

PARTIAL CLUSTERING USING A TIME-VARYING FREQUENCY MODEL FOR SINGING VOICE DETECTION

L. Regnier, G. Peeters *

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

ABSTRACT

We propose a new method to group partials produced by each instrument of a polyphonic audio mixture. This method works for pitched and harmonic instruments and is specially adapted to singing voice. In our approach, we model time-varying frequencies of partials as a slowly varying frequency plus a sinusoidal modulation. The parameters obtained with this model plus some common Auditory Scene Analysis principles are used to define a similarity measure between partials. This multi-criterion based measure is then used to build the input similarity matrix of a clustering algorithm. Clusters obtained are groups of harmonically related partials. We evaluate the ability of our method to group partials per source when one of the sources is a singing voice. We show that partial clustering is a promising approach for singing voice detection and separation.

Index Terms— Singing voice detection, Source separation, Polyphonic music analysis, Vibrato detection.

1. INTRODUCTION

Singing voice melody and singer characteristics are some of the most memorable elements of a song, especially in rock and pop music where the voice always carries the main melody line. Many studies in Music Information Retrieval attempt to separate, transcribe or characterize the singing voice because these offer a large number of applications.

Singing voice separation and sung melody transcription are related problems: when the fundamental frequency of the voice is known *a priori*, the voice can be extracted by tracking frequencies in harmonic ratio to the fundamental ([1] and [2]). Auditory Scene Analysis (ASA)[3] is an alternative to separate sources using principles and constraints present in the natural auditory system. ASA type cues used by Mellinger [4] to group fragments of the spectrum originating with a single source are: common onset, beginning of sound energy, harmonicity and several others. Srinivasan [5] also use harmonicity, dynamics and onset to group part of the spectrum into auditory blobs. Other approaches are based on statistical methods such as Independent Component Analysis (ICA) and Non-negative Matrix Factorization (NMF).

In our approach, we propose to group partials produced simultaneously by each source without any prior knowledge of fundamental frequency, model of instrument or model of the auditory system. In the case of pitched and harmonic instruments, it consists in tracking groups of harmonically related partials. The motivation resides in the fact that these groups carry more source specific cues than individual partials. Our model is specially adapted to separate the

lead vocal from the instrumental background within a song. The only considerations taken into account are some characteristics of the pitch of the singing voice: vibrato and portamento. To highlight these characteristics, time-varying frequencies of partials are modeled as a slowly varying frequency (portamento) plus a sinusoidal modulation (vibrato) as explained in section 2. Parameters given by this model, plus some ASA principles (onset and harmonicity) are used to build a similarity measure between partials presented in section 3. This measure is then used to group harmonically related partials with the help of a usual clustering algorithm. The ability of parameters, the similarity measure and clustering to group partials into sources is evaluated when one of the sources is a singing voice. The last section provides conclusions.

2. MODEL OF TIME-VARYING FREQUENCIES FOR SINGING VOICE

2.1. Characteristics of the pitch of singing voice

The pitch of the singing voice varies considerably over the duration of a note compared to most pitched instruments. We distinguish two kinds of pitch variations, both representative of Western melodic voices, and perceived as expressive attributes used to highlight voice-leading.

Vibrato (Figure:2): Vibrato is one of the most distinctive element of the singing voice. It refers to the periodic variation of pitch characterized by a rate and an extent (or depth) [6]. For the singing voice, the average vibrato rate is around 6 Hz and its extent ranges from 60 to 200 cents (100 cents = 1 semitone).

Portamento or Legato (Figure:3): Portamento and legato are smooth transitions from one note to another without interrupting sound. In both cases, it is a slow and smooth variation of the pitch.

2.2. Fundamental frequency model

We model the frequency $f(t)$ of a sung tone, of mean \bar{f} , by a frequency slowly varying through time $d_f(t)$ (representing the portamento) plus a periodic modulation $s(t)$ (representing the vibrato) [7]. In this model, we suppose the vibrato rate r and vibrato extent A are constants.

$$f(t) = \bar{f} \cdot (d_f(t) + s(t)) + \epsilon(t), \text{ where} \quad (1)$$

$$s(t) = A \cdot \sin(2\pi r t + \phi_0), \quad (2)$$

$d_f(t)$ is a low-order polynomial and $\epsilon(t)$ is the modeling error.

