SINGING VOICE CHARACTERIZATION FOR AUDIO INDEXING

Hélène Lachambre, Régine André-Obrecht and Julien Pinquier

Institut de Recherche en Informatique de Toulouse, UMR 5505 CNRS INP UPS UT1
118 route de Narbonne, 31062 Toulouse Cedex 9, France
{chambre, obrecht, pinquier}@irit.fr

ABSTRACT

To extract the content of audio documents, the first step in many approaches is to segment the signal in primary components, such as music and speech. Very few attention has been brought to the detection of the singing voice.

In this paper, we propose simple parameters (vibrato and harmonic coefficient) and an original segmentation based on a sinusoidal segmentation to characterize the singing voice. This information is then mixed with those issued from a speech/music decomposition.

We test this classification system on a database composed of various types of sound. We first test our system in a classification task, then in a detection task. In both cases, the results are good. In our classification system, the only misclassifications are due to very rare musical styles. In the detection task, our system misses some of the singing voice segments, but we observe very few false-alarm.

1. INTRODUCTION

In the automatic indexation process of audiovisual documents, one of the first step is to precise what kind of information is present and where it is located. Regarding the audio track, many techniques are developed to detect the presence of music, speech or other prominent sounds [1, 2]. Very few work are dedicated to indicate the presence of singing voice [3]. It is a difficult problem in the sense that its characteristics are halfway these of music and these of speech. Most occurrences of singing voice are coupled to music but in some cases, as in rap for example, it may be confused with speech!

Previously, we have proposed a speech/music classification system [4], on which our study is based. The interest of this system is that it exploits simple and robust parameters and a rule-based decision without training. We follow the same strategy to study the singing voice component: we develop a system without training, based on few - and simple - parameters and robust thresholds.

Our method consists in introducing an original segmentation based on a "sinusoidal segmentation" of the signal [5] and to extract very simple but discriminative parameters (vibrato and harmonic coefficient [6]) in correlation with this new segmentation. Some decision rules give the information of presence or absence of the singing voice.

We finally merge the information of presence or absence of singing voice with those issued from the previous speech/music decomposition. This way, we know if we have speech or music, and, if there is some music, if it is purely instrumental music or if there is someone singing.

In section 2, we describe the basic elements of our approach: the parameters and a segmentation named "sinusoidal segmentation". In section 3, we present our original segmentation, how we adapt the extraction of the parameters to it, the speech/music system and the decision strategy. Experiments and results are gathered in section 4.

2. FUNDAMENTAL FEATURES

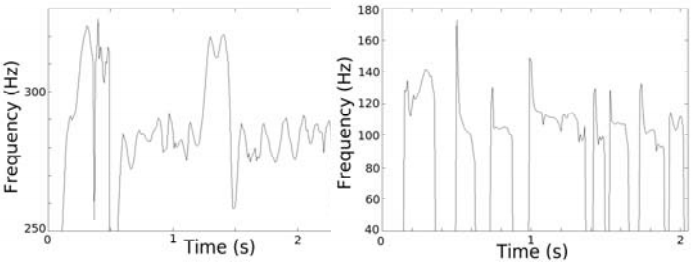
Two parameters are very interesting for the characterization of the singing voice: the vibrato and the harmonic coefficient. Coupled with the "sinusoidal segmentation", their discrimination power for the detection of the singing voice increases.

In this section, we present the original processes and uses of these three component (vibrato, harmonic coefficient and sinusoidal segmentation). Their extension is examined later in our study.

2.1 Vibrato

The vibrato is a variation of the frequency of an instrument or of the voice. The particularity of the vibrato of the voice is that it is present only when we are singing (and not during the speech). It is a natural phenomenon which is then always present (see [7, 8]), and at a very precise rate: between 4 and 8 Hertz. It is always possible to create an artificial vibrato on some musical instruments (strings and wind instruments), but it will be at a different rate.

On figure 1, the fundamental frequency is extracted along 2 seconds excerpts from singing voice (a) and speech (b). Vibrato is observed only on the figure 1.a.



