo voice and background music that could form an instant accompanied singing signal $\boldsymbol{x}_t$. If the accompanied signal is formed from a generative function $\boldsymbol{x}_t = f(\boldsymbol{s}_t, \boldsymbol{b}_t)$, $1 \le t \le T$, the probability $p(\boldsymbol{x}_t \mid \boldsymbol{c}_{s,i}, \boldsymbol{c}_{b,j})$ can be computed by

$$p(\boldsymbol{x}_t \mid \boldsymbol{c}_{s,i}, \boldsymbol{c}_{b,j}) = \iint_{\boldsymbol{x}_t = f(\boldsymbol{s}_t, \boldsymbol{b}_t)} G(\boldsymbol{s}; \boldsymbol{\mu}_{s,i}, \boldsymbol{\Sigma}_{s,i})\, G(\boldsymbol{b}; \boldsymbol{\mu}_{b,j}, \boldsymbol{\Sigma}_{b,j}) d\boldsymbol{s} d\boldsymbol{b}, \tag{4}$$

where $G(\cdot)$ denotes a multi-variate Gaussian density function. When using such a measurement, vector clustering is considered effective if the probability $p(\mathbf{X}|\mathbf{C}_s,\mathbf{C}_b)$ is as high as possible.

In most popular songs, substantial similarities exist between the music in the non-vocal

regions and the accompaniment of the vocal regions. Therefore, although the background music **B** is unobservable, its stochastic characteristics can be approximated from the non-vocal regions. This assumption enables us to construct the background music codebook $\mathbf{C}_b$ by applying the *k*-means clustering algorithm to the feature vectors of the non-vocal regions directly. Accordingly, based on the codebook $\mathbf{C}_b$ and the observable accompanied voice **X**, it is sufficient to derive the solo voice codebook $\mathbf{C}_s$ via a maximum likelihood estimation as follows:

$$\mathbf{C}_s^* = \arg\max_{\mathbf{C}_s} p(\mathbf{X} \mid \mathbf{C}_s, \mathbf{C}_b). \tag{5}$$

Equation (5) can be solved by the EM algorithm, which starts with an initial codebook $\mathbf{C}_s$ and iteratively estimates a new codebook $\hat{\mathbf{C}}_s$ such that $p(\mathbf{X}|\hat{\mathbf{C}}_s,\mathbf{C}_b) \geq p(\mathbf{X}|\mathbf{C}_s,\mathbf{C}_b)$. It can be shown that the need to increase the probability $p(\mathbf{X}|\hat{\mathbf{C}}_s,\mathbf{C}_b)$ can be satisfied by maximizing the auxiliary function

$$Q(\mathbf{C}_s, \hat{\mathbf{C}}_s) = \sum_{t=1}^{T}\sum_{i=1}^{K_s}\sum_{j=1}^{K_b}\delta(i=i^*, j=j^*)\log p(\boldsymbol{x}_t \mid \boldsymbol{c}_{s,i}, \boldsymbol{c}_{b,j}), \tag{6}$$

where $\delta(\cdot)$ denotes a Kronecker delta function, and

$$(i^*, j^*) = \arg\max_{i,j} p(\boldsymbol{x}_t \mid \boldsymbol{c}_{s,i}, \boldsymbol{c}_{b,j}). \tag{7}$$

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

$$\hat{\boldsymbol{\mu}}_{s,i} = \frac{\sum_{t=1}^{T}\sum_{j=1}^{N}\delta(i=i^*, j=j^*)p(\boldsymbol{x}_t \mid \boldsymbol{c}_{s,i}, \boldsymbol{c}_{b,j}) \cdot E\{\boldsymbol{s}_t \mid \boldsymbol{x}_t, \boldsymbol{c}_{s,i}, \boldsymbol{c}_{b,j}\}}{\sum_{t=1}^{T}\sum_{j=1}^{N}\delta(i=i^*, j=j^*)p(\boldsymbol{x}_t \mid \boldsymbol{c}_{s,i}, \boldsymbol{c}_{b,j})}, \tag{8}$$