2.2.1. Computation of parameters

• The **mean frequency** \bar{f} is given by the mean of $f(t)$.
In order to get equivalent values for two partials with frequencies

*This work was supported by the French project Oseo “Quaero” and as a part of “DISCO”, funded by ANR.

harmonically related ($f_i(t) = n \cdot f_j(t), n \in \mathbb{N}^*$), parameters are computed on $f(t)/\bar{f}$.

- The quantity $f(t)/\bar{f}$ is low-pass filtered with a cutoff frequency $f_c = 4\text{Hz}$. The result of the filtering process is a curve representing the **relative frequency variation** $d_f(t)$ parameterized with a third-order polynomial $P(X) \in \mathbb{R}[X]$. We denote $P_{df}(t)$ the estimation of $d_f(t)$ obtained by the evaluation of $P(X)$.

- The sinusoidal component $s(t)$ is obtained by subtracting the relative frequency deviation $d_f(t)$ from the relative frequency: $s(t) = f(t)/\bar{f} - d_f(t)$. Let be $S(\varphi)$ the Fourier transform of $s(t)$ and $\omega = \arg \max_{\varphi} S(\varphi)$. The **vibrato rate** is given by $r = 2\pi\omega$. The

relative vibrato extent is given by $A = S(\omega)$ as shown on Figure 1. This value can be converted into cents: $A_{cents} = 1200 \cdot \log 2(A+1)$. Finally, the **phase at the origin** is given by: $\phi = \arcsin(s(0)/A)$ if $\frac{\delta}{\delta t} f(0) > 0$, $\phi = \pi - \arcsin(s(0)/A)$ otherwise.

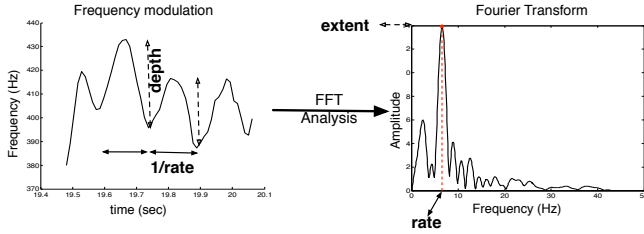


Fig. 1. (Left) $s(t)$: modulation. (Right) rate and extent on $S(\omega)$.

The estimation of $f_p(t)$ the time-varying frequency of partial p is noted $\hat{f}_p(t)$. Thus we have:

$$f_p(t) = \hat{f}_p(t) + \epsilon_p \text{ with} \quad (3)$$

$$\hat{f}_p(t) = \bar{f}_p \cdot (P_{df,p}(t) + A_p \cdot \sin(2\pi r_p t + \phi_p)) \quad (4)$$

2.2.2. Model validation

The proposed model has been evaluated on the set of partials described in section 4.1. We compute the relative absolute modeling error ϵ_{rel} and the root mean square error ϵ_{RMS} , on a frequency of length T , using:

$$\epsilon_{rel} = \frac{1}{T} \sum_{t=1}^T \frac{|f(t) - \hat{f}(t)|}{f(t)}, \epsilon_{RMS} = \sqrt{\frac{1}{T} \sum_{t=1}^T (f(t) - \hat{f}(t))^2}$$

We have obtained the following error values: $\epsilon_{rel} = 0.81\%$ and $\epsilon_{RMS} = 17.61 \text{ Hz}$.

Figures 2 and 3 show respectively the modeling results obtained on a sung tone with vibrato and a sung portamento.

According to the very low modeling error rate, the proposed model estimates with a high precision the parameters of vibrato as well as the portamento. These parameters are used on one hand to discriminate singing partials as in [8], and on the other hand to compare partials.

3. PARTIAL CLUSTERING

In this section, we apply the previous model (see eq(1)) on time-varying frequencies of partials. Partial is extracted using pm2 [9]. The goal of the clustering is to group partials produced by the same

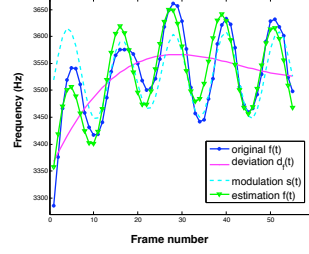


Fig. 2. Vibrato.

