SEPARATION OF HARMONIC SOUNDS USING LINEAR MODELS FOR THE OVERTONE SERIES

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

A signal processing method is described, which separates harmonic sounds by applying linear models for the overtone series of sounds. Time-varying sinusoidal parameters are estimated in an iterative algorithm which is initialized using a multipitch estimator that finds the number of concurrent sounds and their frequency components. The iterative process then improves the estimates using the least-squares criterion. The harmonic stucture is retained by keeping the frequency ratio of overtones constant over time. Overlapping frequency components are resolved by using linear models for the overtone amplitudes. In practice, the models retain the spectral continuity of natural sounds. Simulation experiments were done using some basic structures for the linear models. These include polynomial, mel-cepstal and frequency-band model. 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]. 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 principles, it is possible to reconstruct separated sounds which are perceptually close to the original ones before mixing.

2. SYSTEM OVERVIEW

The overall structure of the system is presented in Figure 1. The overall approach has been presented in [3], but instead of nonlinear smoothing of the overtone series, linear models and the least-square criterion are studied in this paper.

At first, the input signal is divided into frames and passed to the multipitch estimator. The multipitch estimator takes a single 90–200 ms frame of the acoustic input signal at time and outputs the number of sounds, their fundamental frequencies and rough estimates of the harmonic partials of each sound. The performance of

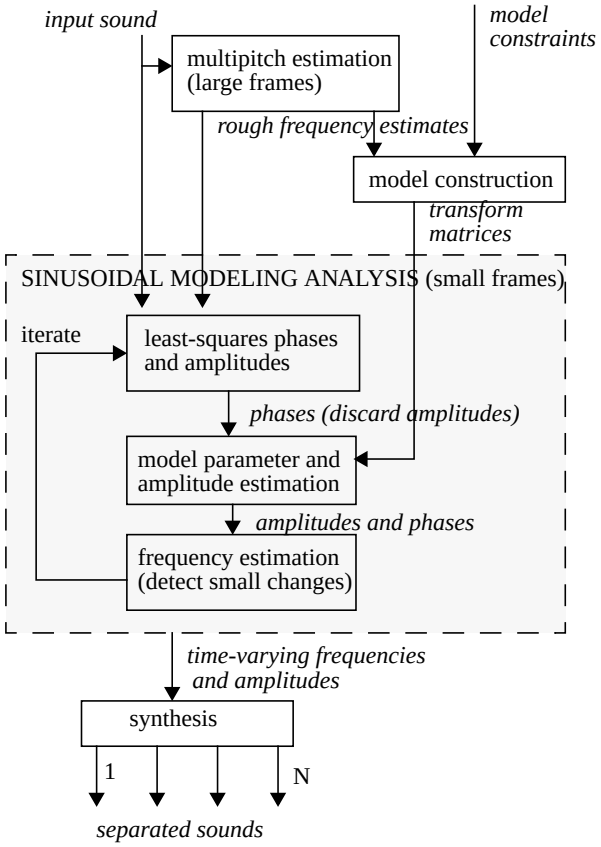

Figure 1. *Block diagram of the separation system.*

the multipitch estimator has been studied in [4]. Harmonic relations of the frequency components were found to be an effective organization force in rich sound mixtures [2].

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.

To estimate the amplitudes of frequency components that overlap with each other, a linear model is used to force the spectral envelope of each sound to be smooth. Using the frequency estimates of the overtone series, linear models are constructed for all the sounds. The models impose some constraints on the amplitude spectra. The models are discussed in Section 5. The sinusoidal

model uses shorter frames than the multipitch estimator, therefore being able to calculate time-varying and significantly more accurate parameter estimates for the components of the sounds proposed by the multipitch estimation system.

Once the parameters of the harmonic components are analyzed in each time frame, the sounds can be synthesized separately. 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 [5].

