Note-based sound source separation of polyphonic recordings

KRISTÓF ACZÉL, ISTVÁN VAJK

Budapest University of Technology and Economics, Department of Automation and Applied Informatics
{aczelkri, vajk}@aut.bme.hu

Keywords: polyphonic music, separation, instrument print, energy split

Decomposing a polyphonic musical piece to separate instrument tracks has always been a challenge. Isolating the tracks is out of reach of today's technology. This article proposes a novel method for the separation of monophonic musical recordings. The architecture of the proposed separation system is given. It uses samples of real instruments for regaining the missing data, thereby allowing for the separation and correction of recordings that cannot be retaken.

1. Introduction

Modifying the musical structure of existing polyphonic pieces would create new dimensions in sound processing. By splitting a recording into its source instrument signals we could fix arbitrary notes in any recordings or simply modify the melody of an instrument in a polyphonic piece.

The problem lies in the fact that although it is possible to record a musical event using many microphones, this is not common for various reasons, apart from some exceptions in pop music. Moreover, multitrack recordings are also mixed down to two channels (stereo) in most of the cases, which practically renders any attempts to modify the original tracks useless. After this step the individual notes in the recording cannot be modified, only the whole signal can be altered using different kinds of filters.

In our research we have developed a sound source separation system that allows for the separation of arbitrary note signals from the remaining part of the mixture. The musical notes of interest can be selected by the user, while the other notes remain in the mixture unaltered. This approach makes our separation system particularly applicable for fixing bad notes in existing recordings.

As reliable automatic musical transcription and instrument recognition is out of reach of today's technology, in our work we allow a reasonable amount of user input and processing time to achieve better separation quality. User input involves entering the musical score (note onset/offset, frequencies, used instruments). Due to the nature of real-life music, this input will never be 100% accurate, even if the user is presented with some kind of hint about the concrete recording (e.g. a spectrogram is plotted and shown to the user). However, it can be precise enough for getting a first rough estimate on the note parameters in the recording

There is a great need for complementary information in sound separation in addition to the raw sound signal that is being processed. The complexity of separation lies in the fact that the information we would like to retrieve from the original signal is actually not present. Significant amount of research work has been devoted to regaining the lost information in different ways.

Model based systems represent a very promising approach. In this category a parametric model of the input sources is established that serves as a set of constraints on the output signals. The model parameters are obtained from the mixture itself. The two main branches of this area are rule-based algorithms [1] that use prior information to build the model, and Bayes estimation [2] where prior information is explicitly given using probability density functions. In music applications, the most commonly used approach is sinusoidal modeling which suits the separation of pitched instruments and voiced speech very well [3].

Unsupervised learning methods usually operate on the basis of simple non-parametric models, and require less information on the original sources. They try to gather information on the source signal structures from the mixed data itself using information-theoretical principles, such as statistical independence between the sources. The most common approaches used to estimate the sources are based on independent component analysis (ICA), non-negative matrix factorization (NMF), and sparse coding. These algorithms usually factorize the spectrogram (or other short-time representation of the signal) into elementary components. This is followed by clusterization that builds the separated output channels from the elementary components.

This paper proposes a separation system that is based on the model-driven approach. A global architecture is given for the proposed separation system, while its parts are also discussed in detail. The established model, the *instrument print* is elaborated along with the Simplified Energy Split algorithm that distributes the energy of the mixture between the output channels. The separation system is capable of separating note signals sharing the same fundamental frequency which is unsolved by most of the other separation approaches.

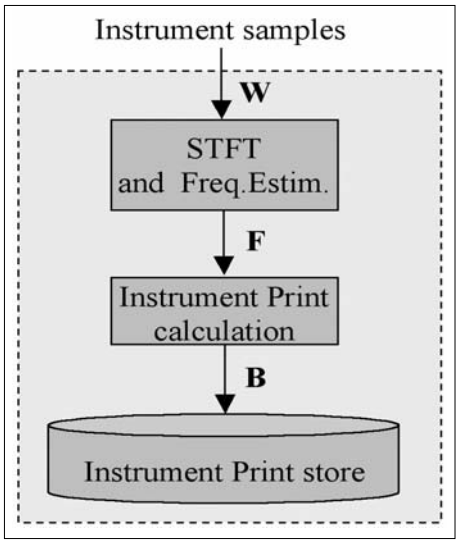


Fig. 1. Block diagram of instrument print creation

W	*Simple waveform*
F	*Frequency Estimated spectrogram:* In addition to the STFT amplitudes ($c_{k,t}$) and phases ($\varphi_{k,t}$) it contains the true frequency ($f_{k,t}^{true}$) of the respective bins.
B	*Bandogram:* A spectrogram split to subbands, in which the energy is summed. Only these sums are stored, no information amplitudes or phase information.