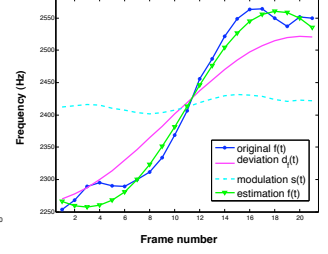


Fig. 3. Portamento.

source at a given time. Working with harmonic instruments, it consists of grouping harmonically related partials. For this we first define a measure of dissimilarity between partials based on ASA principles. This multi-criterion based measure is then used to build the input similarity matrix of the clustering algorithm.

3.1. Partial features and similarity

The goal is to define a dissimilarity measure between two partials p and q that is close to zero when partials are harmonically related at a given time. Partial is first compared on a set of K criteria using comparison functions $\phi_i(p, q)$, for i from 1 to K . The dissimilarity measure is given by the aggregation of the comparison functions $m(p, q) = \psi(\phi_1(p, q), \dots, \phi_K(p, q))$ where ψ is the aggregation function.

3.1.1. Comparison functions

The time interval of partial p is noted: $I_p = [beg(p), \dots, end(p)]$ with $beg(p)$ and $end(p)$ the starting and ending frame of p . $|I_p|$ denotes its length. Later, we compare two partials p and q on their shared part $I_{p,q} = I_p \cap I_q$ of length T .

Strict dissimilarity based on time extent partials

Partials are compared on their shared part $I_{p,q}$ so that we set up a strict dissimilarity based on the length $I_{p,q}$:

$$\phi_1(p, q) = 1 - \frac{|I_{p,q}|}{\max(|I_p|, |I_q|)}$$

The strict dissimilarity leads to:

$$\phi_1(p, q) = 1 \Rightarrow m(p, q) = 1.$$

Dissimilarity based on onset

Unrelated sounds seldom start or stop at exactly the same time. We define a criterion based on the length of the gap between the two starting frames:

$$\phi_2(p, q) = \frac{|beg(p) - beg(q)|}{\max(|I_p|, |I_q|)}$$

Dissimilarity based on frequency variation 1: Vibrato

We compare the characteristics of vibrato (rate and relative extent) computed on $I_{p,q}$ using eq(1):

$$\phi_3(p, q) = \frac{1}{2} \left(\frac{|A_p/f_{0p} - A_q/f_{0q}|}{\max(A_p/f_{0p}, A_q/f_{0q})} + \frac{|r_p - r_q|}{r_{max}} \right)$$

Where r_{max} is the maximum vibrato rate depending on the partial analysis frequency. For example, with an hop size of 0.0116 sec, $r_{max} = \frac{1}{2} \cdot \frac{1}{0.0116} = 43 \text{ Hz}$.

Dissimilarity based on frequency variation 2: Portamento

Relative frequency deviation of p and q on $I_{p,q}$ are estimated with third-order polynomials ($P_p(x)$ and $P_q(x)$ respectively) with real coefficients in $[0, 1]$. The dissimilarity based on relative slow frequency variation is defined as follow:

$$\phi_4(p, q) = \frac{1}{4} \sum_{x=0}^3 |P_p(x) - P_q(x)|$$

Dissimilarity based on harmonicity

The more p and q are harmonic, the more the values of $\frac{f_p(t)}{f_q(t)} = h(t)$ are constants for $t = 1 \dots T$. We define the harmonicity dissimilarity as the standard deviation of h .

$$\phi_5(p, q) = \begin{cases} \sigma\left(\frac{f_p(t)}{f_q(t)}\right) & \text{if } \bar{f}_p > \bar{f}_q \\ \sigma\left(\frac{f_q(t)}{f_p(t)}\right) & \text{otherwise} \end{cases}$$

3.1.2. Multi-criterion aggregation and dissimilarity measure

The aggregation is done using a linear opinion pool [10] if the first comparison function is not equal to one:

$$m(p, q) = \psi(\phi_1, \dots, \phi_K) = \begin{cases} 1 & \text{if } \phi_1(p, q) = 1 \\ \sum_{i=1}^K w_i \cdot \phi_i(p, q) & \text{otherwise} \end{cases}$$

The following weights: $w_i = [1/9, 1/9, 2/9, 2/9, 2/9]$ have been chosen empirically.

3.1.3. Matrix of dissimilarity

The dissimilarity matrix is obtained by computing $m(p, q)$ on all pairs of partials.

3.2. Clustering

The clustering is done using an agglomerative hierarchical algorithm. The distance between any two clusters is taken to be the average of all distances between pairs of partials (average linkage method), resulting in clusters with close variance. If the task were to find all partials produced by one instrument along the song duration, the single linkage method would have been more appropriate. On the other hand, the complete linkage method produces very specific classes and does not fit our problem.

