SPECTRO-TEMPORAL POST-SMOOTHING IN NMF BASED SINGLE-CHANNEL SOURCE SEPARATION

Emad M. Grais and Hakan Erdogan
{grais, haerdogan}@sabanciuniv.edu

Faculty of Engineering and Natural Sciences,
Sabanci University, Orhanli Tuzla, 34956, Istanbul.

ABSTRACT

In this paper, we propose a new, simple, fast, and effective method to enforce temporal smoothness on nonnegative matrix factorization (NMF) solutions by post-smoothing the NMF decomposition results. In NMF based single-channel source separation, NMF is used to decompose the magnitude spectra of the mixed signal as a weighted linear combination of the trained basis vectors. The decomposition results are used to build spectral masks. To get temporal smoothness of the estimated sources, we deal with the spectral masks as 2-D images, and we pass the masks through a smoothing filter. The smoothing direction of the filter is the time direction of the spectral masks. The smoothed masks are used to find estimates for the source signals. Experimental results show that, using the smoothed masks give better separation results than enforcing temporal smoothness prior using regularized NMF.

***Index Terms*—** Single channel source separation, nonnegative matrix factorization, and speech-music separation.

1. INTRODUCTION

In single channel source separation problems, only one observation of the mixed signal is available. The solution of this problem usually relies on training data for each source signal. The training data are used to train a set of representative vectors for each source signal. The mixed signal is then decomposed with the trained representative vectors. The decomposition results are used to find an estimate for each source. The intuitive assumption of the decomposition results is the temporal smoothness and continuity between the consequent frames. In [1, 2, 3], the continuity and smoothness were enforced within the NMF decomposition by using different regularized NMF cost functions. In [4], the continuity was enforced within the decomposition algorithm with a penalized least squares approach. Enforcing continuity and smoothness within the decomposition algorithm needs to define a cost function for the temporal continuity, which makes the decomposition algorithm slightly more complicated.

In this work, we propose a simple and effective method to enforce temporal smoothness on the estimated source signals. In this work, NMF [5] is used to train a set of basis vectors for each source by decomposing the magnitude spectra of their training data. After observing the mixed signal, a regular NMF is used to decompose its magnitude spectra with the trained basis vectors for all sources. The NMF decomposition results are used to build a spectral mask. The spectral mask explains the contribution of each source signal in the mixed signal. To enforce temporal smoothness on the estimated source signal, we pass the spectral mask through a smoothing filter. The spectral mask is treated as a 2-D image signal. In this work, we investigate three different types of smoothing filters. First filter, is the median filter. The second filter, is the moving average low pass filter. The third, is the Hamming windowed moving average filter, which we note it as Hamming filter for short. Here, we have more freedom to choose any length for the filter, which means we can consider smoothness between more than two consequent frames. We also have different ways of smoothing the spectral mask. The final estimates for the source signal magnitude spectrograms are found by element-wise multiplication of the smoothed spectral mask with the magnitude spectrogram of the mixed signal. That means, the entries of the estimated magnitude spectrogram for each source are the scaled version of their corresponding entries in the mixed signal magnitude spectrogram.

The remainder of this paper is organized as follows: In section 2, a mathematical description of the single channel source separation problem is given. In section 3, we give a brief explanation about NMF and how it is used in source separation. In section 4, we explain our main contribution in this paper, which is the smoothed spectral mask approach. In the remaining sections, we present our observations and results of our experiments.

