SEPARATION OF HARMONIC SOUND SOURCES USING SINUSOIDAL MODELING

Tuomas Virtanen, Anssi Klapuri

Signal Processing Laboratory, Tampere University of Technology
P.O.Box 553, FIN-33101 Tampere, FINLAND
tuomasv@cs.tut.fi, klap@cs.tut.fi

ABSTRACT

In this paper, an approach for the separation of harmonic sounds is described. The overall system consists of three components. Sinusoidal modeling is first used to analyze the mixed signal and to obtain the frequencies and amplitudes of sinusoidal spectral components. Then a new method is proposed for the calculation of the perceptual distance between pairs of sinusoidal trajectories, according to the implications of recent psychoacoustic knowledge. A procedure for classifying the sinusoids into separate sound sources is presented. The system is not designed to separate sounds that have same fundamental frequencies. However, a solution to detect single colliding sinusoids is given. Demonstrations of synthesized signals are available at http://www.cs.tut.fi/~tuomasv/demopage.html.

1. INTRODUCTION

Separation of mixed sounds has several applications in the analysis, editing and manipulation of audio signals. These include for example structured audio coding, automatic transcription of music, audio enhancement, and computational auditory scene analysis. Until now, main part of the research in sound separation has taken place in the computational auditory scene analysis [1].

When two sounds overlap in time and frequency, separating them is difficult and there is no general method to resolve the component sounds. In this paper, we take the approach to decompose signals into spectral components and then assign them to sound sources using a set of perceptual association cues. While this is a very complicated task, some assumptions usually have to be made, in our case assuming that the sound sources are harmonic and do not have the same fundamental frequencies. After these assumptions, it is possible to reconstruct the separated sounds which are perceptually close to the original ones before mixing.

1.1 System overview

Calculations proceed as follows. First, the system uses sinusoidal modeling to represent signals with sinusoidal trajectories. Second, some breaks caused by amplitude modulation, transients or noise in resulting sinusoids are removed by interpolating trajectories. Third, the system estimates the perceptual closeness of the trajectories by calculating the difference of scaled amplitudes and frequencies and the harmonic concordance of the trajectories. Then the trajectories are classified into sound sources. The system can determine which of the trajectories are result of colliding harmonics, and then split these trajectories in two. Finally, when trajectories are classified and split, the system is able to synthesize the two sounds separately.

The classification part itself is currently the most undeveloped part of the system. In simulations, the classifier assumes two sound sources and uses their onset difference to initialize both classes. Thus the onset difference of the sounds had to be at least 100 ms. However, the methods themselves can be used also in more complex tasks, and for simultaneously onsetting sounds. However, as long as only sinusoidal modeling is used, it is difficult to obtain good results for a large number of mixed sounds, because some of their harmonic partials are likely to be undetected.

2. SINUSOIDAL MODELING

Theoretically, any signal can be represented with an infinite number of sinusoids. However, it is reasonable to use sinusoids only for stable, periodic signals, because the number of required sinusoids is the same as the number of partials in the sound.

The idea of sinusoidal modeling is to use sinusoids with time-varying frequencies and amplitudes to represent harmonic signals. To obtain these sinusoids from an original signal, their short-time spectra is analyzed by taking discrete Fourier transform of the windowed signal to locate the prominent peaks in the amplitude spectra. To detect a large number of peaks reliably in the magnitude spectrum, we use a statistical method called F-test developed by Thomson [2]. The method employs a set of orthogonal windows called discrete prolate spheroidal sequences, and then uses the F-test to determine whether there exists a line components in the spectrum a particular frequency. The F-test is a measure of the ratio between estimate of spectral peaks at their respective frequencies to that of continuous, non-harmonic part of spectrum.

Once amplitude and frequency of each peak are estimated at each frame, the detected peaks are tracked together in interframe trajectories. A peak continuation algorithm tries to find the most suitable peak in the next frame, which has the frequency and amplitude close to the existing trajectory in current frame [3,4,5]. The result of the peak continuation algorithm is a set of sinusoidal trajectories with time-varying frequencies and amplitudes.

