SINGING VOICE DETECTION IN MONOPHONIC AND POLYPHONIC CONTEXTS

Hélène Lachambre, Régine André-Obrecht, Julien Pinquier

IRIT - Université de Toulouse
118 route de Narbonne, 31062 Toulouse Cedex 9, France
{lachambre, obrecht, pinquier}@irit.fr

ABSTRACT

In this article, we present an improvement of a previous singing voice detector. This new detector is in two steps.

First, we distinguish monophonies from polyphonies. This distinction is based on the fact that the pitch estimated in a monophony is more reliable than the one estimated in a polyphony. We study the short term mean and variance of a confidence indicator; their repartition is modelled with bivariate Weibull distributions. We present a new method to estimate the parameters of these distributions with the moment method.

Then, we detect the presence of singing voice. This is done by looking for the presence of vibrato, an oscillation of the fundamental frequency between 4 and 8 Hz. In a monophonic context, we look for vibrato on the pitch. In a polyphonic context, we first make a frequency tracking on the whole spectrogram, and then look for vibrato on each frequency tracks.

Results are promising: from a global error rate of 29.7 % (previous method), we fall to a global error rate of 25 %. This means that taking into account the context (monophonic or polyphonic) leads to a relative gain of more than 16 %.

1. INTRODUCTION

Our work takes place in the general context of music description and indexation. In this process, many steps are necessary, including melody extraction, instruments, genre, artist, or singer identification. For all these tasks, it can be useful to have a precise information about the presence or absence of singing voice.

The singing voice detection has been a research subject for around 10 years, it is a relatively recent subject. Recent work have been conducted to find the best features to describe singing voice [1, 2, 3]. Other works have been addressing more specific music style contents [4], or have been interested in the accompanied singing voice detection [5].

In a previous work [6], we presented a singing voice detector based on the research of vibrato on the harmonics of the sound: we made the tracking of the harmonics present in the signal, and we looked for the presence of vibrato on each harmonic tracking. In this work, we propose to first separate monophonies from polyphonies. Since this classification is very efficient, we aim at taking advantage of this knowledge to improve the singing voice detection, which is still based on the research of vibrato. The whole process is summarized in figure 1.

The monophony/polyphony classifier is based on the fact that the estimated pitch is more reliable in the case of a monophony than in the case of a polyphony. We analyse a confidence indicator issued from the YIN pitch estimator [7]. The classification process uses bivariate Weibull models. We present a new method to estimate their parameters.

Then the singing voice detection is differentiated, depending of the results of the first step. We still look for vibrato on the frequency tracking in polyphonic context but, in monophonic context, we look for vibrato on the pitch.

In parts 2 and 3, we describe respectively the monophony/polyphony classifier and the previous singing voice detector. In part 4, we present the adaptation of the singing voice detector to each case: monophonic and polyphonic. Then in part 5, we present our corpus, experiments and results. Finally we conclude and give some perspectives in part 6.

2. MONOPHONY/POLYPHONY CLASSIFICATION SYSTEM

2.1 Parameters

In [7], de Cheveigné and Kawahara present a pitch estimator named YIN. This estimator is based on the computing of the difference function $d_t(\tau)$ over each signal frame t:

$$d_t(\tau) = \sum_{k=1}^{N} (x_k - x_{k+\tau})^2 \quad (1)$$

with x the signal, N the window size and τ the shift time.

For a periodic signal, its period T should be given by the first zero of $d_t(\tau)$. This is not always possible, notably due to imperfect periodicity [7]. The authors propose to use instead the Cumulative Mean Normalised Difference:

$$d'_t(\tau) = \begin{cases} 1 & if\ \tau = 0 \\ d_t(\tau) / \left[{}^1\!/\!_\tau . \sum_{k=1}^{\tau} d_t(k) \right] & otherwise \end{cases} \quad (2)$$

The pitch is given by the index T of the minimum of $d'_t(\tau)$. The authors precise that the lower $cmnd(t) = d'_t(T)$ is, the more confident the estimation of T is. So we consider $cmnd(t)$ as a confidence indicator.