$$\hat{\boldsymbol{\Sigma}}_{s,i} = \frac{\sum_{t=1}^{T}\sum_{j=1}^{J}\delta(i=i^*, j=j^*)p(\boldsymbol{x}_t \mid \boldsymbol{c}_{s,i}, \boldsymbol{c}_{b,j}) \cdot E\{\boldsymbol{s}_t\boldsymbol{s}_t' \mid \boldsymbol{x}_t, \boldsymbol{c}_{s,i}, \boldsymbol{c}_{b,j}\}}{\sum_{t=1}^{T}\sum_{j=1}^{J}\delta(i=i^*, j=j^*)p(\boldsymbol{x}_t \mid \boldsymbol{c}_{s,i}, \boldsymbol{c}_{b,j})} - \boldsymbol{\mu}_{s,i}\boldsymbol{\mu}_{s,i}', \tag{9}$$

where the prime operator ($'$) denotes the vector transpose, and $E\{\cdot\}$ denotes the expectation. Detailed derivations of Equations (8) and (9) can be found in Nadas et al. (1989), Rose et al.

(1994), and Tsai and Wang (2006). Figure 3 summarizes the procedure for generating a solo voice codebook.

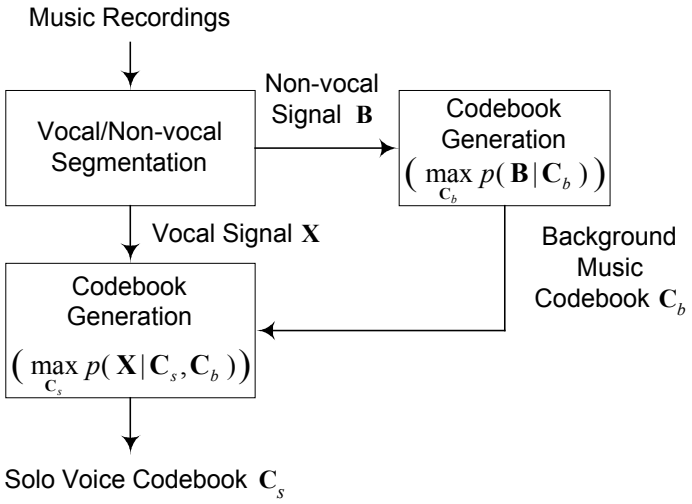


Fig. 3. Procedure for generating a solo voice codebook

5. Stochastic Matching and Decision

In the testing phase, the system determines the language sung in a music recording based on the language-specific codebooks and bigram language models. As shown in Figure 4, a test music recording is first segmented into vocal and non-vocal regions. Then, the feature vectors from the non-vocal regions are used to form a codebook $\mathbf{C}_b$, which simulates the characteristics of the background accompaniment in the vocal regions. For each candidate language L, the associated solo voice codebook $\mathbf{C}_{s,l}$, $1 \le l \le L$ and the background music codebook $\mathbf{C}_b$ are used to tokenize the feature vectors of the vocal regions $\{\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_T\}$ into a codeword index sequence $V^{(l)} = \{v_1^{(l)}, v_2^{(l)}, \ldots, v_T^{(l)}\}$, where T is the total length of the vocal regions, and $v_t^{(l)}$, $1 \le t \le T$ is determined by

$$v_t^{(l)} = \arg_i \left[\max_{i,j} p(x_t \mid \boldsymbol{c}_{s,i}^{(l)}, \boldsymbol{c}_{b,j}) \right]. \tag{10}$$

Each of the codeword index sequences $V^{(l)}$, $1 \le l \le L$ is then converted into a basic phonological unit sequence $W^{(l)} = \{w_1^{(l)}, w_2^{(l)}, \ldots, w_{N^{(l)}}^{(l)}\}$ by smoothing and merging the adjacent identical indices.