Table 1. Notation of the sound separation block diagram

2. Overview of the separation process

The separation system operates in frequency domain. For that reason all signals have to be transformed between time- and frequency domains at the inputs and outputs of the system.

In addition to the usual STFT we also employ the frequency estimation method proposed by Brown [7] which provides a much more precise spectral image than the standard STFT spectrogram. This method is elaborated in [8] in more detail.

The system can operate in two modes. In the first mode, the *instrument print creation mode,* the system takes sample waveforms from real-life instruments and transforms them to a representation that will later be useful for separation purposes. For this purpose we propose the instrument print model that is based on the bandogram representation of instrument sounds [8]. A bandogram is similar to a spectrogram in many respects, it can be obtained by summing the latter in certain frequency ranges. For separation purposes we need sets of instrument bandograms from each instrument playing in the musical piece to be separated.

Fig. 1 depicts the signal flow and blocks of the process, while the notations used are explained in *Table 1.* Section 3 elaborates the details of bandogram and instrument print creation.

The second operation mode of the system, the *separation mode,* is depicted in *Fig. 2.* It extracts individual note signals from the source recording using three inputs: the original music, the musical score that is entered by the user, and the instrument prints that were created in the first operation mode. The most important blocks are the Simplified Energy Splitter (SES) and the beating-correction. The task of the SES is the redistribution of the energy in the recording between the output channels. The signals created by the SES often suffer from beating. This phenomenon is already present in the original recording; however, it gets noticeable only in the separated signals. The beating correction step targets to eliminate this artifact, which is covered in Section 5.

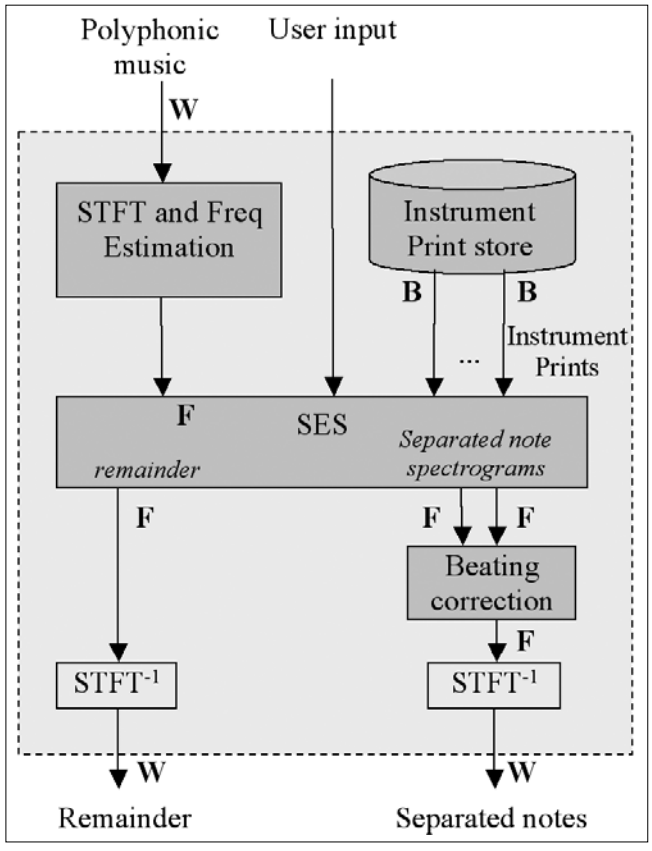
3. Instrument Prints

Even today we do not have complete knowledge about the process of human hearing. Most of the research concludes that our brain stores some kind of memory of instruments [9]. This extra information helps us recognize the melody in a complex mixture. In the process of sound separation as well we need extra information, therefore we try to copy the operation of the human brain. Our instrument model reflects this idea.

The proposed instrument model, the *instrument print,* is a set of instrument samples originating from the same instrument using different frequencies and intonations (blowing strength of the flute, hardness of the piano key hit etc.) Each print may contain one or more orthogonal intonation dimension, depending on how 'freely' a specific instrument can be played.