(a) Song (b) Speech

Figure 1: Variation of the fundamental frequency for 2 second extracts of singing voice (a) and speech (b).

To detect the vibrato, the actual method [9] is to apply a DFT to the fundamental frequency. If a maximum is present between 4 and 8 Hertz, the presence of vibrato is confirmed.

The disadvantage of this method is that we need to extract the fundamental frequency. In the case of a polyphonic extract (multiple instruments, multiple voices or both), we do not know how to define the fundamental frequency. We will describe in part 3.2 how we deal with this problem.

2.2 Harmonic Coefficient

Considering the decomposition of the spectrum into trigonometric series (contribution of a frequency and its harmonics), the idea is to search the most important series and to measure its importance: the harmonic coefficient H_a.

This method was first used to have a better estimation of the fundamental frequency F_0 [10], but it can also be used in our study since H_a is higher in the presence of a singing voice [6]. It is calculated using a combination of the temporal and the spectral autocorrelations [10]:

- Temporal autocorrelation R^T:

$$R^T(\tau) = \frac{\sum_{n=0}^{N-\tau-1} [\tilde{s}(n) \cdot \tilde{s}(n+\tau)]}{\sqrt{\sum_{n=0}^{N-\tau-1} \tilde{s}^2(n) \cdot \sum_{n=0}^{N-\tau-1} \tilde{s}^2(n+\tau)}} \quad (1)$$

  with s the signal, $\tilde{s}$ its zero-mean version, and N the window size.

- Spectral autocorrelation R^S:

$$R^S(\tau) = \frac{\sum_{\omega=0}^{N-\omega_\tau-1} [\tilde{S}(\omega) \cdot \tilde{S}(\omega+\omega_\tau)]}{\sqrt{\sum_{\omega=0}^{N-\omega_\tau-1} \tilde{S}^2(\omega) \cdot \sum_{\omega=0}^{N-\omega_\tau-1} \tilde{S}^2(\omega+\omega_\tau)}} \quad (2)$$

  with S the magnitude spectrum of s, $\tilde{S}$ its zero-mean version and $\omega_\tau = N/\tau$ the potential F_0s.

The two autocorrelations are combined:

$$R(\tau) = \beta \cdot R^T(\tau) + (1-\beta) R^S(\tau) \quad (3)$$

The fundamental frequency F_0 is estimated by maximizing $R(\tau)$, and the harmonic coefficient is its weight:

$$H_a = \max_\tau R(\tau) = R(\frac{1}{F_0}) \quad (4)$$

In our approach, we use only H_a. Experimentally, [10] finds $\beta = 0.5$ as the optimal value, which we will also use in our study.

2.3 The sinusoidal segmentation

This segmentation, developed by [5], is the result of an automatic frequency tracking (see fig. 2). A sinusoidal segment is defined by four parameters: two temporal indexes -beginning and end-, and two vectors: one giving the values of the tracked frequency and the other giving their power. Note that the length of the vectors depends of the temporal indexes.

The sinusoidal segments are particularly significant, and their study provides new features to discriminate singing voice from speech and instrumental music, which are all harmonic sounds.

To find the sinusoidal segments, we use the following algorithm:

- compute the spectrogram every 10 ms, with a 20 ms Hamming window,
- convert the frequency in cent (100 cent = 1/2 tone):

$$f_{cent} = 1200 \cdot \log_2 \left(\frac{f_{Hz}}{440 \cdot 2^{\frac{3}{11}-5}} \right) \quad (5)$$

- smooth the spectrogram with a 17 cent window,
- detect the maxima of the spectrogram: the frequencies $(f_t^i, i = 1,...,I)$ and their log amplitude $(p_t^i, i = 1,...,I)$,
- compute the distance between two points of the spectrogram (at the instant t and $t-1$):

$$d_{i_1,i_2}(t) = \sqrt{\left(\frac{f_t^{i_1} - f_{t-1}^{i_2}}{C_f} \right)^2 + \left(\frac{p_t^{i_1} - p_{t-1}^{i_2}}{C_p} \right)^2} \quad (6)$$

  Two points $(t, f_t^{i_1})$ and $(t+1, f_{t+1}^{i_2})$ are connected (they belong to the same sinusoidal segment) if $d_{i_1,i_2}(t) < d_{th}$. C_f, C_p and d_{th} are found experimentally: $C_f = 100$ (1/2 tone), $C_p = 3$ (power divided by 2) and $d_{th} = 5$ (our experiments have confirmed the values given by [5]).