3. SIGNAL 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

$$\hat{s}(t) = \sum_{n=1}^{N} \sum_{k=1}^{K_n} a_n^k(t) \cos(2\pi f_n^k(t) t + \phi_n^k(t)), \quad (1)$$

where N is the number of mixed sounds, K_n is the number of harmonic components of sound n, and $a_n^k(t)$, $f_n^k(t)$ and $\phi_n^k(t)$ are the amplitude, frequency and phase of the harmonic component k of sound n at time t. The frequencies of one sound are tied together so that the ratio of the frequencies is constant even though the fundamental frequency varies. Perfect harmonicity is not assumed.

For natural sounds, the amplitudes of adjacent frequency components depend on each other, which can be observed as a smoothness of the spectrum. The exact amplitudes of single frequency components do not need to be estimated, since it is sufficient to estimate a lower-order model of the spectrum. This is implemented to the signal model by using a linear model for the amplitude vector a_n of the overtone series of sound n:

$$a_n(t) = X_n y_n(t), \quad (2)$$

where X_n is the transform matrix of the model and y_n is the parameter vector of the sound n at time t. The transform matrix maps the parameters values from a lower-dimensional space to a K_n-dimensional space. This kind of structure itself allows a wide range of different models. Some basic transform matrices were studied. These include a trivial identity matrix which assumes independent frequency components, a mel-filterbank matrix, and polynomial and Fourier-transform matrices. The different models are compared in Section 5. From now on, time is omitted to make the equations more readable.

4. ITERATIVE PARAMETER ESTIMATION

The short-time Fourier transform $\hat{S}(f)$ of the model $\hat{s}(t)$ is given by the parametric expression

$$\hat{S}(f) = \sum_{n=1}^{N} \sum_{k=1}^{K_n} \frac{a_n^k}{2} \left(e^{i\phi_n^k} W(f - f_n^k) + e^{-i\phi_n^k} W(f + f_n^k) \right), \quad (3)$$

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

4.1. 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 [6]:

$$\hat{S}(f) = \sum_{n=1}^{N} \sum_{k=1}^{K_n} [p_n^k R_n^k(f) + p_n^{K+k} R_n^{K+k}(f)], \quad (4)$$

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

$$\begin{cases} p_n^k = \frac{a_n^k}{2} \cos\phi_n^k & k \in [1, K_n] \\ p_n^{K+k} = \frac{a_n^k}{2} \sin\phi_n^k & k \in [1, K_n] \end{cases}, \quad (5)$$

and $R_{n,k}$, the known expressions related to the Fourier transform of the window function are

$$\begin{cases} R_n^k(f) = W(f - f_n^k) + W(f + f_n^k) \\ R_n^{K+k}(f) = i[W(f - f_n^k) - W(f + f_n^k)] \end{cases}. \quad (6)$$

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

$$p = (\Re^H \Re)^{-1} \Re^H S \quad (7)$$

where matrix $\Re$ is composed of column vectors R_k of all the sounds, and the elements of p contain the corresponding parameters.

4.2. Least-square solution for the model parameters

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 7 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 given by the multipitch estimator. For a group of overlapping frequency components, only one amplitude and phase are solved by leaving out all but the strongest components. The amplitudes are discarded, and the least-square phase is used for all the components of the group.

The amplitudes of the components are solved by using the similar least-squares solution for the linear model of Equation 2. For sound n, the relation of phase vector p_n and amplitude vector a_n in Equation 5 can be expressed with a phase matrix P:

$$p_n = P_n a_n,$$

where

$$P_n = \frac{1}{2} \begin{bmatrix} diag(\cos\phi_n^1, \ldots \cos\phi_n^K) \\ diag(\sin\phi_n^1, \ldots \sin\phi_n^K) \end{bmatrix}. \quad (8)$$

Now the Fourier transform of the model can be expressed as:

$$\hat{S}(f) = \sum_{n=1}^{N} R_n p_n = \sum_{n=1}^{N} R_n P_n a_n = \sum_{n=1}^{N} R_n P_n X_n y_n . \quad (9)$$

