Features extraction and temporal segmentation of acoustic signals

S. Rossignol[1,2], X. Rodet[1], J. Soumagne[2], J.-L. Collette[2] and P. Depalle[1]
rossigno@ese-metz.fr,rod@ircam.fr,soumagne@ese-metz.fr
[1]IRCAM 1, place Igor-Stravinsky, 75004 PARIS, FRANCE *
[2]Supélec 2, rue Édouard Belin, 57078 METZ, FRANCE

Abstract

This paper deals with temporal segmentation of acoustic signals and features extraction. Segmentation and features extraction are aimed at being a first step for sound signal representation, coding, transformation, indexation and multimedia. Three interdependent schemes of segmentation are defined, which correspond to different levels of signal attributes. A complete segmentation and features extraction system is shown. Applications and results on various examples are presented. The paper is divided into four sections: a definition of the segmentation schemes; and description of the techniques used for each of them.

1 – The three segmentation schemes

- The first scheme, named *source* scheme, concerns mainly the distinction between speech, singing voice and instrumental parts on movie sound tracks and on radio broadcasts. *Other* sounds, such as street sounds and machine noise, are also considered. The obtained data are propagated towards the following two schemes.
- The second scheme is refering to the *vibrato*. It is necessary to suppress vibrato from the source signal (singing voice and instrumental part only) to perform the following scheme.
- This last scheme leads to segmentation into *notes* or into *phones*, or more generally into stable or transient sounds according to the nature of the sound: instrumental part or singing voice excerpt or speech (when speech is identified on scheme 1, then there is no need to try to detect vibrato).

The data propagation between schemes ($1 \rightarrow 2 \rightarrow 3$ or $1 \rightarrow 3$) shows the dependencies between the results. Other segmentation characteristics (some are described in the features below) such as short silences/sound (energy transients), transitory/steady, voiced/unvoiced (voicing coefficient), harmonic (inharmonicity coefficient) and so forth will follow. Most of these characteristics are needed for getting notes and phones segmentation results on scheme 2.

2 – *source* segmentation scheme

This work is focusing on the identification of two sound classes. The two classes are: music (that is to say singing voice and/or instrumental part) and speech. Accordingly, features have been examined: they intend to measure distinct properties of speech and music. They are combined into several multidimensional classification frameworks. The preliminary results on real data are given here. Six features have been examined and retained [Scheirer and Slaney 97]. They are based on:

The spectral "flux", which is defined as the 2-norm of the difference between the magnitude of the Short Time Fourier Transform (STFT) spectrum evaluated at two successive sound frames. Notice that each evaluation of the STFT is normalized in energy.

The spectral centroid, which is defined as the "balancing point" of the spectral magnitude distribution.

The zero-crossing rate (ZCR), which is the number of time-domain zero-crossings within a frame.

The six features are:

- **features 1 & 2:** mean and variance of the spectral "flux"

*This work is supported by CNET/RENNES (Centre National d'Études des Télécommunications), FRANCE

- **features 3 & 4:** mean and variance of the spectral centroid
- **features 5 & 6:** mean and variance of the ZCR

The frames are 18 ms long. Means and variances are computed in a one-second segment.

Music can be regarded as a succession of periods of relative stability, notes and phones, in spite of the presence of short signals such as percussions (inducing high-frequency noise). Speech is rather a rapid succession of noise periods, such as unvoiced consonants, and of periods of relative stability, like vowels. Then, the selected features have interesting properties. They give very different values for voiced and unvoiced speech; and they are relatively constant within a segment of musical sound. The variances are higher for speech than for music, whereas the means tends to be greater for music than for speech, as shown in the following figure.