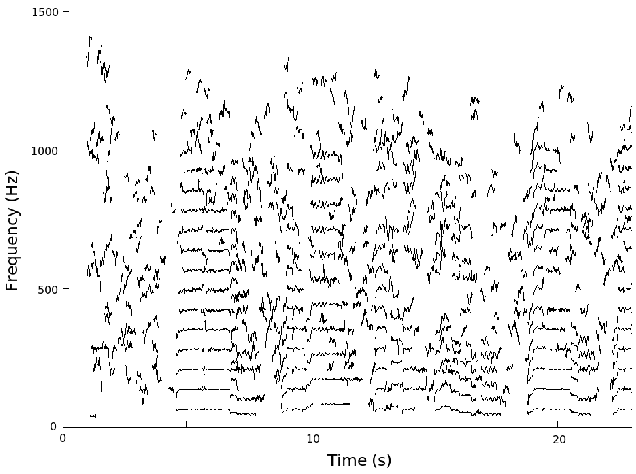


Figure 2: Example of a sinusoidal segmentation for a 23 s extract of a monophonic song a Capella: each curve is a sinusoidal segment.

3. THE SPEECH / MUSIC / SONG SYSTEM

Our global system is based on three correlated binary decisions:

- a speech/non speech decision,
- a music/non music decision,
- a singing voice/non singing voice decision.

The speech music system results from a previous study which needs only 4 parameters and an adequate rule-decision. The singing voice detection implies a new temporal segmentation and a reformulation of the vibrato.

In this section, we describe the temporal segmentation, we recall the speech and music parameters and we precise the singing voice parameters, to conclude with a set of rule based decisions.

3.1 Temporal segmentation

The inspection of some spectrograms of speech, music and songs lead us to propose this new segmentation, derived from the sinusoidal segmentation (see fig. 3). Obviously, during a stable harmonic sound (for example a note), the fundamental frequency and its harmonics begin and end at the same time. Therefore we analyze the temporal correlation between the sinusoidal segments, and more precisely between the beginnings and the ends of the segments.

To make this segmentation, we process as follow:

- extract the sinusoidal segments (see 2.3),
- find all the temporal extremities of the sinusoidal segments, but distinguish the beginnings from the ends,
- place a limit at the instant t if there are:
  - at least 2 extremities at $\boldsymbol{t}$,
  - AND at least 3 beginnings or 3 ends between $\boldsymbol{t}$ and $\boldsymbol{t+1}$ (beginning of a note or end of a note).

A temporal segment is defined by two successive limits, found by the algorithm presented above. We immediately see (fig. 3) that there is two types of segments:

- the long and stable segments,
- the short segments.

For music, each long segment should correspond to a note, while a transition between two notes is represented by a succession of short segments.

We focus our analysis on the long segments, which are more discriminative in the analysis of the singing voice: a segment is long if it is longer than 100 ms.

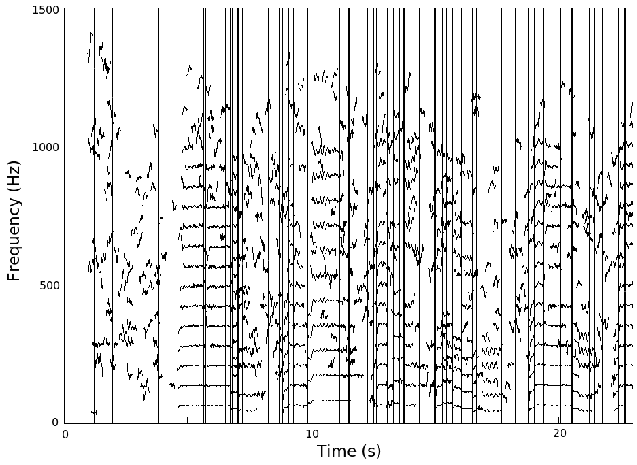


Figure 3: Temporal segmentation of the extract of fig. 2: the vertical lines are the temporal limits of the segments.