By denoting the product of matrices R_n, P_n and X_n by

$$G_n = R_n P_n X_n , \quad (10)$$

the Equation 9 can be expressed as a single matrix multiplication by

$$\hat{S}(f) = Gy , \quad (11)$$

where

$$G = \left[G_1 \ \ldots \ G_N \right] \text{ and } y = \begin{bmatrix} y_1 \\ \ldots \\ y_N \end{bmatrix} . \quad (12)$$

The least-squares solution for all the model parameter vectors y is be obtained by:

$$y = (G^H G)^{-1} G^H S . \quad (13)$$

The ampitudes a_n of each sound are obtained by multiplying the transform matrix X_n of the sound by the corresponding components y_n of y:

$$a_n = X_n y_n . \quad (14)$$

4.3. Frequency estimation

The spectrum estimator in Equation 3 is nonlinear in terms of f_k. Depalle and Hélie [6] 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 method was extended for harmonic sounds in [3]. The change of the frequency of the components of a sound was tied to the lowest harmonic, retaining the harmonic structure of a sound and resulting in a more reliable estimation of the frequency, since all the components can be used in estimation. The proposed model-based separation system utilizes exactly same estimation method, and is therefore not presented here.

4.4. Iteration

Successive amplitude and phase estimation, model parameter and amplitude estimation, and frequency estimation stages are repeated. If the frequencies estimates given by the multipitch estimator are good, the number of needed iterations is very low, only two or three. If the frequencies are varying strongly, more iterations are needed, but this decreases the robustness of the system, since the propability to converge to a false direction is possible. 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.

The estimates obtained in previous frames can be used to initialize the calculations in subsequent frames, instead of using the multipitch estimator. This makes the computational complexity of the algorithm practically applicable. However, it was found that to ensure the stability of the frequencies, the frequencies should be initialized in every frame by the rough estimated given by the multipitch estimator.

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. In experiments it was found that in the presence of noise and other sounds, using an average of four most reliable components gives the best estimation of the frequency.

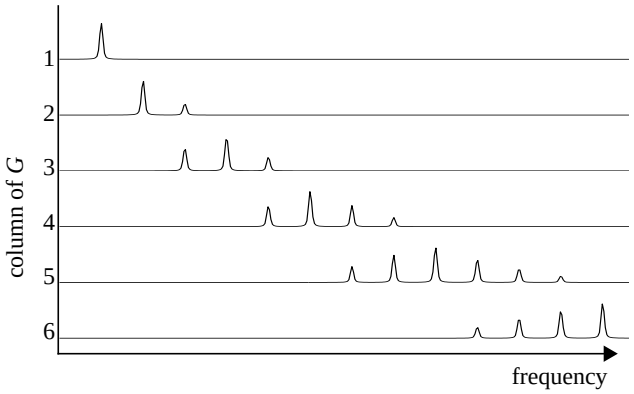

Figure 2. *The magnitudes of six first basis functions of a bandwise model for a perfectly harmonic sound.*

5. TRANSFORM MATRIX AND MODEL PARAMETERS

The proposed method allows a wide range of different models. If the transform matrix is an identity matrix, we get the normal least-squares solution for the amplitudes. If the transform matrix is a vector, the components of which are unity, an equal amplitude is obtained for all the components of each sound. The practical solution is somewhere between these two extreme cases. If there is only one sound present and no interfering sounds, the identity matrix can be used, since it is possible to estimate reliably the amplitudes of individual frequency components. If there is more than one interfering sound present, some of the components are propably overlapping and the rough spectral shape has to be utilized to estimate amplitudes.

Intuitively the easiest way to understand the meaning of the transform matrix is to consider amplitude vector a_n to be represented with a linear combination of the columns of the transform matrix X_n. When phase and shape of the window function are taken into account, this can be extented to the final estimation where the observed spectrum is represented as a linear combination of the columns of matrix G.