There may be e.g. a 'warmth' and a 'loudness' dimension for saxophone notes, the values of which range from 1 to 10. These dimensions cannot always be defined in mathematical terms; very often they can only be labeled

Fig. 2. Block diagram of the separation phase



by subjective ones, like the two above. In short, an instrument print is much like a collection of samples of different f_0 fundamental frequencies and different values in the intonation space. It can be illustrated by the following function:

$$\underline{\mathbf{A}}(\underline{\mathbf{M}}, f, f_0, t) \tag{1}$$

where $t, m_x, f_0 \in \mathrm{R}^+$,
$0 < m_x < m_{x,max}$,
$0 \le t < \infty$,
$0 < f, f_0 \le 20000 Hz$.

This function shows how amplitudes change over time (t) over the frequency range (f) for a specific note played on a certain f_0 fundamental frequency and played with intonation $\underline{\mathbf{M}}$.

In reality it is sufficient to store the sum of the amplitudes in certain frequency bands. This representation is called a *bandogram*. The subbands are aligned on a logarithmical frequency scale. A bandogram can be defined as:

$$A_{\mathbf{M},f_0,b,t} = \sum\nolimits_{\rho(f_{k,t}^{true}, f_0, b)} c_{k,t} \,, \tag{2}$$

where $c_{t,k}$ and $f_{t,k}^{true}$ are the amplitude and estimated true frequency of the k[th] component, $\rho(f, f0, b)$ is true if the distance of f and f_0 is exactly b subbands, where b is calculated as

$$b = \left\lfloor \log_{\sqrt[R]{2}} \frac{f_0}{f_{k,t}^{true}} \right\rfloor . \tag{3}$$

The width of the frequency bands is specified by R, that is, the number of bands per octave. In our experiments we concluded that R=12 provides good enough resolution in frequency, and it is also easy to understand as an octave consists of 12 semitones. In reality we cannot store all the possible samples of an instrument. Missing samples can be calculated by interpolation.

4. The separation problem

As the solution for the separation problem is extremely complex, if at all possible, here we propose a simplified solution that makes the separation possible at the expense of slightly lower quality. Let $\underline{\mathbf{c}}_{r\tau} = \left\{ c_{r\tau,k} \cdot e^{\gamma_{r\tau,k}} \right\}$ denote the spectrum of the recording at time $r\tau$ ($r \in \mathbf{N}$), $\underline{\mathbf{s}}_{i,r\tau}^{orig} = \left\{ s_{i,r\tau,k}^{orig} \cdot e^{\sigma_{i,r\tau,k}} \right\}$ and $\underline{\mathbf{d}}_{r\tau} = \left\{ d_{r\tau,k} \cdot e^{\delta_{r\tau,k}} \right\}$ denote the spectrum of the i[th] note and the noise component, respectively. The original separation can be formed as:

$$\underline{\mathbf{c}}_{r\tau} = \sum\nolimits_{\forall i} \underline{\mathbf{s}}_{i,r\tau}^{orig} + \underline{\mathbf{d}}_{r\tau} \,, \tag{4}$$

where

$$c_{r\tau,k}, s_{i,r\tau,k}^{orig}, \sigma_{i,r\tau,k}, \gamma_{r\tau,k} \in \mathrm{R}^+ .$$

As (4) cannot be solved without any further constraints, it will be simplified in a way that the resulting quality loss is barely noticeable. Previous research [10-12] concluded that the human ear is insensitive to the phase information of sound signals as long as phase continuity is guaranteed between subsequent frames.

Based on these findings, (4) can be modified by eliminating the unknown $\sigma_{i,r\tau,k}$ and $\delta_{r\tau,k}$ phases:

$$\gamma_{r\tau,k} = \sigma_{i,r\tau,k} = \delta_{r\tau,k} \,. \tag{5}$$

thereby modifying (4) to the following form:

$$\left| c_{r\tau,k} \right| = \sum\nolimits_{\forall i} \left| s_{i,r\tau,k} \right| + \left| d_{r\tau,k} \right| , \tag{6}$$

where we are looking for the values of $|s_{i,r\tau,k}|$ and $|d_{r\tau,k}|$ for each $i,r\tau$ and k, if $|c_{r\tau,k}|$ and $\gamma_{r\tau,k}$ are known.

