Shifted NMF using an Efficient Constant-Q Transform for Monaural Sound Source Separation

Rajesh Jaiswal†, Derry Fitzgerald†, Eugene Coyle† and Scott Rickard*

†*Department of Electrical Engineering*
Audio Research Group
Dublin Institute of Technology

**Department of Electronic Engineering*
University College Dublin

E-mail: rajesh.jaiswal@student.dit.ie

Abstract — **Non-negative Matrix Factorisation (NMF) based algorithms have found application in monaural audio source separation due to their ability to factorize audio spectrogram into additive parts-based basis functions, which typically corresponds to individual notes or chords in music. These separated basis functions are usually greater in number than the active sources, hence clustering is needed for individual source signal synthesis. Although, many attempts have been made to improve the clustering of the basis functions to sources, much research is still required in this area. Recently, Shifted NMF based methods have been proposed as a means to avoid clustering these pitched basis functions to sources. However, the Shifted NMF algorithm uses a log-frequency spectrogram with a fixed number of frequency bins per octave which compromises the quality of separated sources. We show that by replacing the method used to calculate the log-frequency spectrogram with a recently proposed invertible Constant Q Transform (CQT), we can considerably improve the separation quality of the individual sound signals.**

Keywords — **Frequency basis function, shifted NMF, Single channel Blind Source Separation.**

I Introduction

Single channel sound source separation (SSS) is a process in which individual sound sources from a monaural audio mixture are estimated. Notably, separating sound sources from a single channel audio mixture has long been a hard problem to solve due to the complex overlapping of sources in time and frequency. However, segregating sound sources in a mono signal would assist in many audio applications which involve editing and manipulation of audio data such as pitch modification and automatic music transcription.

Estimation of individual sources from a monophonic mixture is usually done using a time-frequency representation of the time-domain audio signal $x(n)$. The widely used Short-time Fourier Transform (STFT) is typically used to obtain a magnitude spectrogram $\mathbf{X}$ for further processing. In SSS algorithms, matrix factorisation [5, 6] of the magnitude spectrogram is usually used as the first processing step to analyse the factorised data for various audio applications. Given a discrete-time signal $x(n)$, the complex spectrogram ($\hat{\mathbf{X}}$) is obtained using the STFT.

NMF [1] is a factorisation technique which approximately decomposes a magnitude spectrogram into additive parts-based decompositions known as basis functions where individual basis functions typically correspond to a note or chord. In other words, NMF approximately factorises a non-negative matrix $\mathbf{X}$ of size $n \times m$, i.e. magnitude spectrogram, into factors $\mathbf{A}$ and $\mathbf{B}$ such that

$$\mathbf{X} \approx \mathbf{AB} \tag{1}$$

where $\mathbf{A}$ is $n \times r$ matrix and $\mathbf{B}$ is $r \times m$ matrix, with $r < n, m$. Matrix $\mathbf{A}$ contains r frequency basis functions and the corresponding time activation functions are in matrix $\mathbf{B}$. The commonly used Kullback-Leibler (KL) divergence cost-function is

used to find the basis function as documented in [1].

$$D(A||B) = \sum_{i,j}(A_{ij}log\frac{A_{ij}}{B_{ij}} - A_{ij} + B_{ij}) \quad (2)$$

The NMF basis functions approximately represents the spectral envelop of individual notes played by the instrument in piece of music. Thus, each NMF basic function can be used to reconstruct the original note. However, for real world music mixtures, the number of notes played by each instrument or source considered is not constrained to one. As a result, this separation method results in multiple notes per instrument. Hence, clustering of the basis functions into a number of active sources is required to synthesize the separate sources. Much research has been carried out to cluster these basis functions into sources with considerable success [5, 6].

Shifted Non negative Matrix factorisation (SNMF) was proposed by Fitzgerald et al [9] as a means of avoiding the problem of clustering provided that a log-frequency spectrogram is used to obtain the basis function. The SNMF algorithm [9] assumes that the timbre of a note does not change for all the pitches produced by an instrument. Therefore, if a log-frequency spectrogram is used, the frequency basis function of one note can be used to approximate that of another note by translating the frequency basis function up or down in frequency as required. However, the use of a log-frequency spectrogram comes at a price, as will be seen in section II. There is no true inverse to a log-frequency spectrogram. This can adversely affect the sound quality of the separated signal. However, recently a method was proposed in [4] which allows for an improved inverse transform from log-frequency spectrograms. It is proposed to use this in conjunction with SNMF to determine if improved sound quality in the separated sources can be obtained.