To achieve accurate results when detecting the higher partials of harmonic sounds that exhibit amplitude or frequency modulation, a short time step between frames is needed. However, to detect two sinusoids whose frequencies are close to each other, the required window length is not depending on the wavelength of the sinusoids, but the wavelength of the frequency differences. This means that a long window is needed to detect all the harmonics. These two requirements are fulfilled by using a relatively long window and a large overlap factor so that time increment is small compared to the window size. If the harmonics of different sounds

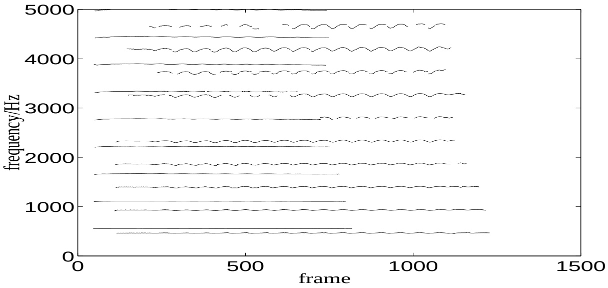

Figure 1. Sinusoidal trajectories of a signal consisting of oboe and violin sounds starting at times 100 and 300 ms.

are very close to each other, which often happens in musical signals, the separation of trajectories requires more sophisticated modeling. For example, iterative separation has been discussed in [6]. Our system is not iterative, it only locates the peaks in the amplitude spectrum and uses them as possible locations of sinusoidal trajectories. The resulting trajectories of sinusoidal modeling are illustrated in Figure 1. In our system, the upper frequency limit of trajectories was set to 5kHz, because harmonics above that have usually too small amplitudes to estimate them reliably.

2.1 Break interpolation

All the trajectories that result from the sinusoidal modeling are not usually representing entire partials of the sound. The most common estimation errors are breaks in trajectories. They can be caused by transients or noise occurring at the same time, or the amplitude of the harmonic itself is so low that the trajectory can not be estimated. This happens usually with signals with strong amplitude modulation, so the amplitude actually can go to zero. This phenomena can be easily seen in higher harmonics of violin of Figure 1.

If time difference between two sounds is small and frequencies and amplitudes of the sounds are close to each other, human auditory system connects the sounds. Our system tries to model this by connecting trajectories which are close to each other. The breaks between trajectories are interpolated. The interpolation also increases the robustness of system, as one harmonic is represented with one long trajectory instead of many short ones.

The trajectories to be connected are detected by comparing the onset and offset times, frequencies and amplitudes near the breaks and finding those trajectories which are most propable belonning to the same partial. Then the breaks are removed by interpolating the frequencies and amplitudes over the break. In our system, we are using linear interpolation since it seems to work well enough for separation purposes. Not every break is interpolated, because allowing too long breaks or big amplitude/frequency differences causes wrong trajectories to be connected. In this example, the total number of trajectories reduced from 87 to 48.

3. MEASURE OF PERCEPTUAL DISTANCE

In his work, Bregman [7] lists the following associaton cues in human auditory organisation:

1. Spectral proximity (closeness in time or frequency)
2. Harmonic concordance

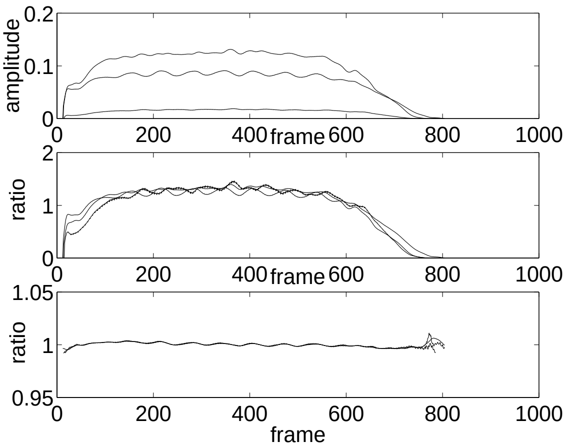

Figure 2. The upper plot shows the amplitudes of three first harmonics of the oboe. In the middle plot are scaled curves of these amplitudes. In the third plot are scaled curves of the frequencies of the same harmonics.

3. Synchronous changes of the componenets: a) common onset, b) common offset, c) common amplitude modulation, d) common frequency modulation, e) equidirectional movement in spectrum
4. Spatial proximity.

In this study, we focus on the synchronous changes of the components, the harmonic concordance is taken into account, too.