In the case of a monophony, the estimated pitch is confident, so $cmnd(t)$ is low and do not vary much. *A contrario*, in the case of a polyphony, the estimated pitch is not reliable: $cmnd(t)$ is higher and varies much more. These considerations lead us to use the two following parameters: the short term mean and variance of $cmnd(t)$, noted $cmnd_{mean}(t)$ and $cmnd_{var}(t)$, computed over 10 frames centred on frame t.

© EURASIP, 2009

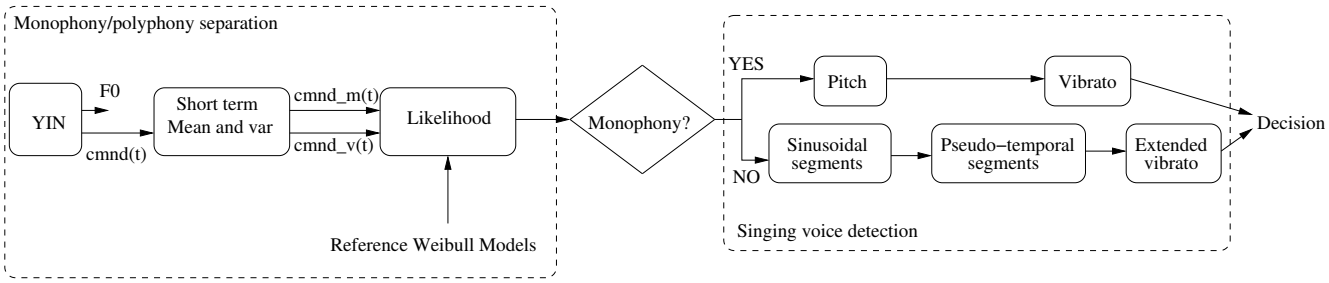


Figure 1: Global scheme of the system.

2.2 Modelling

The bivariate repartition of $(cmnd_{mean}, cmnd_{var})$ is modelled with the bivariate Weibull distributions proposed in [8]:

$$F(x,y) = 1 - \exp\left(- \left[\left(\frac{x}{\theta_1} \right)^{\frac{\beta_1}{\delta}} + \left(\frac{y}{\theta_2} \right)^{\frac{\beta_2}{\delta}} \right]^{\delta} \right) \quad (3)$$

for $(x,y) \in \mathbb{R}^+ \times \mathbb{R}^+$, with $(\theta_1, \theta_2) \in \mathbb{R}^+ \times \mathbb{R}^+$ the scale parameters, $(\beta_1, \beta_2) \in \mathbb{R}^+ \times \mathbb{R}^+$ the shape parameters and $\delta \in]0,1]$ the correlation parameter.

To estimate the five parameters $(\theta_1, \theta_2, \beta_1, \beta_2, \delta)$ of each bivariate distribution, we use the moment method. The moments are given by Lu and Bhattacharyya in [9]:

$$E[X] = \theta_1 \Gamma(1/\beta_1 + 1) \quad (4)$$

$$E[Y] = \theta_2 \Gamma(1/\beta_2 + 1) \quad (5)$$

$$Var(X) = \theta_1^2 \left(\Gamma(2/\beta_1 + 1) - \Gamma^2(1/\beta_1 + 1) \right) \quad (6)$$

$$Var(Y) = \theta_2^2 \left(\Gamma(2/\beta_2 + 1) - \Gamma^2(1/\beta_2 + 1) \right) \quad (7)$$

$$\begin{aligned} Cov(X,Y) = \theta_1 \theta_2 . \\ [\Gamma(\delta/\beta_1 + 1)\Gamma(\delta/\beta_2 + 1)\Gamma(1/\beta_1 + 1/\beta_2 + 1) \\ -\Gamma(1/\beta_1 + 1)\Gamma(1/\beta_2 + 1)\Gamma(\delta/\beta_1 + \delta/\beta_2 + 1)] \\ \div \Gamma(\delta/\beta_1 + \delta/\beta_2 + 1) \end{aligned} \quad (8)$$

with $\Gamma(x)$ the gamma function.

