SINGER VERIFICATION: SINGER MODEL .VS. SONG MODEL

L. Regnier, G. Peeters

IRCAM, Sounds Analysis-Synthesis team, CNRS-STMS
1 place Stravinsky, 75004 Paris

ABSTRACT

This paper proposes a method to verify the singer identity of a given song. The query song is modeled as a GMM learned on the features extracted from sustained sung notes of the song. Each note is described by the shape its spectral envelope and by the temporal variations in frequency and amplitude of its fundamental frequency. The singer identity is verified with two approaches: the model of the query song is compared to a singer-based GMM or compared to the GMM of another song performed by the same singer. The comparison is done using a dissimilarity measurement given by the Kullback Leibler divergence. When the two types of features are combined, the proposed approach verifies the singer identity of a given a cappella song with an error rate lower than 8% when the whole song is considered and an error rate lower than 10% when a short excerpt of the song (i.e. 15 consecutive sustained notes) is considered.

1. INTRODUCTION

The singing voice is the element of a song that attracts the most attention of listeners. Therefore, information on the singers voice is essential to organize, browse, and classify music collections. The singer recognition task encompasses the identification and recognition. Usually, an identification system ([1], [2],[3]) is trained to recognize the singer of a given song from a set of N possible singers composing the data set. The singer verification ([4],[5]) aims to decide whether or not a claimed singer performs a given song. Unlike the singer identification, the performance of a verification system does not depend on the number of singers in the dataset neither on the distribution of the singers over the dataset.

Both tasks require a good understanding of what defines the signature of a singer. Most studies on the recognition of singers model the singer signature using features related to the time extracted from the (short-term) amplitude spectrum. In our research we propose to complete the description of the singer signature by introducing new features related to the style and the technique of the singers. For this purpose we propose to describe some intonative aspects of the singing voice by describing the temporal variations of the fundamental frequency (f_0) of the sung melody. The quasi-sinusoidal variations of the f_0 are related to the *vibrato and tremolo* while the continuous variations of the f_0 are related to the *portamento*. The features used to describe these intonative elements, named INTO in the following, have been proven to be efficient to distinguish the singing voice among the other musical instruments in [6]. We investigate in this paper if the intonative elements are singer-specific and can thus be used to model the signature of a singer.

In this study a song is represented by a succession of sustained sung notes. From each note we extract the INTO features and some "timbral" features (cepstral coefficients derived from the spectral envelope estimated with the "true envelope": TECC). Each song is modeled as a GMM learned using either the INTO or the TECC features extracted from all notes composing the song. Each singer of the data set is also modeled as a GMM. The song-to-song and/or song-to-singer similarity is given by the Kullback Leilber (KL) divergence computed on their GMMs. The similarity measurements are used to perform two cases of verification: verify if a song is performed by a claimed singer (singer-level), and verify if two songs are performed by the same singer (song-level). The comparisons are performed on models obtained using the same type of feature. Finally, to combine information conveyed by INTO and TECC features a new similarity measurement is defined by summing the similarity measures obtained for each type of feature.

The paper is organized as follows. In Sec.2, the task of verification is presented. Sec.3 gives the details of the proposed method. The experiments conducted and the results obtained on a cappella recordings are presented in Sec.4. Sec.5 draws some conclusions and suggests some future works.

2. GENERAL APPROACH FOR VERIFICATION

2.1. Theoretical backgrounds

Singer verification is a binary classification problem that aims to decide whether or not a claimed singer has performed a given song. The formalism of verification task has been first defined in the speech community. A detailed description of verification methods is given in [7]. In our case, for a song x represented by X and a singer c with corresponding model S_c the system has to chose between two hypothesis: $\mathbf{H_0}$, x is sung by singer c and $\mathbf{H_1}$, x is not sung by singer c. Then, $\mathbf{H_0}$ is verified if (1) is satisfied. Applying Bayes rule, (1) can be expressed in the log-domain as (2).

$$P(S_c|X) > P(\bar{S}_c|X) \quad (1)$$

$$\log(p(X|S_c)) - \log(p(X|\bar{S}_c) > \Lambda(X) \quad (2)$$

The decision threshold $\Lambda(X)$ encompasses the prior probabilities and an additional threshold for the hypothesis validation. The log-likelihood ratio is compared to a threshold τ and the claimed singer is accepted if $\Lambda(X) > \tau$ and rejected otherwise.

The choice of the impostor model $\bar{S}_c$ (or background model) is a major issue in verification problems which is related to the problem of score normalization. Numerous alternatives have been proposed to solve this problem as described in [8]. In general, the background model is selected to represent the population of expected impostors depending on the application. Using several background models is a solution to better model the impostor population.