3.1 Measuring the amplitude and frequency changes

When the measument of common amplitude and frequency modulation was studied, we found out that in some cases, modulation can be expressed with two quantities, modulation frequency and index. However, to present amplitude or frequency modulation only with two quantities is usually not enough. Because modulation usually varies in time domain, we would need several measurements to cover the changes within time. Also, the changes of the overall long-time intensity of the sound sometimes makes it hard to measure the modulation characteristics of the sound.

Different harmonic partials have a wide range of amplitudes values and sometimes their long-time progress is not similar. However, by scaling the amplitude of each partial by it's average, the resulting curves are quite close to each other. In case of frequencies this method is even more accurate, because frequencies do not change so much over time as amplitudes, as illustrated in figure 2. The mean square error between these scaled frequencies measures the frequency distance between the sinusoidal trajectories:

$$d_f(i, j) = \frac{1}{t_2 - t_1 + 1} \sum_{t = t_1}^{t_2} \left(\frac{f_i(t)}{f_i} - \frac{f_j(t)}{f_j} \right)^2 \tag{1}$$

where $f_i(t)$ is frequency of trajectory p_i at time t. Times t_1 and t_2 are chosen so that both trajectories p_i and p_j exists at times $t_1<t<t_2$. Scaling coefficients f_i and f_j are the average frequencies of trajectories p_i and p_j, calculated over times t_1 and t_2.

The same principle is used to obtain the perceptual distance caused by amplitude differences $d_a(i,j)$ using amplitudes $a_i(t)$ and

$a_j(t)$ and their averages a_i and a_j.

These quantities measure all the synchronous changes of the components which were listed in the beginning of this chapter. Of course, long-term movements in the spectrum are weighted more, but if the frequencies or amplitudes do not change a lot, which is often true for musical sounds, the quantity measures well the smaller changes like frequency modulation.

3.2 Measure of harmonic concordance

The frequency f_p of a harmonic partial is close to an integral multiple of the fundamental frequency f_0 of the harmonic sound. In our system we do not know the fundamental frequencies of the sounds even though they could be estimated in the original signal [8]. In our system, we do not try to calculate fundamental frequencies of the sounds in the mixture. Instead, we developed a measure for modeling the harmonic concordance of any two sinusoidal trajectories. If we have two sinusoidal trajectories belonning to one harmonic source, the ratio of the frequencies of the trajectories is a ratio of small positive integers:

$$\frac{f_i}{f_j} = \frac{a}{b} \qquad (2)$$

where f_i and f_j are frequencies of the sinusoidal trajectories p_i and p_j and which are a^{th} and b^{th} harmonic of a sound.

Because we do not know which trajectory belongs to which sound or which trajectory is which harmonic partial, we assume that the fundamental frequency can not be smaller than the minimum frequency found in sinusoidal modeling. That way, we obtain upper limits for a and b:

$$a = 1, 2, \ldots, \left\lfloor \frac{f_i}{f_{min}} \right\rfloor, \; b = 1, 2, \ldots, \left\lfloor \frac{f_j}{f_{min}} \right\rfloor \qquad (3)$$

where f_{min} is the minimum frequency found in sinusoidal modeling.

After determining the limits for a and b, we calculate all the ratios for possible a and b and choose the one which the best estimate for the ratio of the frequencies. The harmonic distance quantity is then the error between ratio of the frequencies and the ratio of a and b. To give equal weight for ratios below and above unity, we use the absolute value of the logarithm of the ratio to measure the harmonic distance between trajectories:

$$d_h(i, j) = min \left| \log\left(\frac{f_i / f_j}{a / b} \right) \right| \qquad (4)$$

with a and b having the restrictions described in (3)

3.3 Overall perceptual distance between trajectories

The overall perceptual distance between any two trajectories is a weighted sum of frequency, amplitude and harmonic distances:

$$d_{all}(i, j) = w_f d_f(i, j) + w_a d_a(i, j) + w_h d_h(i, j) \quad (5)$$

Due to the physics of sound production, frequencies usually do not vary as much as amplitudes. Thus the frequency distance has to be weighted more than the amplitude distance. The harmonic distance is calculated in a different way than the others, thus it has a different scaling. Because perceptual organisation of simultaneous trajectories is largely based of harmonic concordance, the harmonic distance is weighted so that it has the biggest effect.