3.2 Parametrization

3.2.1 The speech-music parameters

As in [4], we exploit four parameters to detect speech and music: 4 Hz modulation of energy mod_{4Hz}, entropy modulation mod_H, stationary segment duration l, and number of segments n.

The 4 Hz modulation energy and the entropy modulation are used to detect speech. The 4 Hz modulation energy characterize the fact that about 4 syllables per second are uttered when we speak; the entropy modulation distinguishes the fact that speech signal is acoustically less structured than music.

The parameters l and n are issued from a segmentation of the signal in stationary segments described in details in [11]. They are used to detect music. n is the number of segments per second. The duration l is the mean length of the 7 longest segments in one second. For more details, see [4].

3.2.2 The singing voice parameters

The harmonic coefficient H_a is calculated as presented in section 2.2.

As we saw in part 2.1, the vibrato is an important characteristic of the singing voice. In order to use it in polyphonic music, we exploit the fact that it affects not only the fundamental frequency, but also its harmonics.

As each long sinusoidal segment corresponds to one harmonic frequency, we introduce the parameter "*vibr*" to quantify the proportion of long sinusoidal segments affected by the vibrato: the sinusoidal segments created by the presence of singing will show vibrato, while other segments (instrumental or spoken) will not. In presence of singing, we should find vibrato on the fundamental frequency of the singing and on its harmonics, so *vibr* should be high, while in all other cases, *vibr* should be low, due to the fact that there is no vibrato.

As said previously, a long temporal segment is longer than 100 ms, and we decide that the duration of a long sinusoidal segment is more than 50 ms. The value *vibr* is calculated for each segment; in the case of a short temporal segment, $vibr = 0$. It results, for the long segments:

$$vibr = \frac{\sum_{s \in \Gamma} l(s)}{\sum_{s \in \Omega} l(s)} \tag{7}$$

with:

Ω the set of the long sinusoidal segments present in the current long temporal segment,

Γ the set of the long sinusoidal segments with vibrato - i.e. with a maximum between 4 and 8 Hertz,

$l(s)$ the duration of the sinusoidal segment s.

Nota: be careful to distinguish between *temporal* and *sinusoidal* segments. *vibr* is calculated for each *temporal* segment; s is a *sinusoidal* segment.

Finally, H_a and *vibr* are averaged on 1 s in order to be on the same temporal scale as the parameters used for speech and music detection.

3.3 Decision

The global decision results from three: presence or absence of speech, of music, and of singing. For the 4 Hz modulation of energy, entropy modulation, stationary segment duration and number of segments, the decisions are taken according to the rules studied in [4]. Two new decision rules relative to the parameters H_a and *vibr* are introduced to complete the decision module and deal with the new class.

$$Speech = (mod_H \geq \lambda_1) \& (mod_{4Hz} \geq \lambda_2) \& (H_a \geq \lambda_5) \quad (8)$$

$$Singing = (not(Speech)) \& (vibr \geq \lambda_6) \quad (9)$$

$$Music = Singing \cup ((n \leq \lambda_3) \& (l \geq \lambda_4)) \quad (10)$$

$$Noise = (not(Speech)) \& (not(Music)) \quad (11)$$

The four thresholds λ_1, λ_2, λ_3, λ_4 are given by [4] while the thresholds λ_5 and λ_6 are determined experimentally (see section 4.1):

$\lambda_1 = 0.5$,
$\lambda_2 = 2.5$,
$\lambda_3 = 17$,
$\lambda_4 = 50ms$.

Note that we extend the music class: $Singing \subset Music$, and that it is impossible to have Speech and Singing at the same time. Finally, something which is neither music nor speech is classified as noise.