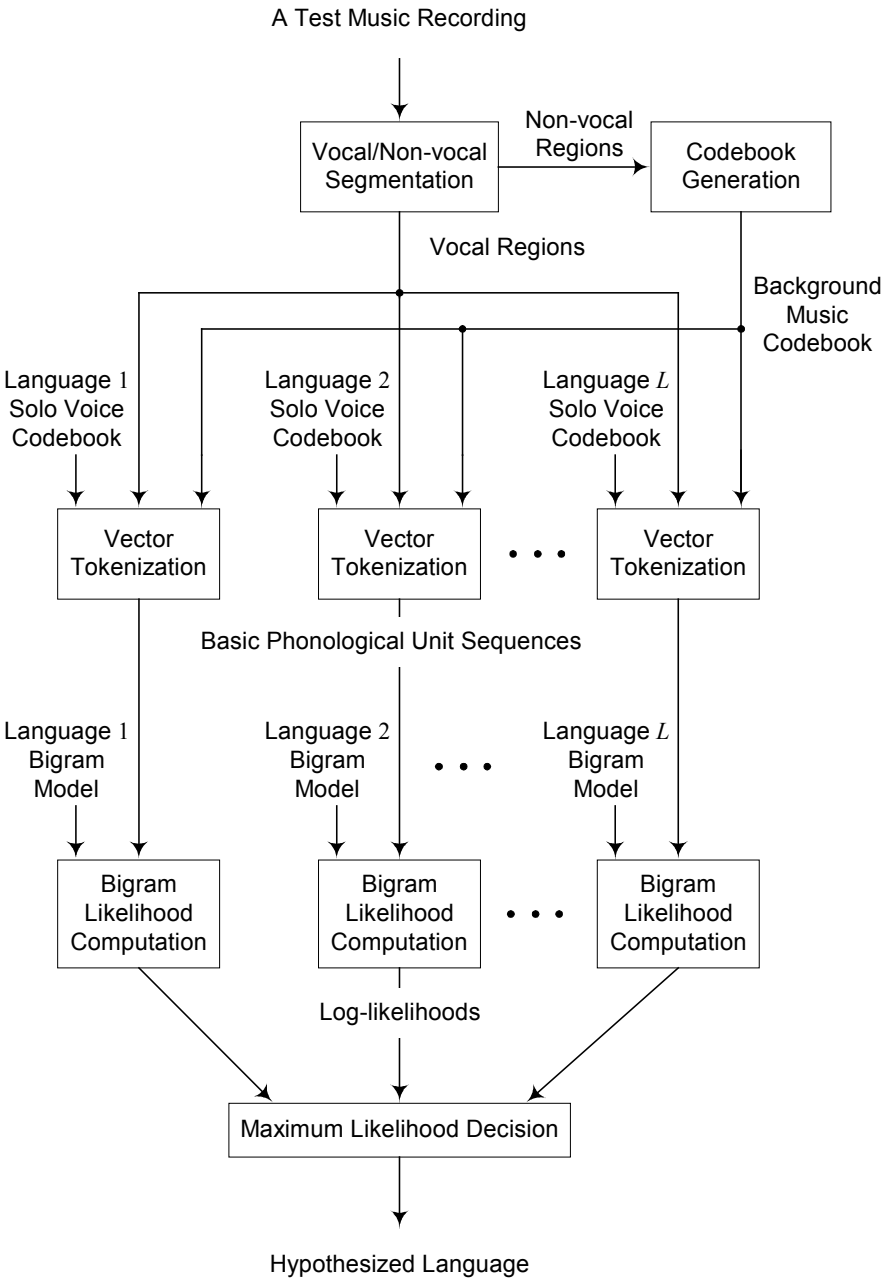


Fig. 4. Procedure for hypothesizing the language of an unknown test music recording

For each language l, the dynamics of the basic phonological unit sequence $W^{(l)}$ are examined using the bigram language model $\lambda^{(l)}$, in which the log-likelihood $\log p(W^{(l)}|\lambda^{(l)})$ that $W^{(l)}$ tests against $\lambda^{(l)}$ is computed by

$$\log p(W^{(l)} \mid \lambda^{(l)}) = \frac{1}{N^{(l)}} \sum_{t=1}^{N^{(l)}} \log p(w_t^{(l)} \mid w_{t-1}^{(l)}). \tag{11}$$