From equations 4, 5, 6 and 7, we extract θ_1, θ_2, β_1 and β_2, the parameters of the two marginal distributions.

We have shown [10] that equation (8) is equivalent to the following equation:

$$f(\delta) = \delta B(\delta/\beta_1, \delta/\beta_2) = C \quad (9)$$

with $B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$ the Beta function and C a constant depending from θ_1, θ_2, β_1, β_2 and $Cov(X,Y)$, which are known.

From equation (9), finding δ is equivalent to finding the zeros of the expression $f(\delta) - C$. As we have shown that $f(\delta)$ is strictly decreasing function [10], we may find easily the unique zero by dichotomy.

2.3 Classification and results

A bivariate Weibull model is learned for each class, on the training corpus (described in part 5.1). These models are thereafter named the reference models.

The classification is done every second. Since $cmnd_{mean}(t)$ and $cmnd_{var}(t)$ are computed every 10 ms, we have 100 2-dimension vectors every second. The decision is taken by computing the likelihood of 100 consecutive vectors (1 second) to each reference model. The assigned class is the one which maximizes the likelihood.

Results given by this method are very good: we have a global error rate of 6.3 % on the corpus presented in part 5.1. This is why we take this method as a preprocessing stage before looking for the presence of singing voice.

3. SINGING VOICE DETECTION

3.1 Vibrato

Vibrato is a well-known [11, 12] property of the human singing voice. In general, the vibrato is defined as a periodic oscillation of the fundamental frequency. In the specific case of the singing voice, this oscillation is at a rate of 4 to 8 Hz. So, on a given frequency tracking vector F, the presence of vibrato is confirmed if there is a maximum between 4 and 8 Hz in the Fourier Transform of F.

In our work, we consider monophonic and polyphonic extracts, so the research of the vibrato on the fundamental frequency is not always possible. However, we note that if the vibrato is present on the fundamental frequency, it is as well present on its harmonics. This is why we make a tracking of all the harmonics present in the signal (see section 3.2). We then look for the presence of vibrato on these harmonics.

3.2 Sinusoidal and Pseudo-temporal segmentations

3.2.1 Sinusoidal segmentation

The tracking of the harmonics (see figure 2), thereafter named "sinusoidal segments", is done with the method described in [13].

The algorithm is the following one:

- compute the spectrum every 10 ms, with a 20 ms Hamming window,
- convert the frequency in cent (100 $cent = 1/2\ tone$), and smooth it with a 17 cent window,
- detect the maxima of the spectrum: the frequencies $(f_t^i, i = 1, ..., I)$ and their log amplitude $(p_t^i, i = 1, ..., I)$,

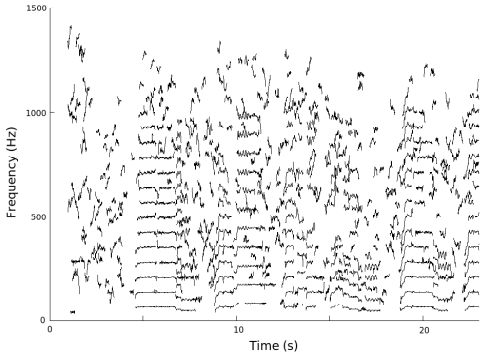
- compute the distance between each pair of consecutive maxima (at the instant t and $t-1$):

$$d_{i_1,i_2}(t) = \sqrt{\left(\frac{f_t^{i_1} - f_{t-1}^{i_2}}{C_f}\right)^2 + \left(\frac{p_t^{i_1} - p_{t-1}^{i_2}}{C_p}\right)^2} \quad (10)$$