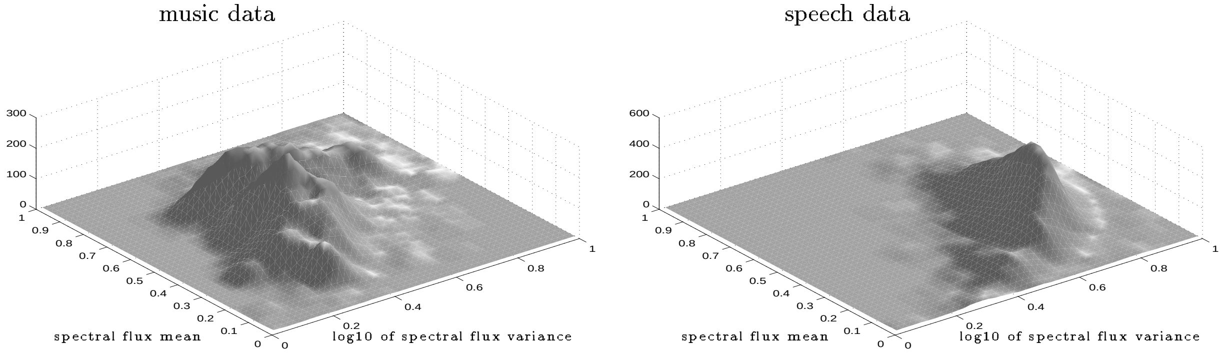


Figure 1: Normalized features 3D histogram.

In order to identify the segments, four classification methods have been tested. Only three have been retained: a Gaussian Mixture Model (GMM) classifier; the k-Nearest-Neighbors (kNN) classifier, with $k = 7$; Neural Networks (NN) classifier. Results gotten with Central Clustering were not conclusive. The GMM is based on the maximum likelihood method, iteratively derived with the expectation-minimization algorithm. Using NN, to get the classification when using features 1 & 2, the hidden layer contains 6 units. When using all the features, it contains 8 units.

The performances of the system have been evaluated using two labeled data sets: each set is 20 minutes long, and contains 10 minutes of speech and 10 minutes of music. For each set, the training is done using 80 % of the labeled samples and the test is operated on the remaining 20 %. The percentages of missclassified segments are given on the two following tables. Table 1 shows results when using only the features 1 & 2. Table 2 presents complete results, when using the all six features. Column 1 lists training errors obtained by comparing the results of a manual segmentation and the results of the training step. Next, testing errors, in column 2, are errors done when using training parameters for segmenting the test part. For the cross-validation, on column 3, the training part of the set number 1 is used, and the test is operated on the testing part of the set number 2 (and vice versa). Only the average percentage values are presented.

	training	testing	cross-validation
GMM	8.0 %	8.1 %	8.2 %
kNN	X	6.0 %	8.9 %
NN	6.7 %	6.9 %	11.6 %

Table 1: Percentage of missclassified segments when using features 1 & 2

	training	testing	cross-validation
GMM	7.9 %	7.3 %	22.9 %
kNN	X	2.2 %	5.8 %
NN	4.7 %	4.6 %	9.1 %

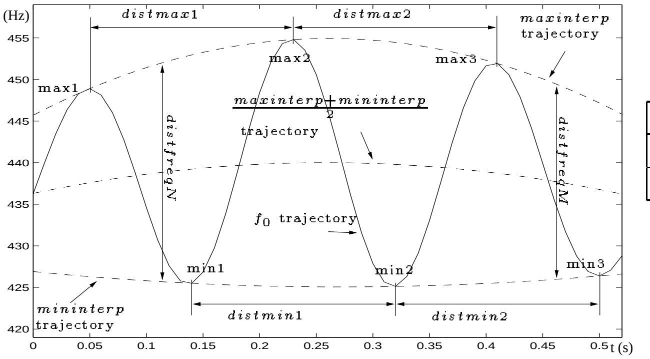
Table 2: Percentage of missclassified segments when using the six features

The two tables indicate that using more features tends to give better results, with noticeable improvement in some cases. Tests on larger data sets are presently done in order to give a final conclusion.