Note that transitions between vocal regions are not considered when computing Equation (11). According to the maximum likelihood decision rule, the identifier should decide in favor of a language that satisfies

$$l^* = \arg\max_{l} \log p(W^{(l)} \mid \lambda^{(l)}). \tag{12}$$

6. Experiments

6.1 Music Database

The music database used in the experiments consisted of 346 tracks from Mandarin, English, and Japanese pop music CDs, which reflected the major languages of popular music in Taiwan. The database roughly covered five genres: soundtracks, country, folk, jazz, and rock. The length of each track was around three minutes. As shown in Table 1, the database was composed of three subsets. Subset DB-1 was used to train the sung-LID system, while subsets DB-2 and DB-3 were used to evaluate the system's performance. In DB-1, there were 60 Mandarin songs, 60 English songs, and 60 Japanese songs, denoted as DB-1-M, DB-1-E, DB-1-J, respectively. In DB-2, there were 20 Mandarin songs, 20 English songs, and 20 Japanese songs, denoted as DB-2-M, DB-2-E, DB-2-J, respectively. There was no overlap between the singers in DB-1 and DB-2.

Subset DB-3 contained 53 pairs of cover versions of songs. Each pair consisted of two songs with similar tunes and accompaniments, but they were sung in different languages. There were 32 pairs of English-Mandarin cover versions and 21 pairs of Japanese-Mandarin cover versions,

denoted as DB-3-EM and DB-3-JM, respectively. Of the 53 pairs, 18 pairs were performed by bilingual singers, i.e., each singer performed two songs with almost identical tunes, but in different languages. None of the singers in DB-3 appeared in DB-1 or DB-2. The purpose of using a dataset like DB-3 is to avoid the bias that arises from tunes, singers, or music styles, which may affect the objectivity of assessing a sung-LID system. For example, some English songs may be easily distinguishable from Mandarin songs due to the significant differences in the musical style of most English and Mandarin pop songs, rather than the underlying languages. Thus, to avoid overestimating the performance of sung-LID, we used DB-3-EM to check if a song with a typical English music style sung in Mandarin could be correctly identified as Mandarin.

Table 1. Music database

Subset			Vocal/Non-vocal Labeling	Purpose
DB-1	DB-1-E	60 Mandarin Songs	Yes	Training
	DB-1-M	60 English Songs	Yes	
	DB-1-J	60 Japanese Songs	No	
DB-2	DB-2-E	20 Mandarin Songs	Yes	Evaluation
	DB-2-M	20 English Songs	Yes	
	DB-2-J	20 Japanese Songs	No	
DB-3	DB-3-EM	32 Pairs of English-Mandarin Cover Songs	Yes	Evaluation
	DB-3-JM	21 Pairs of Japanese-Mandarin Cover Songs	No	

However, during this initial development stage, we only labeled part of the database with vocal/non-vocal boundaries, as shown in Table 1. In addition, to exclude high frequency components beyond the range of normal singing voices, all music data was down-sampled from a CD sampling rate of 44.1 kHz to 22.05 kHz. Feature vectors, each consisting of 20 Mel-scale frequency cepstral coefficients, were computed using a 32-ms Hamming-windowed frame with 10-ms frame shifts.

6.2 Experiment Results

Since the language used in a song is independent of the accompaniment, it is desirable that all non-vocal segments should be identified and removed. To determine the degree that the performance of the proposed sung-LID system is affected by imperfect vocal/non-vocal segmentation, we began the experiments by using only music data with vocal/non-vocal labeling.

First, we used DB-1-E and DB-1-M to train the vocal/non-vocal models and evaluated the performance of the vocal/non-vocal segmentation using DB-2-E, DB-2-M, and DB-3-EM. The performance was characterized by a frame-based accuracy rate computed as the percentage of correctly-hypothesized frames over the total number of test frames. In view of the limited ability of the human ear to detect vocal/non-vocal changes (Delacourt and Wellekens, 2000), all frames that occurred within 0.5 seconds of a perceived switch-point were ignored when computing the accuracy. Using 64 mixture components per GMM along with 60-frame analysis segments (empirically the most accurate configurations), we achieved a segmentation accuracy rate of 79.2%.

