Monaural Music Source Separation: Nonnegativity, Sparseness, and Shift-Invariance

Minje Kim and Seungjin Choi

Department of Computer Science
Pohang University of Science and Technology
San 31 Hyoja-dong, Nam-gu
Pohang 790-784, Korea
{minjekim,seungjin}@postech.ac.kr

Abstract. In this paper we present a method for polyphonic music source separation from their monaural mixture, where the underlying assumption is that the harmonic structure of a musical instrument remains roughly the same even if it is played at various pitches and is recorded in various mixing environments. We incorporate with *nonnegativity, shift-invariance*, and *sparseness* to select representative spectral basis vectors that are used to restore music sources from their monaural mixture. Experimental results with monaural instantaneous mixture of voice/cello and monaural convolutive mixture of saxophone/viola, are shown to confirm the validity of our proposed method.

1 Introduction

The nonnegative matrix factorization (NMF) [1] or its extension such as nonnegative matrix deconvolution (NMD) [2] and sparse coding [3], was shown to be useful in polyphonic music description [4, 5], in the extraction of multiple music sound sources [2, 6], and in general sound classification [7]. Some of these methods regard each note as a source, which might be appropriate for music transcription and work for source separation in a very limited case.

In this paper we present a method for monaural polyphonic music separation, the goal of which is to restore the whole melody generated by each musical instrument from a single channel mixture of several polyphonic musical sounds. We assume that the harmonic structure of a musical instrument approximately remains the same, even if it is played at different pitches and is recorded in different environments. Different musical instruments are assumed to have different spectral characteristics (harmonic structure).

The main idea is to select a few representative spectral basis vectors in the auditory spectrogram of measurement data, assuming that there are some sections in the auditory spectrogram where only a single note from a single source appears. Rather than learning basis vectors, we select a few appropriate nonnegative basis vectors using the sparseness of spectral coefficients. These shift-invariant nonnegative basis vectors are fixed and associated encoding variables are learned by the overlapping NMF [8] which incorporates with the shift-invariant representation, in order to restore music sources. The method is related

to our earlier work [9] and the generalized prior subspace analysis [10]. However, the key distinction lies in a way of selecting shift-invariant basis vectors. Promising results with monaural instantaneous mixture of voice/cello and convolutive mixture of saxophone/viola, are presented to confirm the validity of our proposed method.

2 Overlapping NMF: Nonnegativity and Shift-Invariance

Nonnegative matrix factorization (NMF) is a simple but efficient factorization method for decomposing multivariate data into a linear combination of basis vectors with nonnegativity constraints for both basis and encoding matrix [1].

Given a nonnegative data matrix $\boldsymbol{V} \in \mathbb{R}^{m \times N}$ (where $V_{ij} \geq 0$), NMF seeks a factorization

$$\boldsymbol{V} \approx \boldsymbol{W}\boldsymbol{H}, \tag{1}$$

where $\boldsymbol{W} \in \mathbb{R}^{m \times n}$ $(n \leq m)$ contains nonnegative basis vectors in its columns and $\boldsymbol{H} \in \mathbb{R}^{n \times N}$ represents the nonnegative encoding variable matrix. Appropriate objective functions and associated multiplicative updating algorithms for NMF can be found in [1].

The overlapping NMF is an interesting extension of the original NMF, where transform-invariant representation and a sparseness constraint are incorporated with NMF [8]. Some of basis vectors computed by NMF could correspond to the transformed versions of a single representative basis vector. The basic idea of the overlapping NMF is to find transformation-invariant basis vectors such that fewer number of basis vectors could reconstruct observed data. Given a set of transformation matrices, $\mathcal{T} = \left\{\boldsymbol{T}^{(1)}, \boldsymbol{T}^{(2)}, \ldots, \boldsymbol{T}^{(K)}\right\}$, the overlapping NMF finds a nonnegative basis matrix $\boldsymbol{W}$ and a set of nonnegative encoding matrix $\left\{\boldsymbol{H}^{(k)}\right\}$ (for $k = 1, \ldots, K$) which minimizes

$$\mathcal{J}(\boldsymbol{W}, \boldsymbol{H}) = \frac{1}{2}\left\|\boldsymbol{V} - \sum_{k=1}^{K} \boldsymbol{T}^{(k)}\boldsymbol{W}\boldsymbol{H}^{(k)}\right\|_F^2, \tag{2}$$

where $\|\cdot\|_F$ represents Frobenious norm. The multiplicative updating rules for the overlapping NMF were derived in [8], which are summarized below.

