SEPARATION OF HARMONIC SOUNDS USING MULTIPITCH ANALYSIS AND ITERATIVE PARAMETER ESTIMATION

Tuomas Virtanen, Anssi Klapuri

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

ABSTRACT

A signal processing method for the separation of concurrent harmonic sounds is described. The method is based on a two-stage approach. First, a multipitch estimator is applied to find initial sound parameters which are reliable, but inaccurate and static. In a second stage, more accurate and time-varying sinusoidal parameters are estimated in an iterative procedure, which imposes certain constraints for the amplitudes and frequencies between the components of each sound. The proposed algorithm makes it possible to reconstruct separated sounds that are perceptually close to the original ones before mixing. Experimental data comprised sung vowels and 26 musical instruments. Sound separation was performed for random sound mixtures ranging from one to six simultaneous sounds. The system is able to produce meaningful results in all polyphonies, the quality of separated sounds gradually degrading along with the polyphony. Demonstration 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 enchancement and computational auditory scene analysis [1].

Whereas the human auditory system is very effective in "hearing out" sounds in complex signals, computational modeling of this function has proved to be very difficult [2]. Only few published reports focus on this problem. However, some proposals have been made that apply knowledge about human auditory scene analysis [3], apply iterative estimation and cancellation [4], or aim at more practical goals [5].

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, two properties of harmonic sounds are utilized to estimate the parameters of the underlying sounds: the harmonic structure of the sounds is used in frequency estimation and the spectral envelope continuity of natural sounds is used in amplitude estimation. With these cues, it is possible to reconstruct separated sounds which are perceptually close to the original ones before mixing.

1.1. System overview

The overall structure of the system is presented in Figure 1. The input signal is passed to a multipitch estimator which estimates the number of the sounds present, the fundamental frequencies of the sounds and frequencies of each harmonic component of each sound. The multipitch estimator is explained in detail in Section 2.

The frequencies given by the multipitch estimator are used as a starting point for an iterative sinusoidal analysis system. The sinusoidal model represents the harmonic components of a sound with sinusoids that have time-varying frequencies, amplitudes and phases [6]. The sinusoidal model uses significantly shorter time frames than the multipitch estimator, therefore being able to track parameter changes inside one large multipitch estimation frame.

The exact time-varying frequencies and amplitudes of the components are analyzed using an iterative approach. Starting from the estimates given by the multipitch estimator, the accuracy of the parameters is improved in the least-squares sense, retaining the harmonic structure of the sounds. The spectral envelope of each sound is smoothed using a perceptually motivated mechanism which cuts off single higher-amplitude harmonic partials that result from coinciding partials of different sounds. Once the parameters of the harmonic components are analyzed in each time frame, the sounds can be synthesized separately.

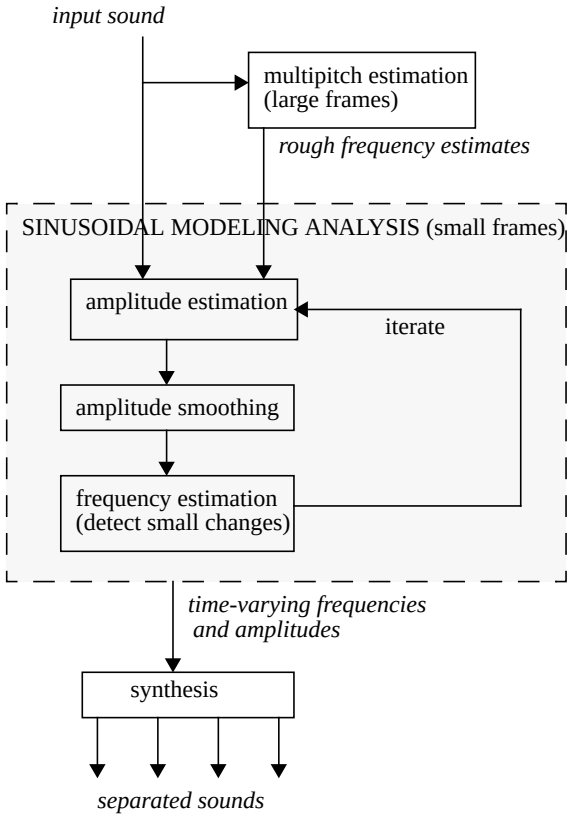

Figure 1. *Block diagram of the separation system.*

2. MULTIPITCH ESTIMATION

The sinusoidal model alone has been used for sound separation in [7]. However, it cannot produce reliable results if the number of concurrent sounds is too large. Harmonic relations of the frequency components, in turn, were found to be an effective organization force in rich sound mixtures. A signal processing method for estimating the multiple fundamental frequencies of concurrent musical sounds has been presented in [8]. The method is able to find reliably the most prominent sounds even in rich sound mixtures.