Onset times are not directly taken into account when calculating these overall distances. Amplitude curves themselves contain the information of onset and offset times. Before onset, the amplitude of a trajectory is always zero. For natural sounds, the amplitude curve usually rises rapidly after the onset, and then starts to slowly decay. After the offset, the amplitude is again zero. The amplitude distance takes all this behaviour into account.

4. TRAJECTORY CLASSIFICATION

After estimating the closeness between each pait of sinusoidal trajectories, they should be classified into separate sound sources. Because we do not have any coordinates common to all the trajectories, but only the distances between each pair of trajectories, we have to classifity these trajectories into classes having minimum error between trajectories inside a class:

$$min\left(\frac{1}{|S_1|} \sum_{i, j \in S_1} d_{all}(i, j) + \frac{1}{|S_2|} \sum_{k, l \in S_2} d_{all}(k, l) \right), \quad (6)$$

$$S_1 \cup S_2 = S, S_1 \cap S_2 = \varnothing,$$

where S_1 and S_2 are trajectory sets of two sounds, S is set of all trajectories, and $|S|$ the cardinality of a set.

An ideal solution for choosing these classes would be to calculate all the possible permutations and choose the best one. However, for a practical number of sinusoidal trajectories, the number of calculations becomes very large (2 to the power of number of trajectories), so some other of classification approach must be taken. An efficient solution to this problem is to choose an initial set of trajectories for each class (or sound, in this case) and then add trajectories one by one to the classification by choosing the trajectory which has the minimum distance to the previous ones.

A good initial set of trajectories can be obtained by choosing all the trajectories whose onset times are close enough to an estimated onset time t_0 of a sound, and then evaluating all the possible subsets of certain number of trajectories. The subset which minimizes the error contains usually sinusoids which have been tracked well and do not contain any estimation errors or colliding sinusoids. To emphasize long, stable trajectories, the length of a trajectory can be used as a scaling factor. The number of trajectories close to onset time is usually so small that all the permutations can be evaluated:

$$e = min\left(\sum_{i, j \in S_1} \frac{d_{all}(i, j)}{\sqrt{length(p_i)}} \right) \qquad (7)$$

$$(i \in S_1; |t_1 - t(i)| < t_{limit}), \; |S_1| = c, S_1 \subset S$$

where t_1 is estimated onset time of sound 1, $t(i)$ is onset time of trajectory i, t_{limit} is maximum distance of trajectory to the onset and c is the size of initial subset. The onset times times can be obtained in many ways. In this system, we used the sum of difference of amplitudes at the beginning of each trajectory which were smoothed using a triangular window. The smoothed amplitudes were summed and local maxima of the resulting curve were selected as onset times. This onset detection method can detect onsets of sounds which do not have a strong transient in the beginning, for example violin.

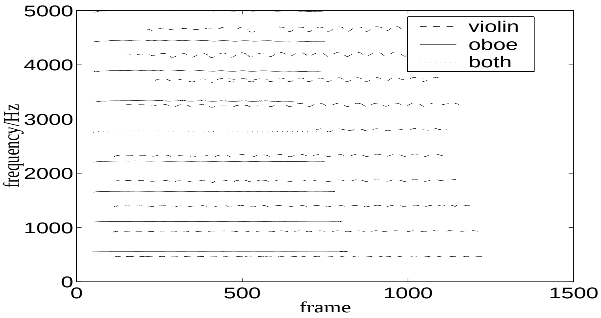

Figure 3. Classified trajectories.

When we have estimated the initial subsets for each sound, we start adding the rest of the trajectories into them one by one, always choosing a subset and trajectory whose distance is the smallest. The distance between a subset and a trajectory is simply the average of distances between the trajectory and the trajectories of the subset. Iteration is continued until all the trajectories are classified. The result of the classification is presented in Figure 3. One colliding trajectory belongs to both sounds. Detection of these colliding trajectories is discussed in the next section.

5. COLLIDING TRAJECTORIES

As mentioned, the harmonics of two sounds often overlap in musical signals. Sinusoidal trajectories which are the result of overlapping harmonics, are called colliding trajectories. Which harmonics are overlapping, depends on the interval of the sounds. With harmonic intervals [7,8] like major third and perfect fifth, many of the low harmonics are overlapping, because the ratio of the fundamentral frequencies is a ratio of small integers. In the case of dissonant intervals, the low harmonics are not overlapping, but they can still be quite close to each other, which may cause estimation errors in sinusoidal modeling.