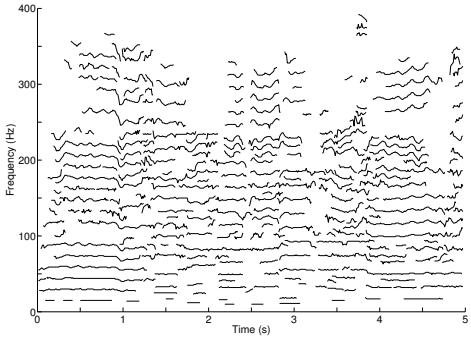
- Two points $(t, f_t^{i_1})$ and $(t+1, f_{t+1}^{i_2})$ are connected if $d_{i_1,i_2}(t) < d_{th}$.

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 [13]).

A set of consecutive connected points forms a sinusoidal segment.



a) Solo singer



b) Polyphonic song

Figure 2: Examples of sinusoidal segmentations for a 23 s extract of a monophonic song a Capella (a) and a 5 s polyphonic song a Capella (b): each curve is a sinusoidal segment.

As we want to work on all the harmonics, we have introduced a pseudo-temporal segmentation, which aims at grouping the temporaly related sinusoidal segments.

3.2.2 Pseudo-temporal segmentation

We assume that in music, for a given note, the fundamental frequency and its harmonics begin and end approximately at the same time. Therefore, we analyse the temporal relations between the beginnings and the ends of sinusoidal segments [6].

Then a limit of a segment is placed at instant t if the two following conditions are respected:

- There are at least 2 extremities of sinusoidal segments at the instant t.
- There are at least 3 beginnings or 3 ends between instants t and $t+1$.

The resulting segmentation is presented on figure 3.

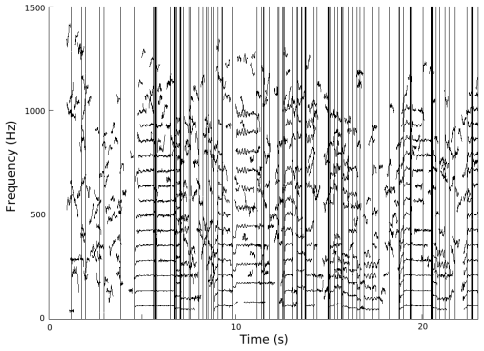


Figure 3: Temporal segmentation of the extract of figure 2 a): the vertical lines are the temporal limits of the segments.

3.3 Extended vibrato

In our previous work [6], we defined the *extended vibrato*, *vibr*, as follows:

$$vibr = \frac{\sum_{s\in\Gamma} l(s)}{\sum_{s\in\Omega} l(s)} \quad (11)$$

with:

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

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

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

vibr characterises each pseudo-temporal segment. For too short pseudo-temporal segments, $vibr = 0$: it is not possible, considering the length of the concerned sinusoidal segments, to determine wether there is or not vibrato. The same value *vibr* is attributed to each frame of the pseudo-temporal segment.

In [6], *vibr* was averaged over one second. The final decision was taken by thresholding the 1-second-averaged *vibr*: a high value (> 0.3) meant the presence of singing voice, while a low value indicated its absence.

4. FUSION

A global remark that can be done on the singing voice detection is that a singer does not sing all the time: there are very short pauses (less then 1/2 second) for respirations. And in the case of a polyphony, there are also long pauses (up to 1 minute or more) for instrumental parts, and short pauses (1 to 3 seconds) with instrumental transitions.

This makes us change our decision strategy (both in monophonic and polyphonic cases): the decision is still taken every second, but instead of thresholding a value computed over one second, the detection of singing voice is now based on the fact that at one (or more) instant of one second, we detect vibrato.

4.1 Monophonic case

In the monophonic case, the estimated pitch is reliable (see section 2). Its value is even more reliable than the sinusoidal segments: the pitch found is either the fundamental frequency or one of its harmonics, so if there is vibrato on the real pitch, it is also present on the estimated pitch. On the other hand, the sinusoidal segments mostly correspond to harmonics of the fundamental frequency, but there are some intruders which are no harmonics. The presence of vibrato on these intruders is not linked to the presence of vibrato on the pitch. This is why our search of vibrato is made on the pitch in the monophonic case.