Some practical structures for the transform matrix were studied. As a starting point, a combination of a polynomial function plus other elementary funtions (exp, log) was used. For example, a third-order polynomial fit is obtained with the matrix

$$(X)_{p,q} = (f_p)^q , \ q \in [0, 3] \quad (15)$$

where f_p is the frequency of the p:th component. A more perceptually-oriented model is obtained by using a matrix which approximates a critical-band-scale filterbank:

$$(X)_{p,q} = max(min(p2^{1-q} - 1, 2 - p2^{-q}), 0) . \quad (16)$$

This transform matrix is intuitively very applicable, because the model parameters correspond to short-time energies within octave frequency bands. The frequency bands were optimized with the experimental setup described in Section 6. The resulted optimal bands were approximately 2/3-octave bands. The magnitude of six first columns of matrix G for the optimized filterbank-model and a completely harmonic sound are illustrated in Figure 2.

In speech and instrument recognition, mel-cepstral features have proven to be very efficient [7]. A transform matrix which outputs cepstral coefficients was realized by applying a discrete cosine transform to frequency components on an approximated mel-scale.

6. 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 production 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, however, restricting the fundamental frequency over five octaves between 65 Hz and 2100 Hz. A number of two to six simultaneous sounds were allotted, and then mixed with equal mean-square levels. The separation was done on the clean signal and also with additive -10 dB pink noise at the frequency band 50-10 000 Hz.

Acoustic input was fed to the separation algorithm. Separated sounds were synthesized from parameters and compared to the original signals. Since the separation algorithm preserves the phase of the signal, an error between the original and synthesized signal could be obtained simply by subtracting the synthesized signal from the original ones. The mean-square level of the error signal was computed over the whole signal and compared to the original to obtain signal-to-residual ratio (SRR) of the separated signal. The separation system uses frequency band 0-5kHz, thus the comparisons were done on that band only.

Five different linear models were compared, one of which has been previously reported in [3]. The models were the trivial identity model, polynomial model, bandwise model and mel-cepstral model. The previously reported model uses perceptually motivated frequency-domain smoothing [8] instead of a linear model. The order of the models was set to be equal, except for the trivial model, in which the order of the model is the number of harmonic components.

The number of generated test signals was 100 for each polyphony and noise condition. The average SRRs are presented in Table 1. The cases where the multipitch estimator estimates the polyphony or any of the notes wrong were removed, since the objective was to study the performance of the separation models. The percentage of removed signals was 20, 35, 68 and 80 for polyphonies 2 to 4, respectively. The actual note-error rate of the multipitch estimator is significantly lower [4].

For all models, the quality of the outputted sounds decreases gradually as the polyphony increases. Bandwise and mel-cepstral models are able to maintain better quality when the number of sounds is increased. For polyphonies four and five the identity model is clearly the worst, since most of the components are colliding. The polynomial model performs poorly on all polyphonies,

Table 1: Signal-to-residual ratios of five different separation models for polyhonies two to five, averaged over noisy and clean conditions, in dB.

	Polyphony			
Model	2	3	4	5
identity	10.5	8.2	6.4	4.8
polynomial	9.1	8.0	7.3	6.5
frequency bands	10.3	9.2	8.4	7.4
mel-cepstrum	10.6	9.3	8.5	7.3
smoothing	10.7	8.9	7.7	6.2

which is explained by the fact that polynomial function do not fit well the spectra of natural sounds.

Most energy of residuals in SRR estimation is caused by small errors in amplitude or phase estimation. Therefore, 10 dB SRR is usually quite good, if the perceptual quality of the synthesized signals is considered. Some separated test signals are available at http://www.cs.tut.fi/~tuomasv/demopage.html.

7. CONCLUSIONS

The proposed linear models and algorithm enables better sound separation, especially for higher polyphonies. The algorithm decreases the degree of freedom of the overtone series of sound, thus achieving better performance in complex sound mixtures.

8. 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] T. Virtanen, A. Klapuri. “Separation of Harmonic Sounds Using Multipitch Analysis and Iterative Parameter Estimation,” IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New York, U.S.A. 2001.

[4] 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.

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

[6] 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.

[7] Eronen, A. “Automatic Instrument Recognition”, MSc thesis, Tampere University of Technology 2001.

[8] 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.