Using the proposed simplification results in a slight quality loss: beating caused by notes that are located on close frequencies is not resolved directly. This artifact is handled by a post-processing step.

5. The Simplified Energy Splitter

This section describes the heart of the separation process, the Simplified Energy Splitter. The SES has the task of redistributing the energy in the source recording between the output channels, the separated note signals, using the instrument prints. The right prints are selected using the score, intonation and instrument information given by the user.

We propose the following iterative algorithm for the separation. We start with the spectrogram ($\underline{\mathbf{c}}$) of the original recording. Each output track will be denoted by its $\underline{\mathbf{s}}_i$ spectrogram (zero in the beginning). In each step a fraction of the amplitude content in the selected bandograms is transferred from the amplitude spectrum of the recording to the separated note signals into the right frequency band. The amount of transferred amplitude is a δ fraction of the energy in the used instrument prints. May the recording no longer contain enough amplitude, then the full remaining energy is transferred. δ can be calculated as:

$$\delta = \frac{A_{i,\mathbf{M}_i,f_{0,i},b}(r\tau - T_{onset,i})}{\sum_{\rho(f_{k,r\tau}^{true}, f_0, b)} c_{[j,i],r\tau,k}} \cdot \frac{1}{J} . \tag{7}$$

Each iteration comprises I substeps, where I denotes the number of instruments in the time frame. In one substep we transfer amplitudes to one output channel only. The amplitude content of the recording in the d[th] iteration and i[th] substep is as follows:

$$c_{[j,i+1],r\tau,k} = \begin{cases} \rho(f_{k,r\tau}^{true}, f_0, b) : \max\left(0, (1-\delta) \cdot c_{[j,i],r\tau,k}\right) \\ otherwise: \quad c_{[j,i],r\tau,k} \end{cases} \tag{8}$$

The amplitude content of the i[th] note is:

$$\underline{\mathbf{s}}_{i,[j+1],r\tau} = \underline{\mathbf{s}}_{i,[j],r\tau} + (\underline{\mathbf{c}}_{[j,i-1],r\tau} - \underline{\mathbf{c}}_{[j,i],r\tau}) \,. \tag{9}$$

The reason for using an iterative algorithm is as follows. In the case of one-step amplitude transfer we may encounter cases where the right amount of amplitude is transferred to the track of a loud output note while the amplitude content of the recording decreases to zero, thereby leaving no amplitude for other notes. This case can be avoided by transferring only a fraction of

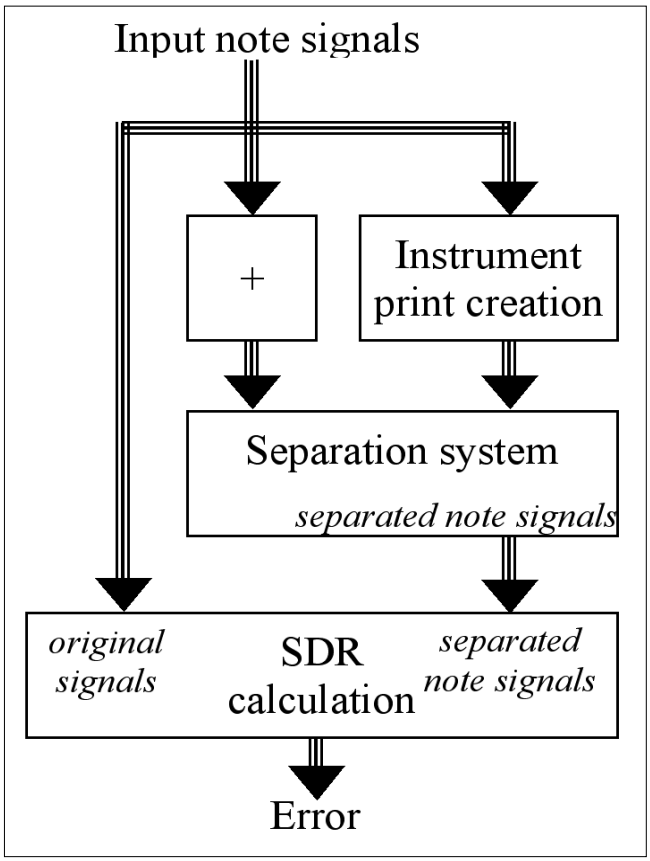


Fig. 3. The test system

the required amplitude in each step, ensuring a fair division of the energy between the outputs. The algorithm is elaborated in earlier publications [8].