The structure of the paper is as follows. Section II briefly outlines the the system model which include the Constant Q transform, the SNMF Algorithm and the signal reconstruction. Section III gives an overview of Experimental set-up followed by results in section IV.

II System model

Figure 1 shows the signal flow in the system model. A test mixture in time domain is first converted into the constant Q domain using the CQT. Thereafter, the shift-invariant property of SNMF is used to determine the instrument basis functions. Furthermore, spectral masking is incorporated to improve the quality of separation. Finally, the separated signals are recovered by the inverse CQT. We will first explain the basic principle involved in calculating the CQT as it would assist in understanding other features in the system model.

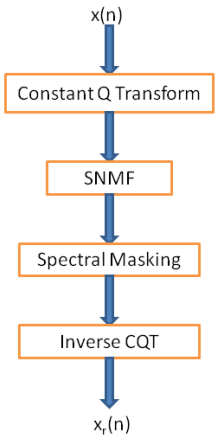

Fig. 1: Signal flowchart of the System model

a) Constant Q spectrogram

A log-frequency spectrogram can be obtained by using the Constant Q Transform, as proposed by Brown and Pluckette [2]. The CQT can be obtained by modulating the input signal with a bank of complex exponentials, whose centre frequencies are geometrically spaced. In the case of Western music, the fundamental frequencies of adjacent notes are geometrically spaced by a factor of $\sqrt[12]{2}$ and so this spacing is used when calculating the CQT.

The CQT can be efficiently calculated in the Fourier domain by taking the transform of the complex exponentials used to obtain the CQT, yielding a transform matrix $\mathbf{Y}$. This results in a matrix where many of the transform coefficients are near zero, and so can be discarded, yielding a sparse matrix. The CQT can then be obtained by multiplying a linear frequency domain spectrogram $\mathbf{P}$ by the sparse matrix $\mathbf{Y}$ as shown in equation:

$$\mathbf{Q} = \mathbf{YP} \quad (3)$$

However, as $\mathbf{Y}$ is not a square matrix, no true inverse of the CQT is possible. An approximate inverse transform was proposed by Fitzgerald [7] with the assumption that the music signals can be sparsely represented in the linear frequency domain. However, the assumption does not hold good for all audio signals and the algorithm was extremely slow in calculating the inverse CQT transform. Recently, Schörkhuber and Klapuri [4] has proposed an extension to the method discussed in

[2, 3] to calculate the CQT in a manner which allows a high quality inverse CQT to be calculated. The algorithm processes each octave in the signal one by one starting from highest to lowest to calculate the CQT coefficients of a given spectrogram. A separate spectral transform matrix $\mathbf{Y}$ is determined which represents frequency bins separated by a maximum of one octave.

Having obtained a linear spectrogram, $\mathbf{X}$, the Constant Q spectrogram is obtained by using the equation 4:

$$\mathbf{C} = \mathbf{YX} \tag{4}$$

where $\mathbf{Y}$ is the spectral transform matrix. The spectral transform matrix remains constant for all the octaves considered for the CQT transform. For the calculation of the CQT coefficients of the highest octave the input signal $x(n)$ is used. Thereafter, for the d^{th} successive octave, $x(n)$ is downsampled by a factor of 2^d. Following the terminology of [4], let $x_d(n)$ represents a signal generated by decimating $x(n)$ by 2^d times. And let $\mathbf{X}_d$ contains the DFT values of $x_d(n)$. Then, the CQT coefficients $\mathbf{C}_d$ for octave d is obtained as

$$\mathbf{C}_d = \mathbf{YX}_d \tag{5}$$

This new method to calculate the CQT coefficients for each octave gives a better inverse CQT than the previously proposed methods because of the following reasons. For b_o number of CQT bins, the signal is repetitively analysed from higher octaves to lower octaves to obtain the CQT coefficients which increases the redundancy of the transform. This increase in redundancy helps in capturing most of the data features required for to obtain a high quality inverse CQT. The redundancy,R_f, is directly proportional to the highest frequency analyzed i.e. the highest octave chosen. The separation quality can be further improved by increasing the number of CQT bins per octave, B_o. However, R_f and B_o are optimally chosen for computational efficiency. An analysis of quality of reconstruction as a function of R_f and B_o can be found in [4].