In Figure 4, we show the results of the application of our method on a real signal with three sources: voice, piano and guitar. Each cluster is represented by a different line type.

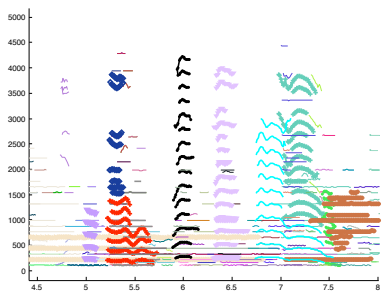


Fig. 4. Example of partial clustering obtained on a real signal

4. EVALUATION

The goal of this study is to evaluate if the use of the proposed parameters, measure of dissimilarity and clustering provide clusters containing a single source. For this, we first evaluate the cluster purity, then we evaluate the impact of clustering on a singing voice partial discrimination task.

4.1. Test-set

The material used in this investigation is a collection of separated multitrack recordings of 15 melodic songs chosen for their variety in artists and music genre. From these, we created our test-data as follows: 1) For each song, partials from the lead vocal, the piano, the guitar and/or bass tracks are extracted on the mono-instrumental tracks. 2) When considering all instruments, instrumental partials are always more numerous than singing partials. For each song, partials from one or two instruments are chosen so that there are the same number of instrument partials as singing partials. 3) Balanced sets are merged into a global set of partials described in Table 1.

	Voice	Instruments		
	Voice	Guitar	Piano	Bass
Label	1	-1	-2	-3
Number of partials	56289	19092	25460	11737
Number of partials	56289	56289		

Table 1. Test-set description.

4.2. Clusters purity

Theoretically, each cluster is constituted by partials coming from the same source at the same time. Thus, labels of partials within a cluster should all be equal. We measure the quality of a cluster according to the original class label with the cluster purity. Cluster purity of cluster C_i of size $|C_i|$ is defined as:

$$purity(C_i) = \frac{1}{|C_i|} \cdot \max_j (|C_i|_{class=j}) \quad (5)$$

Where $|C_i|_{class=j}$ corresponds to the number of elements in C_i classified as belonging to class j . The overall purity of a clustering solution, with I clusters, is expressed as a weighted sum of individual cluster purities:

$$purity = \frac{1}{|C|} \sum_{i=1}^I |C_i| \cdot purity(C_i) \quad (6)$$

When considering two classes (singing and instruments) the overall purity is found to be 0.8944. When four classes are considered (cf. Table 1), the overall purity is 0.7802. We can conclude that partial clustering is efficient to group partials from source.

4.3. Singing partials detection

The problem can be stated as follows: for a given set of partials, find the partials produced by the singing voice. According to singing pitch characteristics, presented in 2.1, most of singing partials have a vibrato with a large extent and a rate around 6 Hz, or an important frequency variation. We first review the method proposed in [8], to locate the singing voice within a song and indicate the problems encountered with it. Then, we show how the partial clustering method proposed in this paper can be used to reduce these problems.

To detect singing partials, in [8], we apply thresholds on the values of vibrato rate ($\tau_{\Delta r}$), vibrato extent (τ_A) and frequency deviation ($\tau_{\Delta df}$). Thresholds leading to the best F-measure (76.52%) have been learnt using a ten-fold cross-validation. The mean values of the three thresholds over the ten folds are:

$\tau_{\Delta r} = 2Hz$. If we consider the mean of singing vibrato rate is equal to 6 Hz, this threshold value leads to: $r \in [4Hz, 8Hz]$. $\tau_A = 0.9\%$, equivalent to 16 cents. $\tau_{\Delta df} = 0.05$, equivalent to 86 cents where $\tau_{\Delta df}$ is taken to be the maximum distance between the peaks of $P_{df}(t)$, the polynomial evaluated.

In general, the Precision measure is higher when $\tau_{\Delta r}$, τ_A and $\tau_{\Delta df}$ are larger and correspondingly the Recall measure is lower. Observations of the obtained results indicate that short partials (shorter than two cycles of the vibrato), and partials with low f (due to the FFT resolution) are often missed. For this reason, we proposed to use groups of harmonically related partials since they carry more source information than individual partials. All partials composing a cluster are labeled as sung if a certain number x of partials within this cluster are detected as sung as illustrated in Figure 5. Sung partials are detected using the method of [8].