The multipitch estimation system is used to initialize the separation system. The multipitch estimator takes a single 90–200 ms frame of the acoustic input signal and outputs the fundamental frequencies and the rough frequency estimates of the harmonic partials of each sound. Detected sounds are outputted in the order of decreasing priority, the first one being aurally the most prominent and reliable.

The iterative parameter estimation stage looks again at the input signal, and calculates time-varying and significantly more accurate parameter estimates for the components of the sounds proposed by the multipitch estimation system. The separation algorithm assumes that the fundamental frequencies given by the multipitch estimation system are correct, and it cannot converge correctly if the initialization is wrong. This is not a very degrading assumption, since the multipitch estimation system in itself was found to be quite accurate, outperforming the average of trained musicians in musical chord identification tasks. Particularly, the firstly detected and most prominent sounds are very reliable even in rich sound mixtures. The overall error rate in four-voice mixtures is 8.1%, and the error rate for the firstly detected sound stays below 1% even in six-voice polyphonies. The multipitch estimation system is also applicable for real musical recordings, is instrument-independent, and is able to estimate the number of concurrent voices in input signals [8].

3. ITERATIVE PARAMETER ESTIMATION

3.1. Sinusoidal model

Mixed sounds are represented using sinusoids with time-varying frequencies, amplitudes and phases. The sinusoids are assumed constant in single analysis frame, so that the local model of the signal $s(t)$ is

$$\hat{s}(t) = \sum_{n=1}^{N} \sum_{k \in S_n} a_k \cos(2\pi f_k t + \phi_k), \quad (1)$$

where N is the number of the mixed sounds, S_n is the set of harmonic components belonging to sound n, and a_k, f_k and ϕ_k are the amplitudes, frequencies and phases of the harmonic component k. The short-time Fourier transform $\hat{S}(f)$ of the model is given by the parametric expression

$$\hat{S}(f) = \sum_{n=1}^{N} \sum_{k \in S_n} \frac{a_k}{2} (e^{i\phi_k} W(f - f_k) + e^{-i\phi_k} W(f + f_k)), \quad (2)$$

where $W(f)$ is the complex-valued Fourier transform of the analysis window translated at frequency f.

3.2. Amplitude and phase estimation

Initial estimates of the frequencies of the harmonic components are given by the multipitch estimator. The objective is to find the parameters for which the model $\hat{S}(f)$ best fits the obserbed spectrum $S(f)$. Least-square solution of amplitudes and phases for the given frequencies can be found using the following linear structure for the spectrum estimator [9]:

$$\hat{S}(f) = \sum_{n=1}^{N} \sum_{k \in S_n} [p_k R_k(f) + p_{K+k} R_k(f)], \quad (3)$$

where each spectral component is represented with two unknown parameters p_k and p_{K+k}, defined as

$$\begin{cases} p_k = \frac{a_k}{2} \cos\phi_k & k \in [1, K] \\ p_{K+k} = \frac{a_k}{2} \sin\phi_k & k \in [1, K] \end{cases} \quad (4)$$

and R_k, the $2K$ known expressions related to the Fourier transform of the window function are

$$\begin{cases} R_k(f) = W(f - f_k) + W(f + f_k) \\ R_{K+k}(f) = i[W(f - f_k) - W(f + f_k)] \end{cases}. \quad (5)$$

The least-square solution for amplitudes and phases is given by the expression

$$p = (\Re^H \Re)^{-1} \Re^H S \quad (6)$$

where the vectors R_k are used as columns of the matrix $\Re$ evaluated at discrete frequency points F_j: $(\Re)_{j,k} = R_k(F_j)$.