A complete implementation of CQT can be found in [4]. In this paper, we have used the MATLAB toolbox of the reference implementation of the above discussed method provided at http://www.elec.qmul.ac.uk/people/anssik/cqt/ to obtain the Constant Q spectrogram.

b) Shifted Non-negative Matrix Factorisation

As noted previously Shifted NMF assumes that a given note basis function can be approximated by translating another note basis function up or down in frequency, provided a log frequency transform, such as the Constant Q Transform is used. Therefore, instead of capturing a single basis function per note of each instrument, SNMF attempts to learn a single basis function per instrument. This single basis function is then translated up or down in frequency to capture all the notes played by an instrument. As SNMF makes use of tensors, we now define the notation used for the tensor parameters in the SNMF model.

The notations for tensor parameters used to define the SNMF model [9] is as per the conventions described in [12]. Calligraphic upper-case letters ($\mathcal{R}$) are used to denote tensors of any given dimension. The contracted product of the two tensors of finite dimension results in a tensor that is a bilinear mapping of individual elements of two tensors in consideration. Let a tensor $\mathcal{R}$ be of dimension $I_1 \times \cdots \times I_S \times L_1 \times \cdots \times L_P$ and tensor $\mathcal{D}$ be of dimension $I_1 \times \cdots \times I_S \times J_1 \times \cdots \times J_N$ then equation 6 denotes the contracted tensor multiplication of $\mathcal{R}$ and $\mathcal{D}$ along the first S modes.

$$\langle \mathcal{RD} \rangle_{\{1,\ldots,S;1,\ldots,S\}} = \sum_{i_1=1}^{I_1} \cdots \sum_{i_1=1}^{I_1} \mathcal{R} \times \mathcal{D} = \mathcal{Z} \tag{6}$$

The dimensions along which the tensors $\mathcal{R}$ and $\mathcal{D}$ are to be multiplied are specified in curly brackets. The resultant tensor $\mathcal{Z}$ will be of dimension $L_1 \times \cdots \times L_P \times J_1 \times \cdots \times J_N$. Indexing of a given tensor is done using lower case letters, such as i and is denoted by $\mathcal{R}(i, j)$.

Having obtained the Constant Q spectrogram $\mathbf{C}$ of size $n \times m$, where m is the number of time frames along the n frequency bins, SNMF can be used to separate the instrument basis functions. In practice, for a given number of r sources the CQT spectrogram $\mathbf{C}$ can be decomposed using the SNMF model as shown in equation 7 :

$$\mathbf{C} \approx \langle \langle \mathcal{RD} \rangle_{\{3,1\}} \mathcal{H} \rangle_{\{2:3,1:2\}} \tag{7}$$

Here, $\mathcal{R}$ is a translation tensor of dimension $n \times k \times n$ for k possible translations. $\mathcal{R}$ translates the instrument basis functions in $\mathcal{D}$ up or down to approximate various notes played by the instrument in question. Then tensor $\mathcal{D}$ of size $n \times r$ contains instrument basis functions for each source. $\mathcal{H}$ is a tensor of size $k \times r \times m$ such that $\mathcal{H}(i\,,\,j\,,\,:)$ represents the time envelope for the i^{th} translation of the j^{th} source, which indicates when a given note is played by a particular instrument.

The cost function used to estimate tensors $\mathcal{D}$ and $\mathcal{H}$ is same as used for NMF and iterative multiplicative update equations can be derived in a manner similar to [9]. To approximately cover all the notes and chords of the instrument, the number of translation k is chosen empirically. The

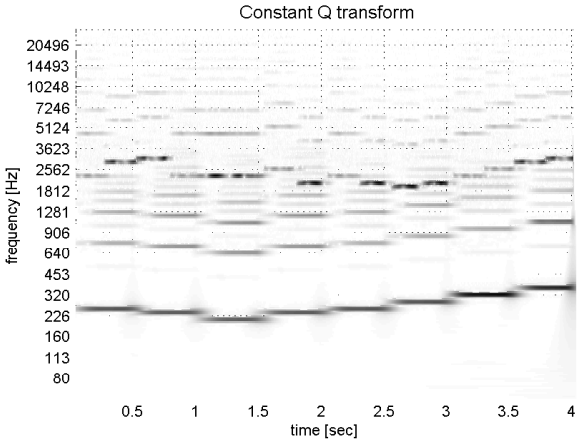

Fig. 2: Frequency basis function of input mixture in constant Q domain.