3 – *vibrato* suppression scheme

An algorithm to detect and to extract the vibrato is proposed below. This is very useful in order to model the vibrato for musical processing of sounds, but also to suppress it from the f_0 trajectory.

The algorithm, whose principle is illustrated by the following figure, is described below. Firstly, the f_0 trajectory local maxima are detected (max list is formed). The $maxinterp$ trajectory is obtained by interpolating maxima points. All the temporal distances between two successive local maxima are calculated ($distmax$ list). Secondly, the $distmax$ variance ($Vmax$) is calculated. The number of periods $distmax$ comprised between $0.15\,s$ and $0.25\,s$ (these values correspond to a $4\,Hz$ to $6.7\,Hz$ vibrato) is evaluated as a percentage $Pmax$. The same processings are done for the local minima. When such a vibrato exists, variances have low values and percentages get high values, because most of the vibrato periods are belonging to the small range between $0.15\,s$ and $0.25\,s$. Thirdly, if a vibrato is detected, the frequency distances between $maxinterp$ trajectory and $mininterp$ trajectory at all the instants are computed ($distfreq$ list). If their mean $Mfreq$ is high, the vibrato is significant and it is necessary to suppress it from the f_0 trajectory. In case of presence of significant vibrato, *notes* and *phones* scheme features based on f_0 trajectory (features 1, 2, 8, 9: see section 4) fail. The new f_0 trajectory, without vibrato, is the average of the $maxinterp$ and $mininterp$ trajectories.
The results are given in the following table. They are obtained for a flute excerpt (no vibrato) and for a singing voice excerpt (vibrato):



	Vmax	Vmin	Pmax	Pmin	Mfreq
song	0.0010	0.0012	85 %	84 %	34 Hz
flute	0.0090	0.0100	50 %	46 %	12 Hz

4 – *notes* and *phones* segmentation scheme

In this work, the strategy of the analysis is composed of three steps. The first step is to extract a large set of features. A feature will be all the more appropriate as its time evolution presents strong and short peaks when transitions occur, and as its variance and its mean remain at very low levels when describing a steady state part. Secondly, each of these features is automatically thresholded. Lastly, a final decision function based on the set of the thresholded features has been built and provides the segmentation marks.
First step. For the *notes* and *phones* scheme, the features used are (including new features):

- **features 1 & 2:** derivative and relative derivative of f_0: At the instant i, the derivative of the function f is $d(i) = |f(i+1) - f(i-1)|$ and its relative derivative is $\delta(i) = d(i)/f(i)$. The trajectory of f_0 is obtained using an IRCAM software, *f0*: for more details see [Doval 94].
- **features 3 & 4:** derivative and relative derivative of energy: The energy is calculated on overlapped frames. The chosen frame duration is 20 ms.
- **feature 5:** measure of the inharmonicity of sine wave partials: an IRCAM software, *additive*, gives us the trajectories of all the N partials between f_0 and half sampling frequency $fe/2$. Then, the relative frequency deviation between the theoritical location of the n*th* partial and its practical location is defined in this work as being: $h_n = |f_n - n.f_0|/(n.f_0)$, with $n\,\epsilon\,[2, N]$. The final measure of the inharmonicity is: $H(i) = \sum_{n=2}^{N} h_n$.
- **feature 6:** voicing coefficient: It uses the magnitude of the Short Time Fourier Transform spectrum $\hat{S}$ computed on a 20 ms windowed signal frame, and the value of f_0 for this frame. A voicing coefficient is obtained for all the N partials between f_0 and $fe/2$. For each partial n, a $2\Delta f$ bandwith is considered. The "signal magnitude spectrum" is truncated between $nf_0 - \Delta f$ and $nf_0 + \Delta f$ and normalized in energy. Similarly, the "frame window's magnitude spectrum" is truncated between $-\Delta f$ and Δf and normalized in energy. The $2\Delta f$ bandwith is chosen for keeping the largest part of the energy of the mainlobe. A high value of v_n, the correlation coefficient of these two truncated and normalized spectra, confirms the presence of a sine wave (corresponding to a partial). Then, the proposed global voicing coefficient is: $V = \sum_{n=1}^{N} v_n$.