3.3. Amplitude smoothing

If sounds are in simple rational number relations, i.e., harmonic relations, some of the harmonic components overlap with each other in frequency domain. In the case of musical signals, this happens often since harmonic intervals are usually favoured over dissonant ones. In the case of dissonant intervals, the low harmonics are not overlapping, but they can still be quite close to each other. Closely spaced frequencies are a serious problem in parameter estimation. In amplitude estimation, matrix $\Re$ in Equation 6 becomes singular and the parameters cannot be solved directly.

The problem of closely spaced components in the amplitude estimation is solved by the following procedure. At first, the components that are too close to each other are detected. This can be done simply by setting a fixed frequency limit, since the rough frequency estimates of all components are known. Only one column in $\Re$ is created to represent each group of overlapping components, and the obtained amplitudes and phases are then used for the whole group. Overlapping partials result in single higher-amplitude partials in the harmonic series of the sounds. This is handled in the last stage, in which the whole harmonic series of all sounds is processed with a perceptually motivated smoothing mechanism [10]. In practice, the amplitude envelope of each sound is first smoothed over a critical-band frequency scale, and then the minimum between the original and smoothed amplitude value is substituted for each partial amplitude. This cuts off single higher-amplitude harmonic partials that are resulted from detected or undetected co-occurring sounds, without corrupting the perceived timbre of the sounds.

If only two harmonic components are overlapping with each

other, but the frequencies of the components are not exactly the same, the amplitude modulation caused by the frequency difference can be used to estimate the amplitudes of both colliding sounds. This was used in the system described in [8].

3.4. Frequency estimation

Spectrum estimator in Equation 2 is nonlinear in terms of f_k. Depalle and Hélie [9] used a first-order limited expansion of the Fourier Transform of the analysis window around each frequency component to find a better estimate of each frequency. Alternative stages of amplitude and frequency estimation were iteratively repeated to converge towards an optimal estimate in the least-squares sense. The mentioned expansion is denoted by

$$W(f \mp f_k) = W(f \mp \hat{f}_k) \mp W'(f \mp \hat{f}_k)\Delta_k + o(\Delta_k^2), \quad (7)$$

where Δ_k is the distance between the correct frequency and its estimate ($\Delta_k = f_k - \hat{f}_k$), and W' is the derivative of the Fourier transform W of the window function. If S is the observed spectrum and $\hat{S}$ the spectrum obtained with frequency estimates $\hat{f}_k$ and fixed amplitudes and phases, and matrix Ω defined by

$$(\Omega)_{j,k} = \frac{a_k}{2}(-e^{i\phi_k}W'(F_j - \hat{f}_k) + e^{-i\phi_k}W'(F_j + \hat{f}_k)), \quad (8)$$

the least-square solution for the frequencies is

$$f = \hat{f} + (\Omega^H\Omega)^{-1}\Omega^H(S - \hat{S}). \quad (9)$$

The frequency estimation is very sensitive to the shape of the window function. In practise, the Fourier transform of the window function must not have sidelobes [9].

3.5. Extension to harmonics sounds by introducing constraints for frequency relations

The frequency partials of real-world harmonic sounds cannot be assumed to be in exact integer ratios. However, it can be assumed that the frequency ratios of a sound remain constant through time, even though the fundamental frequency varies [11].

Let the frequency ratio of the component k to the lowest-frequency partial of the sound it belongs to be denoted by r_k. This value is constant for each partial. Usually r_k is close to the integer number corresponding to the index of the harmonic partial, i.e., unity for the fundamental frequency, five for the fifth harmonic etc. Let us define δ_n to be the distance between the estimated and correct frequency of the lowest frequency component of sound n. The expansion can now be written in the form

$$W(f \mp f_k) = W(f \mp \hat{f}_k) \mp W'(f \mp \hat{f}_k)r_k\delta_n + o(\Delta_k^2),\ k \in S_n. \quad (10)$$

Since the frequencies of a single sound are linearly dependent on each other, Ω can now be written in the form

$$(\Omega)_{j,n} = \sum_{k \in S_n} \frac{a_k}{2}r_k(-e^{i\phi_k}W'(F_j - \hat{f}_k) + e^{-i\phi_k}W'(F_j + \hat{f}_k)) \quad (11)$$

and the solution for $\delta = [\delta_1, \ldots, \delta_N]^T$ is given by

$$\delta = (\Omega^H\Omega)^{-1}\Omega^H(S - \hat{S}). \quad (12)$$

Now we get a better estimation for frequency component by

$$f_k = \delta_n r_k,\ k \in S_n. \quad (13)$$

This method retains the harmonic structure of a sound since the ratios of the frequencies belonging to a common sound do not change. The corrected fundamental frequency value is based on all the harmonic components.