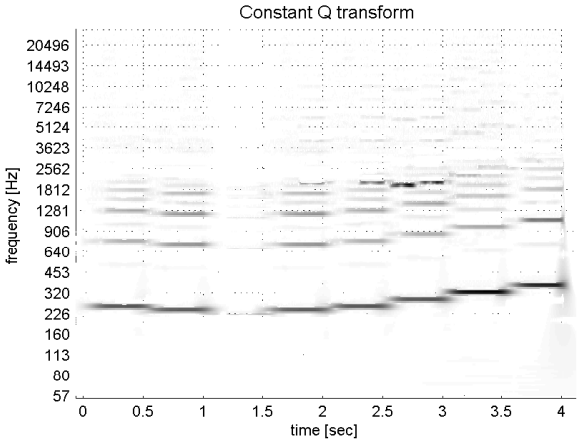

Fig. 3: Separated Source 1

translated (frequency-shifted) version of the instrument basis function approximately captures all the notes played by the instrument considered in a mixture, and so the need for clustering of the NMF basis functions is avoided.

The SNMF algorithm has two notable drawbacks. Firstly, the spectral envelope of notes played by an instrument changes with the pitch, therefore, the assumption that the timbre of any note played by an instrument remains unchanged is not true in general. Secondly, the lack of an inverse CQT results in compromising on the separation quality of the reconstructed signal. However, the shift-invariant property of instrument basis function can be exploited to capture all the notes played by pitched instruments in the audio mixture.

As an example of SNMF using the CQT, figure 2 shows a CQT spectrogram of the input mixture. Figure 3 and 4 show the Constant Q spectrogram of the separated instruments obtained via SNMF. It can seen through visual inspection that the instruments have separated well, though one note has been incorrectly separated.

c) Spectral masking and Signal reconstruction

The SNMF model generates two optimised tensors $\mathcal{D}$ and $\mathcal{H}$ as shown in equation 7. The individual source spectrogram C_r of the r^{th} source is obtained by using the slices as shown in equation 8:

$$\mathbf{C}_r = \langle\langle \mathcal{R}\mathcal{D}(:,r)\rangle_{\{3,1\}}\mathcal{H}(:,r,:)\rangle_{\{2:3,1:2\}} \tag{8}$$

where $\mathcal{D}(:,r)$ and $\mathcal{H}(:,r,:)$ are the slices of tensors $\mathcal{D}$ and $\mathcal{H}$ respectively. Furthermore, to improve the separation quality, an individual mask was created for each of the source spectrograms $\mathbf{C}_r$. A masking filter $\hat{\mathbf{M}}_\mathbf{r}$ can be calculated as shown in equation 9:

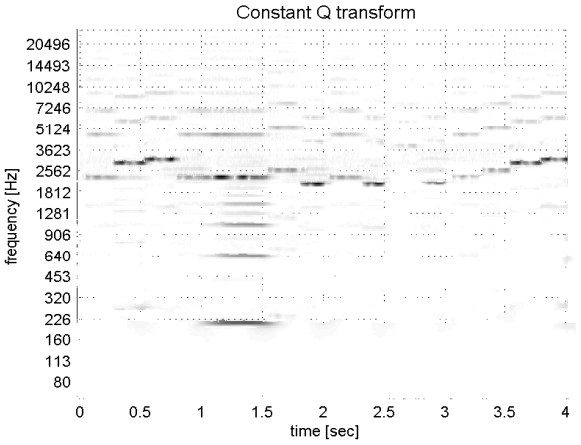

Fig. 4: Separated Source 2

$$\hat{\mathbf{M}}_r = \left(\mathbf{C}_r^2 \oslash \sum^{r} \mathbf{C}_r^2\right) \tag{9}$$

$\oslash$ and $\otimes$ indicate element-wise division and multiplication in equations 9 and 10. The generated masks are then applied to $\mathbf{C}$ to obtain the filtered source spectrograms $\hat{\mathbf{C}}_\mathbf{r}$:

$$\hat{\mathbf{C}}_\mathbf{r} = \mathbf{C} \otimes \hat{\mathbf{M}}_r \tag{10}$$

The recovered source spectrograms are then converted into time domain signal by inverse CQT [4].