It is easy to see that an efficient way to detect colliding sinusoids is to find trajectories which are harmonically suitable for both sounds, or, whose harmonic distance to both sounds is small enough:

$$\frac{1}{|S_1|}\sum_{j \in S_1} d_h(i, j) + \frac{1}{|S_2|}\sum_{k \in S_2} d_h(i, k) < c_{limit} \tag{8}$$

where S_1 and S_2 are sets of trajectories belonning to sounds 1 and 2 and c_{limit} is a constant. If equation is true for trajectory p_i, then it is propable that p_i contains harmonic partials from both sounds.

The detected amplitude that resuls from two or more sinusoids close to each other in frequency is affected by the phase difference of the sinusoids. Because of the frequency and amplitude modulation, which is usually present in natural voices, estimation of the exact amplitudes and frequencies is very complicated, because after sinusoidal modeling we do not have the exact spectral information, only the detected sinusoids.

General solution of the colliding sinusoids problem is above the scope of this paper. This has been addressed e.g. in [6]. However, to achieve better perceptual quality, we approximate the underlying harmonics. The system interpolates the amplitudes of the colliding trajectories using amplitude curves of other, not-colliding sinusoids. The frequencies are left intact.

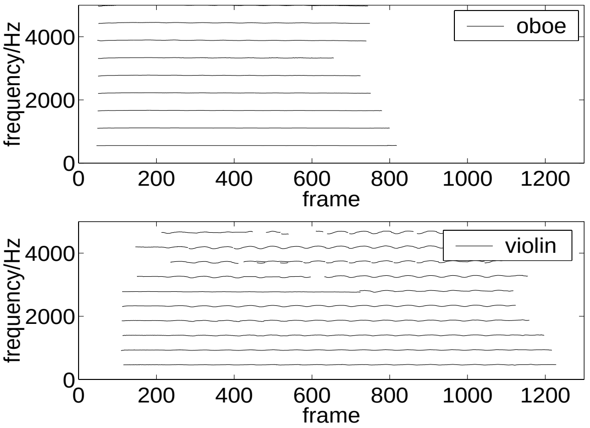

Figure 4. Separated trajectories. The 6th harmonic of the violin which was colliding with the 5th harmonic of the oboe, doesn’t have vibrato.

Finally, when we have detected and splitted the colliding trajectories, we can represent the separated signals and synthesize them. The separated trajectories are presented in Figure 4.

6. CONCLUSIONS

A system was presented that is able to separate harmonic sounds using the synchronicity and harmonic relations of sinusoidal spectral components. Validation experiments and the perceptual quality of separated sounds demonstrate that the presented methods can be used to yield practically applicable results. Remaining problems to be addressed in the future include dynamic detection of the number of the mixed sounds, better estimation of amplitudes of colliding frequency partials, and separation of sounds that have the same onset time.

7. REFERENCES

[1] D. Rosenthal, H. G. Okuno (eds.), “Computational Auditory Scene Analysis”, Lawrence Erlbaum Associates, NJ., 1998.

[2] D. J. Thomson, “Spectrum Estimation and Harmonic Analysis”, Proceedings of the IEEE, 70(9), 1982

[3] X. Serra, “Musical Sound Modeling with Sinusoids plus Noise”, Roads, Pope, Poli (eds.), “Musical Signal Processing”, Swets & Zeitlinger Publishers, 1997

[4] S. N. Levine, “Audio Representation for Data Compression and Compressed Domain Processing”, Ph.D thesis, University of Stanford, 1998

[5] T. Virtanen, “Spectrum Modeling with Sinusoids Plus Noise”, presented in the Seminar on the Content Analysis of Music And Audio, Helsinki, Finland, 1999

[6] T. Tolonen “Methods for Separation of Harmonic Sound Sources using Sinusoidal Modeling”, presented at the AES 106*th* Convention, Munich, Germany, 1999

[7] A. S. Bregman. “Auditory Scene Analysis”. MIT press, 1990

[8] A. Klapuri, “Number Theoretical Means of Resolving a Mixture of Several Harmonic Sounds”, Proceedings of the European Signal Processing Conference, 1998