Cancellations in the recording are much more audible after separation. However, in most cases this artifact can be eliminated by post processing. By comparing the separated signals to the instrument prints the cancellations can be located and amplified back to the right level.

6. Test results

The performance of the separation system was inspected using synthetic tests. Our test system is depicted in *Fig. 3*. We used the instrument sample collection of the University of Iowa [13] that contains 3841 different samples of string, woodwind, brass and keyboard notes.

In each of our tests a random set of instrument note waveforms were selected. The waveforms were converted to instrument prints. The selected samples were then mixed together and fed to the separation system as the input recording. The separated outputs were then compared to the original samples using two different measures.

The first measure is the conventional Signal-to-Distortion Ratio. The original time-domain signal is subtracted from the output, and this residual is compared to the original signal:

$$SDR_i = 10\log_{10}\frac{\sum_n \tilde{s}_i^{orig}(n)^2}{\sum_n\left[\tilde{s}_i(n)-\tilde{s}_i^{orig}(n)\right]^2}, \qquad (10)$$

where $\tilde{s}_i^{orig}$ and $\tilde{s}_i$ denote the waveform of the original and the i^{th} separated signal, respectively. The second measure operates in frequency domain using the same principle:

$$SDR_i^F = 10\log_{10}\frac{\sum_{\forall r\tau}\sum_{k=0}^{K} s_{i,k}^{orig}(r\tau)^2}{\sum_{\forall r\tau}\sum_{k=0}^{K}\left[s_{i,k}(r\tau)-s_{i,k}^{orig}(r\tau)\right]^2} \qquad (11)$$

The results are shown in *Fig. 4*. In the case of two concurrent notes the performance of the system is 15 dB which slowly degrades as the polyphony increases.

Beside the level of polyphony the other important factor influencing the separation quality is the quality of the instrument prints. Our experiments were repeated using incorrect prints sampled from the same instrument type but not of the very same instrument (e.g. using a different brand of piano). In this case the measured quality was 2 dB lower.

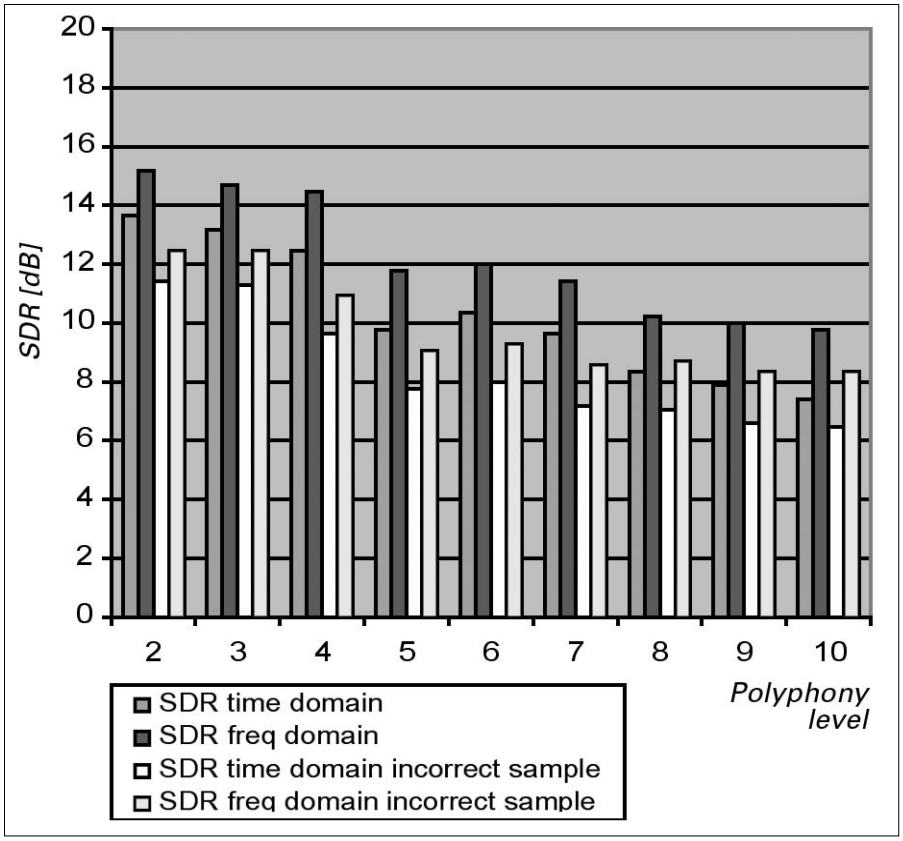


Fig. 4. Test results

7. Summary

We have developed a method for separating the signal of single instrument notes from a recording using pre-recorded instrument prints. The global system architecture of the separation process was given, along with the description of its building blocks. We have established a simple, yet powerful model for storing instrument prints, and the Simplified Energy Splitter was proposed as an algorithm for solving the energy redistribution problem.