Then, sung-LID experiments were conducted using DB-2-E, DB-2-M, and DB-3-EM. During each test, the language sung in a music recording was identified as either English or Mandarin. We evaluated the performance with respect to test recordings of different length. Each track in DB-2-E, DB-2-M, and DB-3-EM was divided into several overlapping clips of T feature vectors. A 10-sec clip corresponded to 1000 feature vectors, and the overlap of two consecutive clips was 500 feature vectors. Each clip was treated as an individual music recording. Sung LID was performed in a clip-by-clip manner, and the technique's performance was evaluated on the basis of its accuracy, which is the percentage of correctly-identified clips over the total number of test clips. In the training phase, the number of codewords used in each language-specific solo voice codebook and the background music codebook were empirically determined to be 32 and 16, respectively. In the testing phase, if the number of non-vocal frames exceeded 200, an

online-created background music codebook was empirically set to have 4 codewords; otherwise, background music codebook was not used. The segment length for smoothing the generated codeword index sequences was empirically set to 5, and the interpolating factor α in Equation (2) was set to 0.9. For performance comparison, we also implemented sung LID without using any background music codebook.

Figure 5 shows the sung-LID results with respect to T = 3000 (30 sec), 6000 (60 sec), 9000 (90 sec), and entire track, in which clips labeled as totally non-vocal were not used for testing. The figure shows that, as expected, the sung-LID accuracy rate improves as the length of the test recordings increases. We also observe that sung LID with solo voice codebooks outperforms the method that does not consider background music. The performance gap increases as the length of the test music recordings increases. This is because longer recordings contain more information about the background music, which can be exploited to obtain a song's underlying phonological units more reliably. However, since automatic vocal/non-vocal segmentation is far from perfect, the sung-LID accuracy rate declines by approximately 10% when the vocal/non-vocal segmentation is performed automatically instead of manually. In addition, we observe that the sung-LID accuracy rates in Figure 5 (b) are significantly lower than those in Figure 5 (a). This reflects the fact that LID for songs translated from another language is more difficult than LID for the songs performed in their original language. Overall, the sung-LID accuracy rate is much lower than the spoken-LID accuracy rate in the literature (e.g., Parris and Carey, 1995; Zissman, 1995). This is mainly attributed to the interference from background accompaniment.

Table 2 shows the confusion probability matrix for the results of sung LID based on the automatic vocal/non-vocal segmentation of each track. The rows of the matrix correspond to the ground-truth of the tracks, while the columns indicate the hypotheses. The majority of errors are caused by mis-identification of English songs. We speculate that such errors might be due to the louder background music that usually exists in English pop songs, compared to Mandarin songs.

The latter often have louder vocals to ensure that Mandarin syllables can be heard and understood because of the lack of tone information. Thus, the lower vocal-to-background ratio may make it difficult to train the English model reliably. In addition, the reason for the bias towards identifying the tracks in DB-3-EM as Mandarin is probably because a large proportion of the singers in DB-3-EM are Chinese. The accents of those singers probably differ significantly from those of the singers in DB-1-E, who are mainly Western; hence, the resulting discrepancy in phonological realizations may also cause the English model to match the test music recordings inaccurately.

Next, we extended the experiments to identify the three languages of interest, namely, Mandarin, English, and Japanese. Since no vocal/non-vocal labeling was available in DB-1-J, DB-2-J, and DB-3-JM, we performed automatic segmentation to mark the vocal/non-vocal boundaries. Specifically, vocal and non-vocal models, trained using DB-1-E and DB-1-M, were used to recognize the vocal/non-vocal boundaries in DB-1-J. Then, the Japanese solo voice codebook and bigram model were generated using DB-1-J with the automatically segmented vocal and non-vocal regions. During each test, the language sung in a music recording was identified as one of the three languages of interest.