4. EXPERIMENTS AND TESTS

4.1 Corpus

In order to assess our system, various audio types have been performed. We have audio extracts containing: pure speech, instrumental music, pure singing, and instruments + singing. The total duration is about 7 hours and the repartition is given in table 1. The fact that there is less pure singing than the other categories is due to the fact that it is less represented in music and therefore more difficult to find. This database is sampled at 16 kHz.

We try to have as different musical styles as possible: classical music, opera, rap, country, reggae, rock, jazz, celtic, electronic music... All these extracts contain various instruments (various strings, flute, electric guitar, harp, piano, drum and some more anecdotal such as accordion or bagpipe) and various size of orchestra and choir (one person, little groups, or big bands). Finally, the singers are professional (opera, rock, country,...) or amateur chorus.

Table 1: Number of files by audio type.

Type audio	Number of files		Duration	
	Train	Test	Train	Test
Pure singing	2	9	8'	22'
Speech	3	12	25'	2h
Music	8	32	25'	2h
Music + Singing	9	36	45'	3h
Total	**22**	**89**	**1h44'**	**7h22'**

We used a part of this corpus, approximately 20 files (1/4 of the files of each type, which represents approximately 1h30) as training set to determine the values of λ_5 and λ_6 (the criterion was the minimisation of the global error). With our corpus, we found the values:

$\lambda_5 = 0.7$,
$\lambda_6 = 0.08$.

4.2 Identification task

The first assessment experiment concerns an identification task: for a given audio extract, which is homogeneous, the system indicates its nature: speech or music, and, in the case of a musical extract, the presence or the absence of singing.

Our system gives a classification every second; we decide if it is a musical or spoken extract according to the prominent class. This choice can always be made without any ambiguity since, for the speech/music classification, the error rate is lower than 10% (see [4] for more detailed results).

The analysis of several songs showed us that a singer does not always sing during the whole song (there may be instrumental interlude). It also showed us that there is a minimum duration for the singing: 1/4 of the total duration of the song. So, if we have detected a musical extract, the presence of singing will be characterized by the fact that we detect it during at least $1/4$ of the duration.

Table 2: Classification of extracts.

Decision /	Presence of singing	Speech	Music	**Number of files**
Pure singing	9	0	9	**9**
Speech	0	12	0	**12**
Instrumental music	3	0	32	**32**
Instruments & Singing	27	0	36	**36**

We can see that the results for the identification of files containing pure speech or pure singing are excellent (see table 2): for a given extract, we always classify it in the good class.

In the case of instrumental music, we sometimes detect the presence of singing while there is no singing (3 files over 32). This is due to the presence of instruments which have the same vibrato as the human voice, such as pan flute or accordion. So these false detection of singing voice are due to the presence of instruments that are rare. The opposite case (missing singing occurrences) happens more often (9 files over 36). It happens when there is not much singing in the file, and the singing is masked by the instruments. But even if we make a mistake about the presence or absence of singing, we still detect correctly that these extracts are musical extracts.

4.3 Detection task

We tested the performances of the system in a detection task: our aim is to decide, at each instant, which components are present in the signal and where.

After the decision (see part. 3.3), we have results at each second for the presence or absence of each component. This scale is appropriate for speech and music, but the detection of singing needs a longer scale because there may be short (0.5 to 1 second) interruption of the singing, notably due to the respiration of the singer: but the singing does not really stops during these interruptions. In order to take into account this fact, we smooth the results we obtain after the decision step: the presence of singing is therefore characterized by the fact that we detected it during at least 2 seconds on 3 consecutive ones.

To evaluate the results of this task, we compare our system to a "classic" one, based on MFCC and GMM. As for our system, it is the combination of three decisions: speech/non speech; music/non music and singing/non singing. For the music and speech decision, see [4]. For the singing, we extract 18 MFCC every 10 ms. Then two models are built to represent the class Singing and the non-class Non-Singing; each model is a GMM with 32 gaussians. In order to be able to compare the results, the learning and tests of this system were conducted respectively with the same files as in our system, and we made the same smoothing of the results.