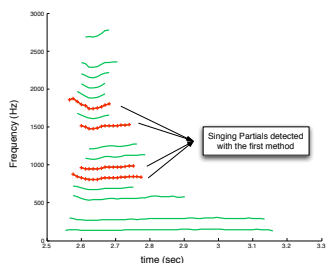


Fig. 5. One cluster of singing partials found using clustering coupled with the method proposed in [8].

4.3.1. Results and comments

Results obtained with the different approaches are compared in Table 2. The first line corresponds to results obtained without clustering.

x	F-measure	Precision	Recall
	76.52 %	85.45%	70.30%
≥ 1	77.23 %	66.97%	91.19%
≥ 2	79.02 %	75.33%	83.08%
≥ 3	77.91%	79.21%	76.65%

Table 2. Results

The first method discriminates singing partials from instrumental partials with high Precision. We conclude that vibrato and portamento are discriminant characteristics for identifying singing partials. However, the vibrato characteristics are not well evaluated on all partials. This is the reason why the Recall is relatively low. Using partial clustering increases considerably the Recall value. As shown in table 2, choosing $x = 2$ ensures a high Recall value of 83%. Increasing the Recall without affecting the Precision would be possible if the Precision given by the first approach was equal to 100% and if singing clusters were perfectly homogenous (purity = 1).

5. CONCLUSION AND FUTURE WORK

In this paper, we have proposed a new method to group, at a given time, partials produced by each instrument of a polyphonic audio

mixture. This method is evaluated when one of the sources is a singing voice. Portamento and vibrato are representative attributes of singing voice. In this investigation, we deal with the time-varying frequencies of partials modeled by a slowly varying frequency (portamento) plus a sinusoidal modulation (vibrato). We have proposed a method to estimate vibrato and portamento parameters with a high precision according to the very low modeling error rate ($< 1\%$).

Using these parameters combined with ASA principles, we have defined a dissimilarity measure between two partials. This multi-criterion based measure is used in an agglomerative hierarchical clustering algorithm to group harmonically related partials. We have evaluated the reliability of our approach on a partial test-set extracted from multitrack recordings. We evaluated the cluster purity, and showed that our method is efficient to group partials coming from the same source. Partial clustering has then been applied to the problem of singing partials detection. Vocal partials are discriminated using specificity of the singing voice: quasi systematic presence of vibrato with large extent, use of portamento and harmonicity. Singing partials are detected using thresholds on vibrato and portamento. An harmonicity criterion is introduced by using the obtained clusters. We have shown that using clusters of partials is a promising idea to improve singing voice partials detection within a mixture.

Finding all partials produced by the voice has applications in singing voice re-synthesis, singing extraction or sung melody transcription.

6. REFERENCES

- [1] A.L.C. Wang, *Instantaneous and frequency-warped signal processing techniques for auditory source separation*, Ph.D. thesis, Stanford University, CA, USA, 1994.
- [2] Y. Li and D.L. Wang, "Separation of singing voice from music accompaniment for monaural recordings," *IEEE Transactions on Audio, Speech, and Language Processing*, 2007.
- [3] A.S. Bregman, *Auditory scene analysis: The perceptual organization of sound*, MIT press, 1990.
- [4] D.K. Mellinger, *Event formation and separation in musical sound*, Ph.D. thesis, Stanford University Stanford, CA, USA, 1992.
- [5] SH Srinivasan and M. Kankanhalli, "Harmonicity and dynamics based audio separation," in *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03)*, 2003, vol. 5.
- [6] J. Sundberg, "Acoustic and psychoacoustic aspects of vocal vibrato," *Vibrato*, pp. 35–62, 1995.
- [7] R. Maher and J. Beauchamp, "An investigation of vocal vibrato for synthesis," *Applied Acoustics*, vol. 30, no. 2-3, pp. 219–45, 1990.
- [8] L. Regnier and G. Peeters, "Singing voice detection in music track using direct vibrato detection," in *ICASSP*, 2009.
- [9] A. R obel and M. Zivanovic, "Signal decomposition by means of classification of spectral peaks," *Proc. of the International Computer Music Conference (ICMC'04)*, pp. 446–449, 2004.
- [10] M. Stone, "The linear opinion pool," *Ann. Math. Statist.*, vol. 32, pp. 1339–1342, 1961.