We first segment the pitch into notes: a transition between two consecutive notes is found if the pitch jumps for more than 1/2 tone. Then, on each note, we look for the presence of vibrato. A second is classified as containing singing voice if there is vibrato on at least 10% of the frames.

4.2 Polyphonic case

In the polyphonic case, there is a strong overlap between the frequencies of the different instruments/singers performing. The estimated pitch is not reliable, it does not correspond to any instrument or singer present. However, we can assume that each sinusoidal segment corresponds to the partial of one instrument. We also assume that the different sinusoidal segments at a given time can come from different instruments.

The value *vibr* is short term averaged over 500 ms (corresponding to 50 values of *vibr*), giving $vibr_{average}$. The decision process is the following: there is singing voice if, during 1 second, there is at least one instant for which $vibr_{average} > 0.15$.

Note that the preprocessing step (separating monophonies) allows us to have a different threshold than in [6], which is more robust and adapted to the polyphonic case.

5. EXPERIMENTS AND RESULTS

5.1 Corpus

Our corpus is home made, and contains monophonic and polyphonic music, with and without singing voice. We have either single instruments or single singers performing for the monophonies, and either multiple instruments, or singers with instrumental background performing for the polyphonies.

The corpus contains extracts from approximately 50 musical recordings, so we consider that the results are not dependant from the recording condition.

The corpus contains also various styles (rock, opera, renaissance, jazz, country...), various instruments (piano, violin, recorder, cello, guitar, brass...), and 12 different singers. Some styles, instruments, and singers are present in the test set and are not in the training set. This allows us to consider also that the results are not dependant from the music type, instrument, and singer.

Table 1: Corpus repartition (training and test sets).

Class	Train Duration	Test Duration	Nb. of tests
Single instrument	25 s	2 min 57 s	177
Single singer	25 s	4 min 38 s	278
Monophony	**50 s**	**7 min 35 s**	**455**
Multiple instruments	25 s	3 min 23 s	203
Singers and instruments	25 s	3 min 10 s	190
Polyphony	**50 s**	**6 min 33 s**	**393**
Total	**2 min 5 s**	**18 min 41 s**	**1121**

The composition of the corpus is summarized in table 1 in terms of duration and of number of test (seconds) for each class.

5.2 Results with a handmade monophony/polypony classification

The previous method, in which we did not distinguish monophonies from polyphonies, gave a global error rate of **29.7 %**.

First, we propose to consider a handmade monophonic/polyphonic classification, in order to evaluate our proposal to differentiate the decision process and to obtain a superior limit of the performances.

The results of this experiment are presented in table 2 for the monophonic case, and in table 3 for the polyphonic case.

The global error rate is **21.7 %**.

Table 2: Confusion matrix - Monophonic context.

	Singing voice	No singing voice
Single singer	**0.83 %**	0.17 %
Single instrument	0.20 %	**0.80 %**

Table 3: Confusion matrix - Polyphonic context.

	Singing voice	No singing voice
Singers and instruments	**0.66 %**	0.34 %
Multiple instruments	0.16 %	**0.84 %**

We first note that having a knowledge on the number of sources performing (one or more) improves the singing voice detection in every cases. The global error rate is improved by more than 8.5 %.

As we could have predicted, it is more difficult to detect the presence of singing voice in polyphonic context. The missed occurrences are due to low singing voice compared to the accompagning instruments.

In the monophonic context, the false alarms are due to wind instruments, on which the performer voluntarily produces a vibrato as a musical effect.

5.3 Results with an automatic monophony/polypony classification

We conduct a second experiment, to evaluate the whole system: the monophonic/polyphonic decision is taken with the method presented in part 2. Then, depending of the output of the first system, the singing voice detection is run with either one or the other method.