3.6. Iteration

Successive amplitude/phase estimation, amplitude smoothing, and frequency estimation stages are repeated until the values converge. In the first time frame, more than ten iterations are needed. In subsequent frames, the previous frame is used to initialize the frequencies, instead of using the multipitch estimator. In this case, only less than five iterations are needed. This makes the computational complexity of the algorithm practically applicable. Figure 2 illustrates the time-varying fundamental frequency of a flute tone.

Like in amplitude estimation, nearby harmonic components of other sounds can disturb the frequency estimation. Large-amplitude partials can even "catch" the analyzed harmonics so that a wrong sound becomes detected. This problem can be solved by choosing only some of the harmonic components of each sound to estimate the frequency changes. The components are chosen so that there is no interfering components at nearby frequencies. If there is not enough such components, we choose components that have large amplitudes compared to those of the interfering components.

4. SYNTHESIS

Once the parameters of the harmonic components have been estimated in each frame, the sounds can be synthesized separately. In synthesis, the frequencies, amplitudes and phases are interpolated from frame to frame, and time-domain signals are obtained by summing up all the harmonic components of each sound [6].

The parameters of the sounds can also be further analyzed, or manipulation can be performed on the parametric data [8]. An example of the synthesized signals is illustrated in Figure 3.

5. EXPERIMENTAL RESULTS

Simulations experiments were carried out to monitor the behaviour of the proposed algorithm. Test material consisted of a database of sung vowels plus 26 different musical instruments comprising plucked and bowed string instruments, flutes, and brass and reed instruments. These introduce several different sound pruduction mechanisms, and a variety of spectra. Semirandom sound mixtures were generated by first allotting an instrument, and then a random note from its whole playing range,

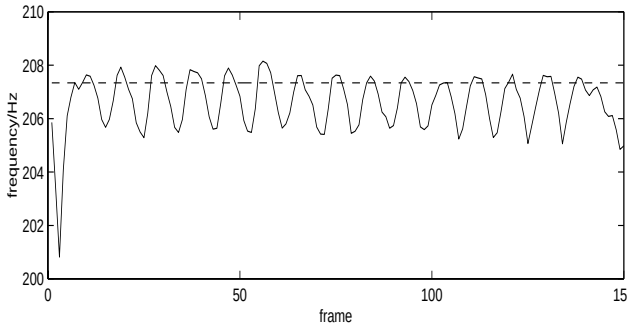

Figure 2. *Fundamental frequency of a flute. The dashed line is the rough estimate given by the multipitch estimator. The solid line is obtained using the iterative estimation procedure.*

however, restricting the pitch over five octaves between 65 Hz and 2100 Hz. A number of one to six simultaneous sounds were allotted, and then mixed with equal mean-square levels. Acoustic input was fed to the separation algorithm. Separated sounds were resynthesized and compared to the originals in informal listening tests.

For two or three-voice polyphonies and dissonant intervals, the system was able to produce good results in almost all cases. In very simple harmonic relations, such as the octave relation, most of the harmonic components of the sounds are overlapping and the error in resynthesized signal was usually clearly audible. Also, the sharp attacks in sounds like guitar and piano were somewhat smeared due to the sinusoidal model used. In spite of the degraded quality of the separated sounds, the system was still able to produce perceptually meaningful results most of the time in all polyphonies ranging from one to six simultaneous sounds.

When the polyphony increased, the perceptual quality of the outputted sounds decreased gradually. In four and five-voice polyphonies, the system was able to produce good results in some cases, but usually the quality of synthesized signals was already remarkably reduced. For six-sound polyphonies the system could still produce applicaple results for some of the notes, but typically at least one of the sounds was defective.