Algorithm Outline: Overlapping NMF [8]

Step 1 Calculate the reconstruction: $\boldsymbol{R} = \sum_{k=1}^{K} \boldsymbol{T}^{(k)}\boldsymbol{W}\boldsymbol{H}^{(k)}$.
Step 2 Update the encoding matrix by

$$\boldsymbol{H}^{(k)} \leftarrow \boldsymbol{H}^{(k)} \odot \frac{\boldsymbol{W}^\top \left[\boldsymbol{T}^{(k)}\right]^\top \boldsymbol{V}}{\boldsymbol{W}^\top \left[\boldsymbol{T}^{(k)}\right]^\top \boldsymbol{R}}, \quad k = 1, \ldots, K, \tag{3}$$

where $\odot$ denotes the Hadamard product and the division is carried out in an element-wise fashion.

Step 3 Calculate the reconstruction $\boldsymbol{R}$ again using the encoding matrix $\boldsymbol{H}^{(k)}$ updated in Step 2, as in Step 1.

Step 4 Update the basis matrix by

$$\boldsymbol{W} \leftarrow \boldsymbol{W} \odot \frac{\sum_{k=1}^{K} \left[\boldsymbol{T}^{(k)}\right]^{\top} \boldsymbol{V} \left[\boldsymbol{H}^{(k)}\right]^{\top}}{\sum_{k=1}^{K} \left[\boldsymbol{T}^{(k)}\right]^{\top} \boldsymbol{R} \left[\boldsymbol{H}^{(k)}\right]^{\top}}. \tag{4}$$

3 Spectral Basis Selection: Sparseness

The goal of spectral basis selection is to choose R representative vectors $\boldsymbol{V}_r = [\boldsymbol{v}_{r_1} \cdots \boldsymbol{v}_{r_R}]$ (R is the number of music sources) from $\boldsymbol{V} = [\boldsymbol{v}_1 \cdots \boldsymbol{v}_N]$ where $\boldsymbol{V}$ is the data matrix associated with the spectrogram of mixed sound. Each column vector $\boldsymbol{v}_t$ corresponds to the power spectrum of the mixed sound at time $t = 1, \ldots, N$. Selected representative vectors are fixed as basis vectors that are used to learn an associated encoding matrix set through the overlapping NMF with sparseness constraint, in order to restore unmixed musical sound.

Our spectral basis selection method is based on the assumption that there are some sections where only a single note from a single source appears. In the spectrogram of mixed sound, solo sections are searched partly through the sparseness value of $\boldsymbol{v}_t$ over time. Our earlier work can be found in [9].

Fig. 1 shows the schematic diagram of the spectral basis selection method, consisting of two parts. The first part is to select several candidate vectors $\boldsymbol{V}_c = [\boldsymbol{v}_{c_1}\, \boldsymbol{v}_{c_2} \cdots \boldsymbol{v}_{c_K}]$ from $\boldsymbol{V}$ using a sparseness measure and a clustering-elimination method. The second part involves determining representative basis vectors from candidate vectors, through the overlapping NMF. More detailed description is summarized below.

Part 1

1. Sparseness calculation: We calculate the sparseness value for input vectors $\boldsymbol{v}_t$ for $t = 1, \ldots, N$, using the measure in [11],

$$\xi_t = \text{sparseness}(\boldsymbol{v}_t) = \frac{\sqrt{m} - (\sum_i |v_{it}|)/\sqrt{\sum_i v_{it}^2}}{\sqrt{m} - 1}, \tag{5}$$

where v_{it} is the ith element of the m-dimensional vector $\boldsymbol{v}_t$.