The results are presented in table 4. The global error rate is now **25 %**. Even with the imprecision introduced by the monophonic/polyphonic detector, the singing voice detection is still better than in the previous system, since it is improved by more than 5 %.

Table 4: Confusion matrix - Whole system.

	Singing voice	No singing voice
Single singer	**0.79 %**	0.21 %
Single instrument	0.26 %	**0.74 %**
Singers and instruments	**0.65 %**	0.35 %
Multiple instruments	0.18 %	**0.82 %**

We see that the singing voice is always more difficult to detect in polyphonic context. Most errors are due to the same causes than in the previous experiment, adding the imprecision of preprocessing step.

6. CONCLUSION AND PERSPECTIVES

In this article, we presented an improvement of our singing voice detector, based on a differentiated strategy depending of the monophonic/polyphonic character of the music. We first presented the monophonic/polyphonic classifier, then the two strategies we adopt to detect the presence of singing voice. The singing voice is detected by the presence of vibrato either on the pitch (monophonic context) or on the frequency trackings (polyphonic context).

The method we propose needs a training phase which is done once for all, to learn the parameters of the models and the classification thresholds. All these parameters are then considered universal. This assumption is confirmed by the very little size of the training set: 50 seconds of monophonic signal and 50 seconds of polyphonic signal.

The results are very promising, since the global error rate falls from 29.7 % to 21.7 % with a handmade separation between monophonic and polyphonic. When the whole system is used, the global error rate is still improved since it falls to 25 %.

Our next work will be now to improve the detection of the singing voice in polyphonic context, which is at the moment the more critical case.

This system could be used in many applications. It can for example be added to speech/music separation systems. In these systems, the singing voice is the cause of many errors, which are mainly solo singing voice recognised as speech. An other application is the music structuration: some parts of a song, such as the chorus, almost always contain singing voice, whereas others, such as transitions, or interludes do not.

REFERENCES

[1] M. Rocamora and P. Herrera, "Comparing audio descriptors for singing voice detection in music audio files," in *Brazilian Symposium on Computer Music, 11th. San Pablo, Brazil*, 2007.

[2] A. Mesaros and Moldovan S., "Method for singing voice identification using energy coefficients as features," in *IEEE International Conference on Automation, Qality and Testing, Robotics*, 2006, pp. 161–166.

[3] N.C. Maddage, Kongwah Wan, Changsheng Xu, and Ye Wang, "Singing voice detection using twice-iterated composite fourier transform," in *Multimedia and Expo, 2004. ICME '04. 2004 IEEE International Conference on*, 2004, vol. 2, pp. 1347–1350.

[4] S.Z.K. Khine, T. L. Nwe, and H. Li, "Singing voice detection in pop songs using co-training algorithm," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2008, pp. 1629–1632.

[5] S. Santosh, S. Ramakrishnan, Vishweshwara Rao, and Preeti Rao, "Improving singing voice detection in presence of pitched accompaniment," in *Proc. of the National Conference on Communications (NCC)*, 2009.

[6] H. Lachambre, R. André-Obrecht, and J. Pinquier, "Singing voice characterization for audio indexing," in *15th European Signal Processing Conference (EUSIPCO)*, 2007, pp. 1563–1540.

[7] A. de Cheveigné and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *Journal of the Acoustical Society of America*, vol. 111, no. 4, pp. 1917–1930, April 2002.

[8] P. Hougaard, "A class of multivariate failure time distributions," *Biometrika*, vol. 73, no. 3, pp. 671–678, 1986.

[9] J.C. Lu and G.K. Bhattacharyya, "Some new constructions of bivariate Weibull models," *Annals of Institute of Statistical Mathematics*, vol. 42, no. 3, pp. 543–559, 1990.

[10] H. Lachambre, "Estimation of Weibull bivariate distribution parameters via the moment method," Tech. Rep., IRIT - SAMoVA, Dec 2008.

[11] 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.

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

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