2.2. Evaluation of verification system

A complete description of the paradigm for the evaluation of verification tasks can be found in [9]. Two types of errors can occur:

accept an identity claimed by an impostor **(False Alarm: FA)** and reject a valid identity **(Miss Detection: MD)**.

To evaluate the performance of a verification system, a certain number of trials are given as inputs to the system which has to distinguish between the *true-trials* (the song is from the claimed singer) and the *false-trials* (the song is not from the claimed singer). For a given threshold τ, the FA and MD rates are coupled to form an operating point. By varying τ a set of points forming a Detection Error Tradeoff (DET) curve is obtained. To report the performance with a single value the Equal Error Rate (EER) (i.e. the value corresponding to MD = FA) can be used.

2.3. From song similarity to singer verification

The log-likelihood measurements in (2) can be replaced by any similarity measure because the only important element is the relative values of these measurements. Thus, the score of a trial can be expressed as:

$$score(x, c) = sim(S_c, X) - sim(\bar{S_c}, X) \quad (3)$$

The verification is performed by comparing this score to a threshold.

The similarity between a song and a singer can be defined in the same way than the similarity between songs developed to generate automatically playlist ([10], [11]) and to identify singers ([2], [1]). A representative collection of works on song similarity is presented and evaluated in [12]. Most of these studies compare songs on the basis of their global timbre. For each song, spectral features are extracted from frames all over the song and are then clustered to group similar frames together. Each clusters is described by its mean and its variance. Finally, an entire song is modeled as a GMM based on song-level features [1]. Based on the assumption that the temporality is not of importance, suggesting that a song played with two different temporal orders is similar to itself, songs are compared by measuring similarity between their models. The (dis)similarity between two models can be computed using the Kullback Leibler (KL) divergence. If song-models are simple Gaussians, KL can be applied directly. If songs are modeled as GMM, the KL divergence can be approximated by the Monte Carlo method ([1], [2]) or by the Earth Moving Distance ([11], [13], [14]). The difference between the two approaches is not significant in term of quality according to [10]. KL divergence, denoted by d_{KL}, is a non-symmetrical dissimilarity measurement, that can be converted into a similarity measure by:

$$sim(X, Y) = e^{-(d_{KL}(X,Y)+d_{KL}(Y,X))} \quad (4)$$

The similarity between a singer and a song is defined using a similar approach: a singer model is learned using features extracted from different songs of this singer. The assumption is that a singer is similar to himself through different songs.

In content-based audio classification (genre or artist classification), the similarity measures are given as inputs of a classifier (e.g. SVM [2] or kNN [15]), which assigns the song to class of the problem. In our approach, we propose to use the similarity measurements as inputs of a verification system.

3. PROPOSED APPROACH

We chose to model a song using a set of sustained sung notes extracted from that song and to model a singer using a set of songs (i.e. using the notes composing its set of training songs). Therefore, the singer and song models rely on an appropriate description of the sung notes.

[1] this term is introduced in [2]

In this study, we work with a cappella songs. The f_0 of the sung melody is estimated using the YIN algorithm and the sustained notes are manually segmented.

3.1. Description of one note

Finding an invariant voiceprint in the sung signals univocally characterizing a singer is at the basis of the singer recognition problem. We propose two complementary approaches to derive information on the singer identity from a sustained note:

3.1.1. Description of spectral contents: timbre

According to the source-filter model, the spectral envelope of a spoken utterance gives the transfer function of the vocal tract, which is obviously a characteristic of the singer identity. This envelope conveys information related to the timbre of the sound. We chose to describe the spectral envelope by a set of Cesptral-Coefficients derived from the "true-envelope" [16] estimated on frames of 40 ms length (with a hop size of 10ms). The DCT of the envelope is computed and the 25 firsts coefficients are retained to form the spectral description of one frame.

3.1.2. Description of frequency variations: style and technique

The singing voice has some particularities. First of all, it is impossible for a singer to sustain a note without slightly varying the pitch. In most cases, the pitch varies around a mean in a periodic way generating a natural **vocal vibrato** (FM). This frequency modulation creates a passive modulation of amplitude (AM) because the vocal tract enhances consistently some frequencies crossed by the FM [17]. The correlation between AM and FM conveys an interesting information on the vocal tract of the singer. The **portamento** is another particularity of singing voice that refers to the smooth transitions between distant notes sung in the same breath. Portamento corresponds to slow and monotonic variation of frequency between the two pitches.