2. Normalization: We normalize input vectors $\boldsymbol{v}_t$ for $t = 1, \ldots, N$ such that each vector has unit Euclidean norm,

$$\boldsymbol{v}_t \leftarrow \frac{\boldsymbol{v}_t}{\|\boldsymbol{v}_t\|}. \tag{6}$$

3. Alignment: We calculate the index $f_i = t^*$ which involves the largest sparseness value among $\{\xi_t\}_{t=1}^{N}$, i.e,

$$t^* = \underset{1 \le t \le N}{\arg\max}\, \xi_t. \tag{7}$$

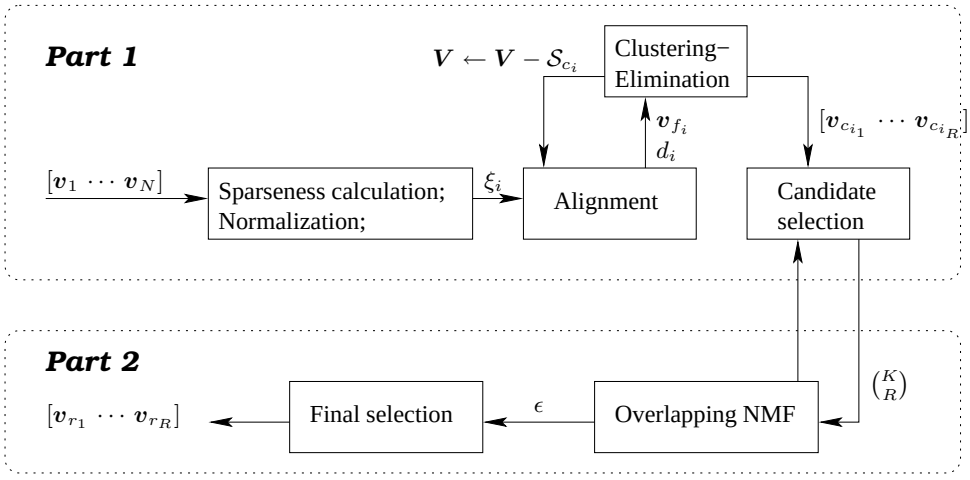

Fig. 1. Schematic diagram of our spectral basis selection method, is shown, where 'Part 1' involves the selection of candidate vectors and 'Part 2' determines a few representative spectral basis vectors from candidate vectors found in 'Part 1'.

The vector $\boldsymbol{v}_{f_i}$ associated with the index $f_i = t^*$, is referred to as a *foundation vector* that has the largest sparseness value among $\{\boldsymbol{v}_t\}$. Then we align each vector $\boldsymbol{v}_j$ in L remaining input vectors (initially $L = N$ but L represents the number of remaining vectors after the clustering-elimination procedure in step 4) with respect to the current foundation vector $\boldsymbol{v}_{f_i}$ such that the Euclidean distance between $\boldsymbol{v}_{f_i}$ and vertically shift-up or -down version of $\boldsymbol{v}_j$, is minimized. In other words, vectors $\boldsymbol{v}_j$ are vertically shifted-up or -down such that their shifted version provides the minimal Euclidean distance from the foundation vector $\boldsymbol{v}_{f_i}$.

4. **Clustering-Elimination:** The goal of the clustering-elimination step is to eliminate vectors belonging to the cluster where the foundation vector is contained, since those vectors are regarded as redundant vectors. To this end, we first apply the k-means clustering method to dichotomize the aligned vectors (including the foundation vector), leading to two groups $\mathcal{S}_{c_i}$ and $\bar{\mathcal{S}}_{c_i}$. The cluster containing the foundation vectors, $\mathcal{S}_{c_i}$, is further grouped into R sub-clusters, producing $\{\boldsymbol{v}_{c_{i_1}}, \ldots, \boldsymbol{v}_{c_{i_R}}\}$ that is a collection of mean vectors of R sub-clusters.
5. **Candidate selection:** Add the mean vector of the cluster $\mathcal{S}_{c_i}$ to the candidate set.
6. **Repeat:** Repeat steps 3-5 with data excluding vectors in $\mathcal{S}_{c_i}$, i.e, $\boldsymbol{V} - \mathcal{S}_{c_i}$, until we choose a pre-specified number of candidate vectors or there is no remaining input vector.

Part 2

This second part involves determining the final representative spectral basis vectors $\{\boldsymbol{v}_{r_1}, \ldots, \boldsymbol{v}_{r_R}\}$ from $K \geq R$ candidate vectors $\{\boldsymbol{v}_{c_1}, \ldots, \boldsymbol{v}_{c_K}\}$ (where K is the integral multiples of R, depending on the number of loops in the clustering-elimination) found in the first part.