Table 3 shows the results of testing each track in DB-2 and DB-3. The sung-LID accuracy rates for DB-2 and DB-3 were 65.0% and 59.6%, respectively. We can see from Table 3 that identifying Mandarin songs is easier than identifying English and Japanese songs. Significantly, a large proportion of Japanese songs were identified as Mandarin. This may be attributed to the inferior training process for Japanese due to the lack of vocal/non-vocal labeling. In addition, it was found (Muthusamy, Jain, and Cole, 1994; Parris and Carey, 1995) that from the spoken-LID point of view, Japanese is confused more often with Mandarin than English. Moreover, as Mandarin and Japanese pop songs are often regarded as East Asian pop music (Chua, 2004), in which songs of one language are often adapted, copied, mixed, and reproduced into songs of another language, the high degree of closeness and influence between Mandarin and Japanese pop songs may also

contribute to the bias towards mis-identifying Japanese songs as Mandarin rather than English. Averagely, although only sixty percent of songs can be identified correctly, the result is much higher than the chance probability (one third).

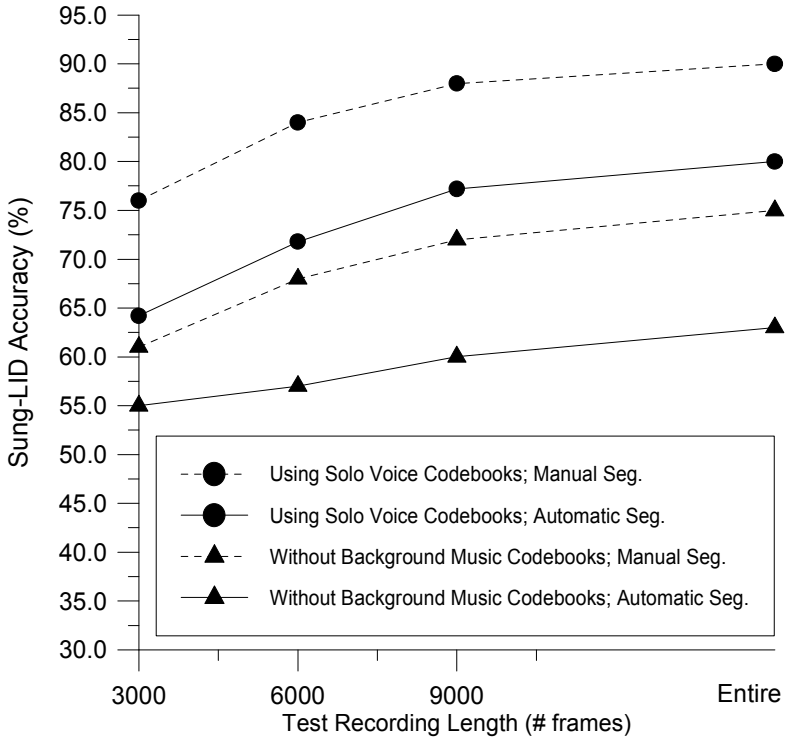
The above experiment results indicate plenty of room left for performance improvement. We believe that the proposed sung-LID system can benefit significantly from improvements in the vocal/non-vocal segmentation as well as the increase in the amount of training data. Although the current system is far from ready to use in practice, it would be possible to build a feasible sung-LID system based on extensions of the framework developed in this study.

7. Conclusions and Future Works

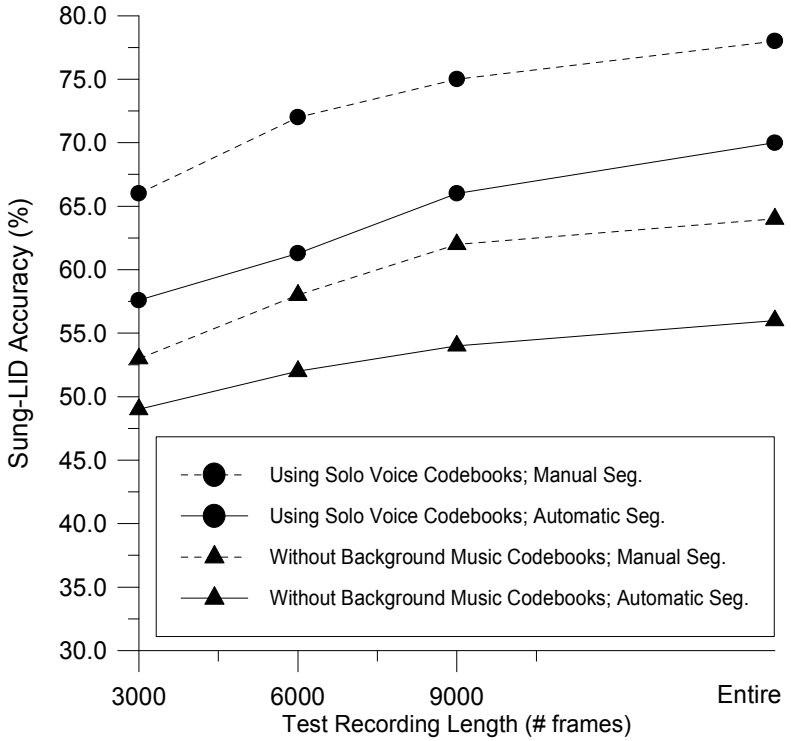
We have examined the feasibility of automatically identifying the language sung in a popular music recording. Moreover, we have shown that the acoustic characteristics of a language can be extracted from singing signals via grammatical modeling of the basic phonological unit sequences output from the vector tokenization of spectrum-based features. To eliminate interference from background music, we have proposed a more reliable codebook generation method for vector clustering based on an estimation of the solo voice's characteristics.