We model the f_0 of a sung tone as the sum of a slow varying frequency ($d_f(t)$) representing the portamento and a periodic modulation ($s(t)$) associated to the vibrato, where the periodic modulation is modeled by an exponential sinusoidal model (ESM) .

$$f_0(t) = d_f(t) + s(t), \text{ with} \quad (5)$$

$$s(t) = a_0.e^{a_1 t}\cos(2\pi r t + \phi_0) \quad (6)$$

The choice of ESM for vibrato model is rather appropriate since some singers have the particularity to attenuate their vibrato towards the end of the note while some others need a lapse of time before setting their vibrato resulting in an exponential slow variation of the amplitude variation over time. In practice, the frequency deviation $d_f(t)$ is estimated with a polynomial of degree 3 (as described in [6]) and is then subtracted from the original f_0. The parameters of the residual, which represents the vibrato, are estimated using High Resolution method (HR). The traditional HR method [18] is applied for one frequency instead of a sum of sinusoids. Finally, the f_0 of one note is described with 4 parameters for the polynomial representing $d_f(t)$ plus 3 parameters (a_0, a_1, r) for the sinusoidal part.

To estimate the parameters of the amplitude variations, the time-varying amplitude function associated to the time varying $f_0(t)$ is also modeled with Eq.(5) and (6). The sinusoidal modulation is interpreted as a **tremolo** while the continuous variation is interpreted as a variation of dynamic (such as a *crescendo*).

The characteristics of the singing voice mentioned is this paragraph are known to help the voice to stand out from the instrumental

accompanied and to add expression. As they can be interpreted as parameters of expressive variations of frequency, we refer to them as intonative (INTO) feature.

3.2. Model of song and singer

In the studies presented in Sec.2.3, song models are built using song-level features extracted from all frames of the song. Here, information on songs is obtained on a set of sustained notes coming from the song instead of the entire song.

3.2.1. Song-based and Singer-based GMM

The song and singer models are obtained using one type of feature at a time. A song model, song-based GMM, is learned from the features of all notes composing the song. A singer is defined by a set of songs and is modeled by a singer-based GMM leaned from all notes of all songs of the set. In this study, GMMs are composed with a mixture of 2 multivariate Gaussians with a diagonal covariance matrix. For the TECC features, models are learned using information obtained on each frame of each note composing the songs. For the INTO features, one note is described by one feature vector. Therefore, the INTO-based models are learned using these feature vectors obtained on all notes composing the songs. The song whose singer identity has to be verified is modeled as song-based GMM, and the similarity between the models is computed between GMMs obtained using the same type of feature.

3.2.2. Models comparison

The distance between two GMM is given by the symmetric version of the KL divergence approximated using the Monte Carlo method (with 200 samples). The distances computed between the GMMs obtained using INTO and TECC features are respectively denoted by d_{INTO} and d_{TECC}. To combine information conveyed by the two types of features we define a new distance between models X and Y as:

$$d_{CUM}(X,Y) = d_{INTO}(X,Y) + d_{TECC}(X,Y) \qquad (7)$$

3.2.3. Verification experiments

On a given data set, the verification experiment is performed using each singer of the dataset as a claimant with the remaining singers acting as impostors and rotating through all singers as described in [7]. The data set is split into training and testing subsets of songs. Training songs are reserved to model singer identities and the evaluation is performed using the remaining songs as query songs. Singer identity can be modeled using one singer-based GMM (**singer-level**) or using a set of q song-based GMM (where q is the number of songs used to trained the singer model). This last approach is referred to as **song-level**. For a query song and a claimed singer, the task is to verify if the claimed singer performs the song. For the singer-level approach, the verification is done directly by comparing the query song to the singer models. For the song-level approach, the task can be interpreted as follow: is the query song performed by the same singer than the song s_c ? (where s_c comes from the claimed singer.)

4. EXPERIMENTS AND RESULTS

4.1. Data Set

The experiment is conducted on a set of 18 singers. For each singer we work with the lead vocal track of 3 songs. We have selected an average of 50 notes per track. Finally, we have 54 songs and a total of 2592 notes. For each singer, 2 songs are used to train the model and the remaining song is used as a query song. All experiments are conducted with a 3 folds cross-validation by rotating training and testing songs. On each note we extract the features described in Sec.3.1. The singer and song models are obtained with the method presented in Sec.3.2.