1. **Overlapping NMF** Repeat the following step for all possible $\binom{K}{R}$ combination. Construct a small set of input vectors $\widetilde{\boldsymbol{V}}$ by random sampling and treat them at input vectors for the overlapping NMF. Choose R candidate vectors from $\{\boldsymbol{v}_{c_1}, \ldots, \boldsymbol{v}_{c_K}\}$ and fix them (denoted by $\widetilde{\boldsymbol{W}}$) as basis vectors. Run the overlapping NMF with these $\widetilde{\boldsymbol{V}}$ and $\widetilde{\boldsymbol{W}}$ to calculate the reconstruction error.
2. **Final selection** Choose spectral basis vectors that give the lowest reconstruction error.

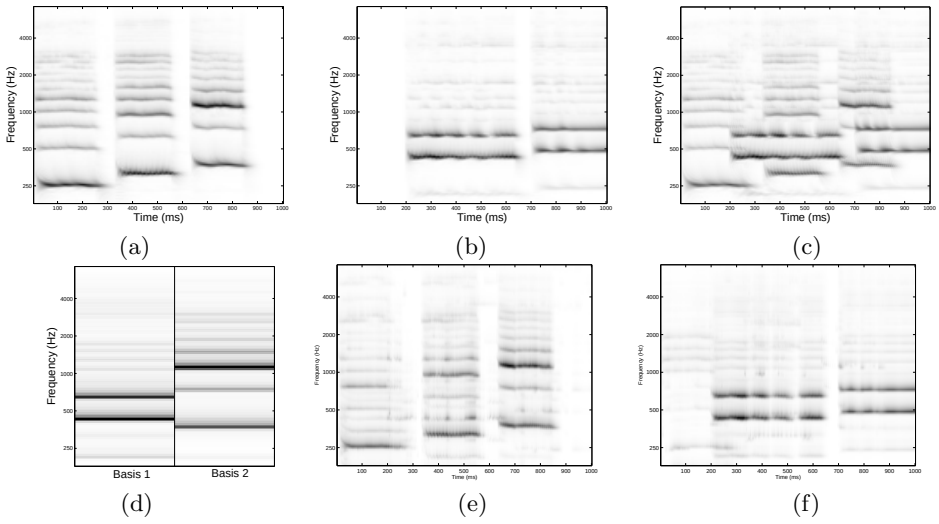

Fig. 2. Auditory spectrograms of original sound of /*ah*/ voice and a single string of a cello are shown in (a) and (b), respectively. Horizontal bars reflect the harmonic structure. One can see that every note is the vertically-shifted version of each other if their musical instrument sources are the same. Monaural mixture of voice and cello is shown in (c) and final two representative spectral basis vectors in (d) which give the smallest reconstruction error in the overlapping NMF are selected by our algorithm in Fig. 1. Each of these two basis vectors is a representative one for voice and a string of cello. Unmixed sound is shown in (e) and (f) for voice and cello, respectively.

4 Numerical Experiments

We present two simulation results for monaural instantaneous mixtures of voice and cello and monaural convolutive mixtures of saxophone and viola. We apply

our spectral basis selection method with the overlapping NMF to these two data sets transformed to auditory spectrograms using the NSL toolbox [12]. Experimental results are shown in Fig. 2 and 3 where figure captions describe detailed results. Note that the mixture in Fig. 3 (c) is a convolutive mixture and we can apply our framework even in that case without any modification if the reverberation time is not too long.