Table 3: Detection rate (% of the duration).

Audio type	Our system	GMM
Speech	89.5 %	94%
Music	93%	91%
Singing	70 %	70.3%

We see from the results (see table 3) that we have no problem regarding the detection of speech (89.5 %) and music (93%): the results are comparable to those from a classic system.

Even if our singing detector is not perfect (70% of the singing is detected), we still have good results in most of the cases and our system is competitive with a classic one. The majority of the non-detection cases occur when we test our system on very rare - and non classical - music style (for example with bagpipe), which can be considered as anecdotal or when the singing is almost masked by the instrumental part. The singing which is not detected as singing is almost never classified as speech (less than 1% of the singing is classified as speech). It is classified either as pure music (3/4) or as noise (1/4) (no component present at this instant).

Our false alarm rate (instruments recognised as singing) is low: 8.5%, to be compared to the one from the GMM system: 19.6%. In our system, these errors are due to instruments such as pan flute or accordion which can have a vibrato.

5. CONCLUSION

In this paper, we presented a method for the detection of the singing voice, based on two simple parameters: the vibrato and the harmonic coefficient, and an original segmentation of the signal, which is made through a temporal and frequential analysis of the spectrogram.

Coupled with the parameters from the Speech/Music classifier [4], the information extracted from the signal allows us to know which component are present: speech, music, singing voice, any combination or none of them.

The performances of our system are comparable to those from a classic system based on MFCC and GMM. The advantage of our is that it does not need any learning, the detection is based on robust parameters and can be applied to any audio excerpt.

Our work will now be to improve the singing detection. We will analyse the possibilities of making our system more robust by improving extraction of parameters and the temporal segmentation. We will also study if combining our system with a classic one could improve both of them.

REFERENCES

[1] E. Scheirer and M. Slaney, "Construction and evaluation of a robust multifeature speech/music discriminator," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1997, vol. 2, pp. 1331–1334.

[2] M. Karjalainen and T. Tolonen, "Multi-pitch and periodicity analysis model for sound separation and auditory scene analysis," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1999, vol. 2, pp. 929–932.

[3] I. Arroabarren, M. Zivanovic, X. Rodet, and A. Carlosena, "Instantaneous frequency and amplitude of vibrato in singing voice," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2003, vol. 5, pp. 537–540.

[4] J. Pinquier, J.L. Rouas, and R. Andre-Obrecht, "A fusion study in speech / music classification," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2003, vol. 2, pp. 17–20.

[5] Toru Taniguchi, Akishige Adachi, Shigeki Okawa, Masaaki Honda, and Katsuhiko Shirai, "Discrimination of Speech, Musical Instruments and Singing Voices Using the Temporal Patterns of Sinusoidal Segments in Audio Signals," in *Interspeech - European Conference on Speech Communication and Technology*. ISCA, 2005, pp. 589–592.

[6] Wu Chou and Liang Gu, "Robust Singing Detection in Speech/Music Discriminator Design," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2001, vol. 2, pp. 865–868.

[7] I. Arroabarren and A. Carlosena, "Voice production mechanisms of vocal vibrato in male singers," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 1, pp. 320–332, Jan 2007.

[8] R. Timmers and P. Desain, "Vibrato: questions and answers from musicians and science," in *Proc. Int. Conf. on Music Perception and Cognition*, 2000.

[9] David B. Gerhard, "Perceptual Features for a Fuzzy Speech-Song Classification," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2002, vol. 4, pp. 4160–4163.

[10] Y.D. Cho, M.Y. Kim, and S.R. Kim, "A spectrally mixed excitation (SMX) vocoder with robust parameter determination," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1998, vol. 2, pp. 601–604.

[11] R. André-Obrecht, "A new statistical approach for the automatic segmentation of continuous speech," *IEEE Transaction on Acoustics, Speech, and Signal Processing*, vol. 36, pp. 29–40, 1988.