4.2. Experiments

First, we evaluate the capacity of our system to verify the singer identity of a given song. The singer identity is verified using either the entire song (*song-based singer verification*) or a short excerpt of the song. To simulate an excerpt of song, we model the song using a set of p consecutive notes. This approach is referred to as *p_notes-based singer verification*.

4.3. Results

4.3.1. Song-based verification

We report in Table 1 the EER obtained for the song-based singer verification for each feature separately and for combined features.

Feature	**Song-level**	**Singer-level**
TECC	17.81	14.49
INTO	13.34	10.68
CUM	9.20	**7.41**

Table 1. Mean EER through the 3 folds cross-validation

The singer-level approach yields the best performance. The TECC lead to a lower performance than the INTO features. When the two types of features are combined the verification is performed with a very low error rate. In practice it is not always straightforward to combine features, especially when they have different sizes, meanings and ranges. We note that the proposed method allows a very simple scheme of feature combination (see Eq.(7)).

Next, we evaluate the performance of the singer verification when using an excerpt instead of the whole song.

4.3.2. p_note-based verification

The length of the expert is defined by the number of notes (p) composing the excerpt. In Fig.1 the EER for singer and song levels experiments are plotted for p varying from 2 to 30. The min and max EER obtained during the 3 folds experiments are reported with errors bars. Like the song-based experiments, the singer-level approach yields the best performance. For a small number of notes ($p < 20$), the TECC features lead to better results than INTO. However, for $p > 5$ the worst EER obtained with INTO feature is always lower than the worst EER obtained using TECC. We also note that the variance of performance through the folds is much smaller for INTO than TECC. The large variation of performance through folds for the TECC can be explained by the "album effect". The spectral envelope is strongly affected by post-processing treatments even when performing a cappella.

Like the experiments performed on whole songs, the combination of both feature types greatly reduces the error rate and the variance of performances through the folds. Fig.2 displays the DET (and EER) curves for $p = 30$ for each FA of feature CUM. As shown on these plots, the FA rate can be bound to 20% for the singer-level and to 40% for the song-level approach. For an application specified in term of cost of MD and FA with a low cost for the false alarms, the proposed method leads to a very low minimum detection cost (MDC).

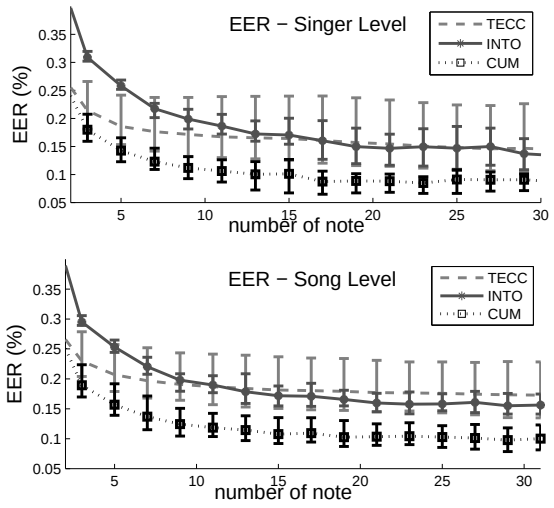

Fig. 1. EER per type of features and combined features

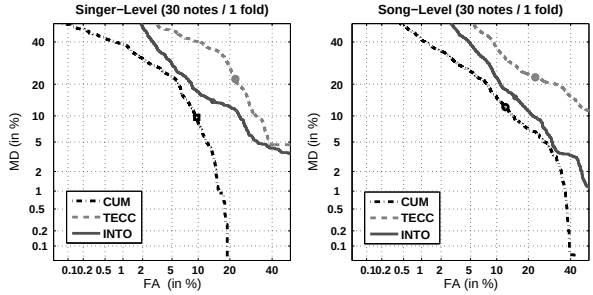

Fig. 2. DET curves for p_note-based verification (p=30)

5. CONCLUSION AND FUTURE WORKS

In this study we described a song by a set of sustained sung notes. From these notes we extracted frame-based features describing the shape of the spectral envelope by mean of cepstral coefficients derived from the true envelope (TECC) and note-based features to describe the variations in frequency and amplitude of the f_0 of the note (INTO). Using these two sets of features we conducted two experiments to verify the singer identity of a given song. The first approach, **singer-level**, verifies the singer identity of a query song using a distance between a singer-based GMM (trained using different songs of the singer) and a song-based GMM. The distance, computed using the Kullback Leibler divergence approximated by Monte Carlo method, is transformed into a similarity measurement to perform a verification task. The second approach, **song-level**, verifies if two songs are performed by the same singer by measuring the similarity between two song-based GMM. For all experiments, the singer-based and the song-based GMM were computed using either TECC or INTO features. On our dataset, INTO features perform better than the TECC features. Since the song-to-song (and song-to-singer) similarity is computed by means of a distance, information conveyed by each type features can be simply combined by summing the distances obtained on each feature type separately. Using the combination of features, the identity of a singer is verified with an EER of 7.5% for the singer-level approach and 9% for the song-level approach.