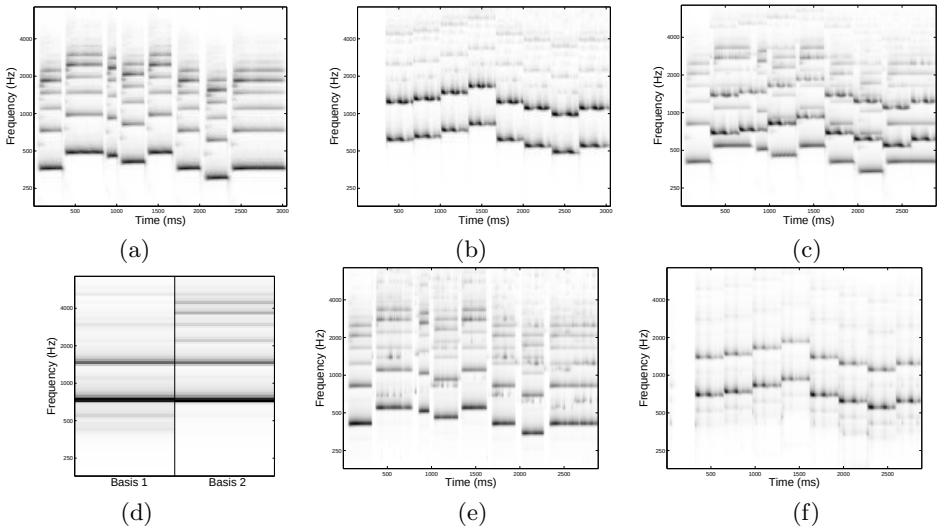

Fig. 3. Auditory spectrograms of original sound of saxophone and viola are shown in (a) and (b), respectively. Every note is artificially generated by changing the frequency of a real sample sound, so that the spectral character of each instrument is constant in all the variations of notes. We mixed these two signals by convolving them with two impulse response signals measured in a studio environment (reverberation time is about 150ms and the frequency response makes a peak at around 27Hz). The monaural convolutive mixture is shown in (c) and finally selected two representative spectral basis vectors are in (d). Unmixed sound is shown in (e) and (f) for saxophone and viola, respectively.

Fig. 4 shows the reusability of our obtained spectral basis vectors. The mixture in Fig. 4 (c) is another part of the same song used in Fig. 3. In this example, we do not have to find out the spectral basis vectors of saxophone and viola again, but can simply reuse the previous results of Fig. 3. Note that if some input data do not satisfy the horizontal sparseness, which means that there is no section occupied by only one instrument, our spectral basis selection method will fail in this case. However we can attack this problem by reusing the previously obtained spectral basis vectors of the same source instruments. Audio demo can be found in `http://home.postech.ac.kr/~minjekim/demo.php`. .

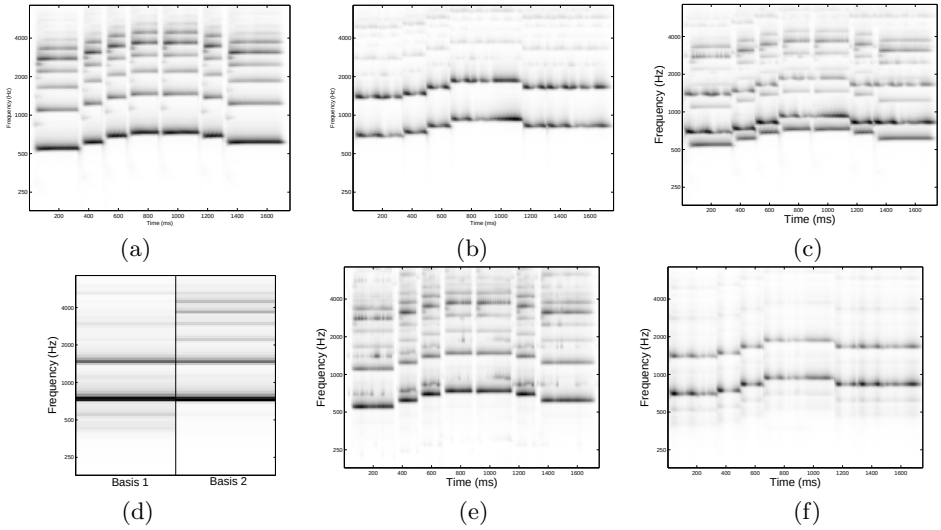

Fig. 4. These figures show the reusability of spectral basis vectors. Auditory spectrograms of original sound of saxophone and viola are shown in (a) and (b), respectively. Every note is generated in the same manner of Fig. 3 but the melody is totally different from it since this is another part of the same song. The mixing process is also the same with the previous experiment. The monaural convolutive mixture is shown in (c). Instead of finding out representative basis vectors, we reused the basis vectors (d) found in previous example. Unmixed sound is shown in (e) and (f) for saxophone and viola, respectively.