- **feature 7:** spectral "flux": Already defined (see the *source* segmentation scheme).
- **feature 8:** distance between two statistical models, on the f_0 trajectory. This feature is able to detect pitch transition between two notes. Two segments of this trajectory (commonly with a 50 ms fixed duration) are considered: one before and one after the current analysis time i (in samples). We assume that these observations are obeing two stochastic processes. The probability $p(i)$ that the distance between the two models exceed a threshold S is computed. Commonly, we consider that the models are obeing normal processes. Accordingly, this probability can be expressed analytically. S is related to a musical characteristic: a pitch gap (such like a quarter-tone).
- **feature 9:** detection of abrupt changes using AR modeling, on the f_0 trajectory. Two segments are considered. One is extended from 50 ms before the current sample j. The other one extends from the previously detected change at sample i ($i < j$). Two AR models are calculated, and their *Cumulative Entropy* (CE) is estimated. When the difference between the CE max on the i to j interval and the CE at j is bigger than a threshold λ, a change is detected and the current change point stays now at j. For more details, see [Basseville and Nikiforov 93].

The aforementioned features are sampled at 100 Hz. The "distance between two statistical models" and the "detection of abrupt changes using AR modeling" methods are also used on energy trajectory.
Second step. Only feature 9 is based on a decision algorithm. For the others, discrimination between meaningful peaks and noise peaks must be done. To discriminate between these two classes, a threshold is automatically estimated (3σ rule). Assuming a normal distribution, one can determine σ when retaining 90% of the smallest samples of the feature. This gave consistent results. Thresholding gives a function with values 0 (steady state) or 1 (a transition is detected).
Third step. Finally, a decision function is computed, based on the particular decisions obtained for each feature. The decision functions (sequencies of 0 and 1) are computationaly added (giving much higher value when several of them gave transitions exactly at the same time). The resulting signal corresponds to a kind of variable density pulse signal. Actually, marks are scattered because the transitions are not instantaneous and because some of the features react rather at the beginning and others rather at the end of transitions. This signal is modified as follows. When the distance between 2 segmentation marks (pulses) is smaller than T ms ($T = 50$ ms), they are replaced by a new one, whose weight is the sum of the 2 previous values, and whose position is the corresponding centre of mass. The original pulse signal is modified step by step following this way. The result is a weighted pulse signal where weight corresponds to the confidence granted to the presence of a transition. The optimized localisation of these pulses (due to the centre of mass calculation) gives good approximation of the exact time of transition. Then, a threshold is applied to eliminate the smallest segmentation marks (false alarms).
This final decision depends on the current application. As an example, it is possible to sing several consecutive vowels with the same pitch: segmentation according to pitch must not lead to different segments. But in the same case segmentation into phones must separate exactly the vowels (one can imagine one vowel sung with f_0 variations...). For the final step, some features among the previous list are selected: the most appropriate for the current application (segmentation into notes or phones or other combination) are chosen. Related to the previous example, features 1, 2, 3, 4, 5, 8 and 9 correspond to pitch segmentation, and features 6 and 7 correspond to phone segmentation.
Others features based on methods such as higher order statistics and cepstral coefficients are under study.

References

[BN93] Michèle Basseville and Igor V. Nikiforov, *Detection of abrupt changes*, PTR Prentice-Hall, 1993.

[Dov94] Boris Doval, *Estimation de la fréquence fondamentale des signaux sonores*, Ph.D. thesis, Université de Paris VI, Mars 1994.

[SS97] Eric Scheirer and Malcolm Slaney, *Construction and evaluation of a robust multifeatures speech/music discriminator*, IEEE Transactions on Acoustics, Speech, and Signal Processing (ICASSP'97), 1997, pp. 1331 – 1334.