III Experimental Set-up

It is proposed to evaluate the performance of SNMF using the original inverse CQT against SNMF using the inverse CQT proposed in [4]. Both these methods were evaluated using a set of 25 monaural input mixtures of 2 instruments. The

25 test signals were generated by using a huge library of orchestral samples and were comprised of mixtures of melodies from a total of 15 different orchestral instruments [13]. The input mixtures were of 4 to 8 seconds in length and were sampled at a rate of 44.1 kHz. To approximate real world signals, it is necessary to have overlapping harmonics and notes played by the individual instruments which overlap in time in the test samples. The test set used was designed with these requirements in mind. The input mixtures were the result of the extensive collection of pitches, ranging from as low as 87 Hz up to 1.5 kHz and all the notes played by the individual instruments in the audio mixtures were in harmony. In some cases, notes of same pitch were played simultaneously by both the instruments in the mixture. The process by which the database was created is detailed in [8].

For the SNMF model, the number of sources was set equal to 2. The SNMF algorithm was run for 50 iterations. An individual spectrogram of each separated source was then obtained through spectral masking followed by reconstruction of the source signal as explained in section II.

For the given 25 test mixtures, the number of allowable translations, k, was varied between 4 and 9. Multiple tests were run for the different number of allowable time shifts and the separated sources with highest separation quality were picked. Figure 5 shows the audio waveforms in the time domain of one test mixture signal and its corresponding separated sources. It can be seen from the waveforms that the original and reconstructed signals closely match with each other. The melodies played by both the sources were found to be separated well. The algorithm also separated the same notes simultaneously played by both the pitched instruments in the mixture with some interference of harmonics related to that note. Thus, the SNMF algorithm discussed in this paper can be used to separate sound sources in a single channel mixture. Performance evaluation of the SNMF algorithms, to measure the separation quality, is carried out in the next section.

IV Results

The original unmixed signals created from the sample library [13] were used as a reference to measure the performance of the SNMF algorithm. The widely used quality measures, signal-to-distortion (SDR), signal-to-interference ratio (SIR) and signal-to-artifacts ratio (SAR) were used for the performance evaluation of the discussed SNMF algorithms. SDR refers to the signal-to-distortion ratio which measures the amount of distortion present in the reconstructed signal, SIR calculates the interference of other sound sources in the separated signal and SAR

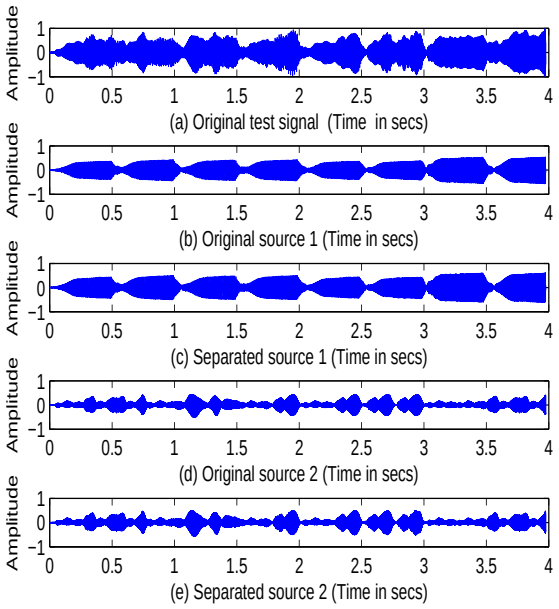

Fig. 5: Original and reconstructed signals in the time domain. The figure shows (a) the original test signal, (b) the original source 1, (c) the separated source 1, (d) the original source 2, (e) the separated source 2.

determines the artifacts present in the separated signal as a result of data processing and reconstruction.