We have demonstrated the potential of the system on synthetic and real-life test cases. Simulation experiments on generated mixtures of pitched real-life musical instruments were carried out. In these experiments we obtained an average SDR above 18 dB for two simultaneous sources, and the quality decreased gradually as the level of polyphony increased.

Example waveforms of the synthetic tests as well as real-life separation results can be downloaded from *http://avalon.aut.bme.hu/~aczelkri/separation*.

Authors

KRISTÓF ACZÉL received his degree in Technical Informatics in 2004 from the Budapest University of Technology and Economics. Currently he is a PhD student at the Department of Automation and Applied Informatics doing research in the field of analysis and manipulation of polyphonic music recordings. He is also working as a software research engineer at Nokia Siemens Networks, where he is involved in the design and development of screen, image and document sharing solutions.

ISTVÁN VAJK received the degree in Electrical Engineering in 1975 from the Budapest University of Technology and Economics, Hungary. He was a postgraduate student at the same university, where he obtained his Ph.D. degree in 1977. Since then, he has been with the Faculty of Electrical Engineering and Informatics, working at the Department of Automation and Applied Informatics. He was given the Candidate of Sciences Degree for the implementation problems of adaptive controllers in 1989 and the Doctor of Sciences Degree for the identification from noise measurements using SVD/EVD algorithms in 2007 from the Hungarian Academy of Sciences. Since 1994 he has been the head of Department of Automation and Applied Informatics. His main interest covers the theory and application of control systems, especially adaptive systems and system identification, as well as real-time software engineering.

References

[1] Every, M.R., Szymanski, J.E., "Separation of synchronous pitched notes by spectral filtering of harmonics". IEEE Transactions on Audio, Speech, and Language Processing, Vol. 14, No. 5, pp.1845–1856., 2006.

[2] Cemgil, A.T., "Bayesian Music Transcription", PhD thesis, Radboud University Nijmegen, 2004.

[3] Virtanen, T., "Sound Source Separation in Monoaural Music Signals", PhD thesis, University of Kuopio, 2006.

[4] Mitianoudis, N., Davies, M.E., "Using Beamforming in the audio source separation problem", 7th International Symposium on Signal Processing and its Applications, pp.89–92., 2003.

[5] Smaragdis, P., Brown, J.C., "Non-Negative Matrix Factorization for polyphonic music transcription", IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, pp.177–180., 2003.

[6] Plumbley, M., Abdallah, S., Blumensath, T., Davies, M., "Sparse representations of polyphonic music", EURASIP Signal Processing Journal, Vol. 86, No. 3, pp.417–431., 2006.

[7] Brown, J.C., Puckett, M.S., "A high resolution fundamental frequency determination based on phase changes of the Fourier Transform", J. of the Acoustical Society of Am., Vol. 94, No. 2, pp.662–667., 1993.

[8] Aczél, K., Vajk, I., "Note separation of polyphonic music by energy split", WSEAS International Conference on Signal Processing, Robotics and Automation, pp.208–214., 2008.

[9] McAdams, S., "Recognition of Auditory Sound Sources and Events. Thinking in Sound: The Cognitive Psychology of Human Audition", Oxford University Press, 1993.

[10] Zwicker, E., Flottorp, G., Stevens, S.S., "Critical band width in loudness summation", J. of the Acoustical Society of Am., Vol. 29, pp.548–557, 1957.

[11] Smith, S.W., The Scientist and Engineer's Guide to Digital Signal Processing, California Technical Publishing, 1997.

[12] Edler, B., Purnhagen, H., "Parametric Audio Coding", IEEE International Conference on Communication Technology, Vol. 1, pp.614–617., 2000.

[13] The University of Iowa, Musical Instrument Samples Database (2008.07.07), http://theremin.music.uiowa.edu