Although this study has shown that the languages sung in different music recordings can be distinguished from one another, the proposed method and the experiments only represent a preliminary investigation into the sung-LID problem. First, due to the limited availability of music data, we only examined the sung-LID system using Mandarin, English, and Japanese pop songs available in Taiwan. To be of more practical use, our future work will consider the case of identifying distinct languages with very similar phonological inventories and phonotactic patterns, such as Spanish vs. Italian and English vs. Dutch or German. Though such a case has been evaluated in spoken-LID research (e.g., Parris and Carey, 1995), the methodology is essentially linguistics-oriented and vulnerable to interference from background accompaniment. With regard

to usability and portability, it would be desired to develop a robust data-driven system for sung-LID. Second, the sung-LID investigated in this study belongs to a *closed-set* recognition problem, which assumes only languages of interest will be encountered, i.e., a forced choice among a set of languages known to the system. However, in practical use, it is inevitable to handle songs not performed in the languages known to the system, i.e., it is required to distinguish the languages known to the system from a large set of languages unknown to the system. Our future work will also study such an *open-set* recognition problem. Third, this work does not deal with the songs containing more than one language, i.e., mixed languages. As multilingual lyrics are commonplace in popular music, the problem of sung-LID in mixed-language songs is also worth a deep investigation. However, to explore the above-mentioned issues, the music database must be scaled up to cover more languages, singers with a wider variety of accents, and richer music styles or genres.



(a) Experiments on DB-2-E and DB-2-M



(b) Experiments on DB-3-EM

Fig. 5. Sung-LID results

Table 2. Confusion probability matrix of the discrimination of Mandarin and English songs

(a) Experiments on DB-2-E and DB-2-M

	Hypothesized	
Actual	Mandarin	English
Mandarin	0.85	0.15
English	0.25	0.75

(b) Experiments on DB-3-EM

	Hypothesized	
Actual	Mandarin	English
Mandarin	0.78	0.22
English	0.37	0.63

Table 3. Confusion probability matrix of the discrimination of Mandarin, English, and Japanese songs

(a) Experiments on DB-2

	Hypothesized		
Actual	Mandarin	English	Japanese
Mandarin	0.75	0.10	0.15
English	0.20	0.65	0.15
Japanese	0.30	0.15	0.55

(b) Experiments on DB-3

	Hypothesized		
Actual	Mandarin	English	Japanese
Mandarin	0.65	0.14	0.21
English	0.29	0.57	0.14
Japanese	0.29	0.21	0.50

8. Acknowledgement

This work was partially supported by National Science Council, Taiwan under Grants NSC93-2422-H-001-0004, NSC95-2221-E-001-034, and NSC95-2218-E-027-020.

References

Akeroyd, M. A., Moore, B. C. J., and Moore, G. A. (2001). Melody recognition using three types of dichotic-pitch stimulus. *Journal of the Acoustical Society of America*, 110(3): 1498-1504.