Here, we test the performance of SNMF using the original CQT against that of SNMF using the recently proposed version of the CQT. $SNMF_{cqt,old}$ refers to the SNMF algorithm which uses a reasonable quality approximate inverse CQT to get the desired separated signal. Details on $SNMF_{cqt,old}$ algorithm can be found in [9]. $SNMF_{cqt,new}$ represents the SNMF algorithm using the new inverse CQT as discussed in this paper. Quality measures for both the listed algorithms were calculated. Table 1 shows the quality measures calculated for $SNMF_{cqt,old}$ and $SNMF_{cqt,new}$. All the results i.e. mean SDR, SIR and SAR, are in dB. Both the SNMF algorithms, $SNMF_{cqt,old}$ and $SNMF_{cqt,new}$, are coded in MATLAB and are tested with the same set of test mixtures, as discussed in section III. From table 1, we can see an improvement of mean SDR of more than 10 dB by using the new CQT in the SNMF algorithm for the same set of audio mixture and test parameters. On listening to the separated signals, the sources can be clearly identified with only small artifacts. These artifacts are due to interference of melodies played by one source with the other in the mixture. Overall, it can be stated that by replacing the method to calculate the CQT in the SNMF algorithm, the separation quality considerably improves. Hence, the al-

clustering	SDR	SIR	SAR
$SNMF_{cqt,old}$	-1.85	14.97	3.46
$SNMF_{cqt,new}$	10.88	25.44	11.47

Table 1: Calculated mean SDR, SIR and SAR for separated sound sources

gorithm $SNMF_{cqt,new}$ outperforms the monaural source separation algorithm $SNMF_{cqt,old}$ demonstrating the utility of the new inverse CQT technique.

V Conclusion

We have presented a sound separation technique that uses the shift-invariant property of SNMF algorithm to separate out sound sources present in a monaural mixture. We have briefly explained the basic principle of the SNMF algorithm and necessity of using a log-frequency spectrogram with SNMF. We, then described the drawbacks of using log-frequency spectrograms. We have further showed how a new method of calculating log-frequency spectrograms overcomes some of these problems. We, then demonstrated that using this new method in conjunction with SNMF results in improved sound quality of the separated sources.

VI Acknowledgement

The authors wish to acknowledge the Dublin Institute of Technology (Dublin, Ireland) for funding under the ABBEST scholarship programme.

References

[1] D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorisation, " *Advances in Neural Information Processing System*, 2000, pp. 556-562.

[2] J. C. Brown, "Calculation of a Constant Q spectral transform," *Journal of the Acoustic Society of America*, vol. 89, no.1, pp 425-434, 1991.

[3] J. C. Brown and M. S. Puckette,"An efficient algorithm for the calculation of a Constant Q Transform," *Journal Acoustic Society America*, vol. 92, no. 5, pp. 26982701, 1992.

[4] A. Klapuri, C. Schörkhuber, "Constant-Q Transform toolbox for music processing," *7th Sound and Music Computing Conference*, Barcelona, Spain, 2010.

[5] M. Spiertz and V. Gnann, "Source-Filter based clustering for monaural blind source separation," *Proceedings of the* 12^{th} *International Conference on Digital Audio Effects*, Italy, 2009.

[6] T. Virtanen, "Sound Source Separation Using Sparse Coding with Temporal Continuity Objective," *International Computer Music Conference*, 2003.

[7] D. FitzGerald, M. Cranitch and M. T. Cychowski, "Towards an inverse Constant Q Transform," *in 120th Audio engineering Society Convention*, Paris, France, 2006

[8] D. FitzGerald, M. Cranitch and E. Coyle, "Extended Non-negative Tensor Factorisation Models for Musical Sound Source Separation," *Computational Intelligence and Neuroscience*, Hindawi Publishing Corp., 2008.

[9] D. FitzGerald, M. Cranitch and E. Coyle, "Shifted Non-negative matrix factorisation for sound source separation," *IEEE Workshop of Statistical Signal Processing, Bordeaux*, France, 2005.

[10] E. M. Burns, "Intervals, Scales and Tuning," *The Psychology of Music*, D. Deutsch, Ed. Academic Press, 1999.

[11] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech and Language Processing*, 2006.

[12] B. W. Bader and T. G. Kolda, "Algorithm 862: MATLAB tensor classes for fast algorithm prototyping," *ACM Transactions on Mathematical Software*, vol. 32, no. 4, pp. 635-653, 2006.

[13] P. Siedlaczek, "Advanced Orchestra Library Set," 1997.