6. CONCLUSIONS

A sound separation system was described that is able to resolve mixtures of harmonic sounds reliably and without *a priori* knowledge about the number of voices or type of sound sources involved. The proposed new principles, frequency-ratio constrained iterative parameter estimation and critical-band frequency scale spectral envelope smoothing were effective and produced perceptually good sound quality. The overall approach of first detecting the fundamental frequencies of the sounds in a longer window and then estimating more accurate time-varying parameters was successful in combining robustness and accuracy.

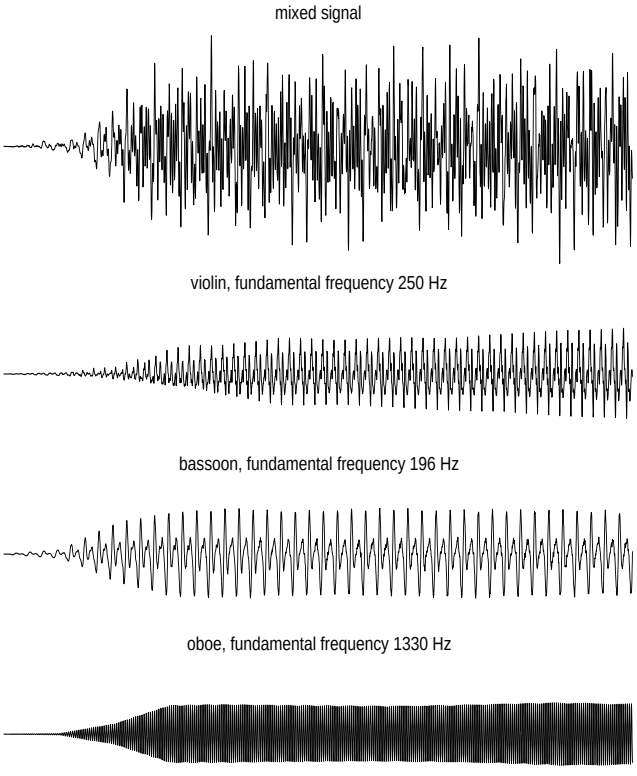

Figure 3. *A separation example: a three-note mixed sound and separated sounds*

7. REFERENCES

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

[2] A. S. Bregman. “Auditory scene analysis: the perceptual organization of sound,” MIT Press, Cambridge, Massachusetts, 1990.

[3] G. J. Brown, M. P. Cooke. “Perceptual grouping of musical sounds: A computational model,” J. of New Music Research 23, 107–132, 1994.

[4] A. de Cheveigné. “Separation of concurrent harmonic sounds: Fundamental frequency estimation and a time-domain cancellation model of auditory processing,” J. Acoust. Soc. Am. 93 (6), 3271–3290, 1993.

[5] M. Goto. “A robust predominant-F0 estimation method for real-time detection of melody and bass lines in CD recordings,” Proc. IEEE International Conf. on Acoust., Speech, and Signal Processing, Istanbul, Turkey, 2000.

[6] T. Virtanen. “Audio signal modeling with sinusoids plus noise”. MSc thesis, Tampere University of Technology, 2001.

[7] T. Virtanen, A. Klapuri. “Separation of Harmonic Sound Sources Using Sinusoidal Modeling,” IEEE International Conference on Acoustics, Speech and Signal Processing, Istanbul, Turkey, 2000.

[8] A. Klapuri, T. Virtanen, J.-M. Holm. “Robust multipitch estimation for the analysis and manipulation of polyphonic musical signals,” In Proc. COST-G6 Conference on Digital Audio Effects, Verona, Italy, 2000.

[9] Ph. Depalle, T. Hélie. “Extraction of Spectral Peak Parameters Using a Short-time Fourier Transform and no Sidelobe Windows,” IEEE 1997 Workshop on Applications of Signal Processing to Audio and Acoustics, New York, 1997.

[10] A. Klapuri. “Multipitch estimation and sound separation by the spectral smoothness principle,” IEEE International Conference on Acoustics, Speech and Signal Processing, Salt Lake City, USA, 2001.

[11] N. F. Fletcher, T. D. Rossing. “The Physics of Musical Instruments” (2nd ed.). Springer–Verlag, New York, 1998.