The set of transformation matrices, $\mathcal{T}$, that we used, is

$$\mathcal{T} = \left\{ \boldsymbol{T}^{(k)} \;\middle|\; \boldsymbol{T}^{(k)} = \overset{k-m}{\longmapsto} \boldsymbol{I}, \quad 1 \leq k \leq 2m-1 \right\}, \tag{8}$$

where $\boldsymbol{I} \in \mathbb{R}^{m \times m}$ is the identity matrix and $\overset{j}{\longmapsto} \boldsymbol{I}$ leads to the shift-up or shift-down of row vectors of $\boldsymbol{I}$ by j, if j is positive or negative, respectively. After shift-up or -down, empty elements are zero-padded.

For the case where $m = 3$, $\boldsymbol{T}^{(2)}$ and $\boldsymbol{T}^{(5)}$ (they means that $k = 2$ and $k = 5$) are defined as

$$\boldsymbol{T}^{(2)} = \overset{2-3}{\longmapsto} \boldsymbol{I} = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}, \quad \boldsymbol{T}^{(5)} = \overset{5-3}{\longmapsto} \boldsymbol{I} = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}. \tag{9}$$

Multiplying a vector by these transformation matrices, leads to a set of vertically-shifted vectors.

5 Discussions

We have presented a method of spectral basis selection for monaural music source separation, where we incorporated with the harmonics, sparseness, clustering, and the overlapping NMF. Rather than learning spectral basis vectors from the data, our approach is to select a few representative spectral vectors among given data and fix them as basis vectors to learn associated encoding variables through the overlapping NMF, in order to restore unmixed sound. The success of our approach lies in the two assumptions. The one is that the distinguished timbre of a given musical instrument can be expressed by a transform-invariant time-frequency representation, even though their pitches are varying. The other is that there is solo sections in a musical sound where the contribution of each source instrument appears. Our experimental results showed that the proposed methods are reasonable in both instantaneous and convolutive mixture cases.

Acknowledgments: This work was supported by ITEP Brain Neuroinformatics Program and ITRC CMEST.

References

1. Lee, D.D., Seung, H.S.: Learning the parts of objects by non-negative matrix factorization. Nature **401** (1999) 788–791
2. Smaragdis, P.: Non-negative matrix factor deconvolution: Extraction of multiple sound sources from monophonic inputs. In: Proc. Int'l Conf. Independent Component Analysis and Blind Signal Separation, Granada, Spain (2004) 494–499
3. Plumbley, M.D., Abdallah, S.A., Bello, J.P., Davies, M.E., Monti, G., Sandler, M.B.: Automatic transcription and audio source separation. Cybernetics and Systems (2002) 603–627
4. Smaragdis, P., Brown, J.C.: Non-negative matrix factorization for polyphonic music transcription. In: Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New Paltz, NY (2003) 177–180
5. Abdallah, S.A., Plumbley, M.D.: Polyphonic music transcription by non-negative sparse coding of power spectra. In: Proc. Int'l Conf. Music Information Retrieval, Barcelona, Spain (2004) 318–325
6. Helén, M., Virtanin, T.: Separation of drums from polyphonic music using non-negative matrix factorization and support vector machine. In: Proc. European Signal Processing Conference, Antalaya, Turkey (2005)
7. Cho, Y.C., Choi, S.: Nonnegative features of spectro-temporal sounds for classfication. Pattern Recognition Letters **26** (2005) 1327–1336
8. Eggert, J., Wersing, H., Körner, E.: Transformation-invariant representation and NMF. In: Proc. Int'l Joint Conf. Neural Networks. (2004)
9. Kim, M., Choi, S.: On spectral basis selection for single channel polyphonic music separation. In: Proc. Int'l Conf. Artificial Neural Networks. Volume 2., Warsaw, Poland, Springer (2005) 157–162
10. FitzGerald, D., Cranitch, M., Coyle, E.: Generalised prior subspace analysis for polyphonic pitch transcription. In: Proc. Int'l Conf. Digital Audio Effects. (2005)
11. Hoyer, P.O.: Non-negative matrix factorization with sparseness constraints. Journal of Machine Learning Research **5** (2004) 1457–1469
12. Ru, P., Chi, T., Shamma, S.: NSL Toolbox (1997)