Chua, B. H. (2004). Conceptualizing an East Asian popular culture. *Inter-Asia Cultural Studies*, 5(2): 200-221.

Cummins, F., Gers, F., and Schmidhuber, J. (1999). Language identification from prosody without explicit features. *Proceedings of the European Conference on Speech Communication and Technology*, Budapest, Hungary.

Delacourt, P., and Wellekens, C. J. (2000). DISTBIC: A speaker-based segmentation for audio data indexing. *Speech Communication*, 32: 111-126.

Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 39: 1-38.

DuPreez, J. A., and Weber, D. M. (1999). Language identification incorporating lexical information. *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, Phoenix, USA.

Eronen, A. (2003). Musical instrument recognition using ICA-based transform of features and discriminatively trained HMMs. *Proceedings of the International Symposium on Signal Processing and Its Applications*, Paris, France.

Harbeck, S., and Ohler, U. (1999). Multigrams for language identification. *Proceedings of the European Conference on Speech Communication and Technology*, Budapest, Hungary.

Hazen, T. J. and Zue, V. W. (1997). Segment-based automatic language identification. *Journal of the Acoustical Society of America*, 101(4): 2323-2331.

House, A. S. and Neuburg, E. P. (1977). Toward automatic identification of the language of an utterance. I. Preliminary methodological considerations. *Journal of the Acoustical Society of America*, 62(3): 708-713.

Jelinek, F. (1990). Self-organized language modeling for speech recognition. *Readings in Speech Recognition*, Palo Alto, CA: Morgan Kaufmann, chap 8, pp. 450-506.

Kim, Y. E. and Whitman, B. (2002). Singer identification in popular music recordings using voice coding features. *Proceedings of the International Conference on Music Information Retrieval*, Paris, France.

Medina, R. A., Smith, L. A., and Wagner, D. R. (2003). Content-based indexing of musical scores. *Proceedings of the Joint Conference on Digital Libraries*, Texas, USA.

Muthusamy, Y. K., Jain, N., and Cole, R. A. (1994). Perceptual benchmarks for automatic language identification. *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, Adelaide, Australia.

Muthusamy, Y. K., Barnard, E., and Cole, R. A. (1994). Reviewing automatic language identification. *IEEE Signal Processing Magazine*, 4: 33-41.

Nadas, A., Nahamoo, D., and Picheny, M. A. (1989). Speech recognition using noise-adaptive prototypes. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 37(10): 1495-1503.

Nakagawa, S., Ueda, Y., and Seino. (1992). Speaker-independent, text-independent language identification by HMM. *Proceedings of the International Conference on Spoken Language Processing*, Alberta, Canada.

Navratil, J. and Zuhlke, W. (1998). An efficient phonotactic-acoustic system for language identification. *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, Seattle, USA.

Navratil, J. (2001). Spoken language recognition – A step toward multilinguality in speech processing. *IEEE Transactions on Speech and Audio Processing*, 9(6): 678-685.

Parris, E. S., and Carey, M. J. (1995). Language identification using multiple knowledge sources. *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, Detroit, USA.

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

Schwenninger, J., Brueckner, R., Willett, D., and Hennecke, M. (2006) Language Identification in Vocal Music. *Proceedings of the International Conference on Music Information Retrieval*.

Tsai, W. H. and Wang, H. M. (2004). Towards automatic identification of singing language in popular music recordings. *Proceedings of the International Conference on Music Information Retrieval*, Baltimore.

Tsai, W. H. and Wang, H. M. (2006). Automatic singer recognition of popular music recordings via estimation and modeling of solo vocal signals. *IEEE Trans. on Audio, Speech, and Language Processing*, 14(1): 330-341.

Tzanetakis, G. and Cook, P. (2002). Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing*, 10(5): 293-302.

Wang, C. K., Lyu, R. Y., and Chiang, Y. C. (2003). An automatic singing transcription system with multilingual singing lyric recognizer and robust melody tracker. *Proceedings of the European Conference on Speech Communication and Technology*, Geneva, Switzerland.

Zissman, M. A. (1995). Comparison of four approaches to automatic language identification of telephone speech. *IEEE Transactions on Speech and Audio Processing*, 4(1): 31-44.