We also evaluated the capacity of our method to verify the singer identity using an excerpt instead of the whole song. For this purpose, a small number of consecutive notes were taken to model the query song. With 15 notes, the distance obtained by combining both feature types performed the verification with en EER of 10% for both approaches.

The method presented in this paper yields very good performances for verification of singer identity on a cappella recordings even when working with an excerpt of the song. To be applied on real recordings, the fundamental frequency of the sung melody has to be extracted, segmented and the sustained notes have to be automatically selected. Theoretically, results found using INTO should remain the same if the melody is correctly transcribed. The TECC can be extracted either on isolated vocals or on accompanied vocals.

6. REFERENCES

[1] A. Mesaros, T. Virtanen, and A. Klapuri, "Singer identification in polyphonic music using vocal separation and pattern recognition methods," in *ISMIR*, 2007, pp. 375–378.

[2] M. Mandel and D. Ellis, "Song-level features and support vector machines for music classification," in *Proc. ISMIR*. Citeseer, 2005, vol. 5.

[3] T. Zhang, "Automatic singer identification," *Multimedia and Expo, 2003. ICME'03. Proceedings. 2003 International Conference on*, vol. 1, 2003.

[4] W.H. Tsai and H.M. Wang, "Automatic singer recognition of popular music recordings via estimation and modeling of solo vocal signals," *Audio, Speech and Language Processing, IEEE Transactions on*, vol. 14, no. 1, pp. 330–341, 2006.

[5] S.Z.K. Khine, T.L. Nwe, and H. Li, "Exploring perceptual based timbre feature for singer identification," *Lecture Notes In Computer Science*, pp. 159–171, 2008.

[6] L. Regnier and G. Peeters, "Partial clustering using a time-varying frequency model for singing voice detection," in *ICASSP*. IEEE, 2010, pp. 441–444.

[7] D.A. Reynolds, "Speaker identification and verification using gaussian mixture speaker models* 1," *Speech communication*, vol. 17, no. 1-2, pp. 91–108, 1995.

[8] F. Bimbot, J.F. Bonastre, and al., "A tutorial on text-independent speaker verification," *EURASIP Journal on Applied Signal Processing*, vol. 2004, pp. 430–451, 2004.

[9] D.A. van Leeuwen, A.F. Martin, M.A. Przybocki, and J.S. Bouten, "Nist and nfi-tno evaluations of automatic speaker recognition," *Computer Speech & Language*, vol. 20, no. 2-3, pp. 128–158, 2006.

[10] J.J. Aucouturier and F. Pachet, "Music similarity measures: Whats the use," in *Proc. of ISMIR*. Citeseer, 2002, pp. 157–163.

[11] B.T. Logan and A. Salomon, "Music similarity function based on signal analysis," Oct. 31 2001, US Patent App. 10/004,157.

[12] JS Downie, "Mirex 2005 contest results," *Available on-line at http://www. musicir. org/evaluation/mirex-results. Retrieved January*, vol. 9, pp. 2006, 2005.

[13] K. West and P. Lamere, "A model-based approach to constructing music similarity functions," *EURASIP Journal on Applied Signal Processing*, vol. 2007, no. 1, pp. 149–149, 2007.

[14] E. Pampalk, A. Flexer, G. Widmer, et al., "Improvements of audio-based music similarity and genre classification," in *proc. ISMIR*. Citeseer, 2005, vol. 5.

[15] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *Speech and Audio Processing, IEEE transactions on*, vol. 10, no. 5, pp. 293–302, 2002.

[16] A. Röbel, "Efficients spectral envelope estimation and its application to pitch shifting and envelope preservation," in *DAF'x*, 2005.

[17] I. Arroabarren and A. Carlosena, "Vibrato in singing voice: the link between source-filter and sinusoidal models," *EURASIP Journal on Applied Signal Processing*, vol. 2004, no. 7, pp. 1007–1020, 2004.

[18] R. Badeau, B. David, and G. Richard, "High-resolution spectral analysis of mixtures of complex exponentials modulated by polynomials," *Signal Processing, IEEE Transactions on*, vol. 54, no. 4, pp. 1341–1350, 2006.