2. PROBLEM FORMULATION

In single-channel source separation problems, we aim to find estimates of source signals $s_i(t)$ that are mixed when a single mixture is available. This problem is usually solved in the short time Fourier transform (STFT) domain. Let $X(t,f)$ be the STFT of $x(t)$, where t represents the frame index and f is the frequency-index. Due to the linearity of the STFT, we have:

$$X(t,f) = \sum_{i=1}^{N} S_i(t,f), \tag{1}$$

where $S_i(t,f)$ is the unknown STFT of source i in the mixed signal, and N is the number of sources in the mixed signal. In this framework [1, 6], the phase angles of the STFT were usually ignored. Hence, we can approximately write the magnitude spectrum of the measured signal as the sum of source signals' magnitude spectra as follows:

$$|X(t,f)| = \sum_{i=1}^{N} |S_i(t,f)|\,. \tag{2}$$

We can write the magnitude spectrogram in matrix form as follows:

$$\boldsymbol{X} = \sum_{i=1}^{N} \boldsymbol{S}_i, \tag{3}$$

where $\boldsymbol{S} = \{\boldsymbol{S}_1, .., \boldsymbol{S}_i, .., \boldsymbol{S}_N\}$ are the unknown magnitude spectrograms of the source signals, and need to be estimated using the observed mixed signal and the training data. The magnitude spectrogram for the observed signal $x(t)$ is obtained by taking the magnitude of the DFT of the windowed signal.

3. NON-NEGATIVE MATRIX FACTORIZATION

Non-negative matrix factorization (NMF) is a matrix factorization algorithm with nonnegativity constraints. The nonnegative matrix $\boldsymbol{V}$ can be decomposed into a nonnegative basis vectors matrix $\boldsymbol{B}$ and a nonnegative gains matrix $\boldsymbol{G}$ as follows:

$$\boldsymbol{V} \approx \boldsymbol{B}\boldsymbol{G}. \tag{4}$$

In this work, the matrices $\boldsymbol{B}$ and $\boldsymbol{G}$ can be found by solving the following generalized Kullback-Leibler divergence cost function [5]:

$$\min_{\boldsymbol{B},\boldsymbol{G}} D\left(\boldsymbol{V} \,||\, \boldsymbol{B}\boldsymbol{G}\right), \tag{5}$$

where

$$D\left(\boldsymbol{V} \,||\, \boldsymbol{B}\boldsymbol{G}\right) = \sum_{k,l} \left(\boldsymbol{V}_{k,l} \log \frac{\boldsymbol{V}_{k,l}}{(\boldsymbol{B}\boldsymbol{G})_{k,l}} - \boldsymbol{V}_{k,l} + (\boldsymbol{B}\boldsymbol{G})_{k,l} \right),$$

subject to elements of $\boldsymbol{B}, \boldsymbol{G} \geq 0$. The solution for equation (5) can be computed by alternating updates of $\boldsymbol{B}$ and $\boldsymbol{G}$ as follows [5]:

$$\boldsymbol{B} \leftarrow \boldsymbol{B} \otimes \frac{\frac{\boldsymbol{V}}{\boldsymbol{B}\boldsymbol{G}}\boldsymbol{G}^T}{\boldsymbol{1}\boldsymbol{G}^T}, \tag{6}$$

$$\boldsymbol{G} \leftarrow \boldsymbol{G} \otimes \frac{\boldsymbol{B}^T \frac{\boldsymbol{V}}{\boldsymbol{B}\boldsymbol{G}}}{\boldsymbol{B}^T\boldsymbol{1}}, \tag{7}$$

where $\boldsymbol{1}$ is a matrix of ones with the same size of $\boldsymbol{V}$, the operation $\otimes$ is element-wise multiplication, and divisions also are element-wise operations.

3.1. Training the bases

The available training data for each source signal is used with NMF to train a set of basis vectors for each source in magnitude spectral domain as follows:

$$\boldsymbol{S}_i^{\text{train}} \approx \boldsymbol{B}_i \boldsymbol{G}_i^{\text{train}}, \tag{8}$$

where $\boldsymbol{B}_i$ is the matrix that contains the set of basis vectors in its columns, $\boldsymbol{S}_i^{\text{train}}$ is the magnitude spectrogram of the training data for source i. The NMF multiplicative update rules in equations (6, 7) are used to solve for $\boldsymbol{B}_i$ and $\boldsymbol{G}_i^{\text{train}}$. In each iteration, we normalize the columns of $\boldsymbol{B}_i$ and find $\boldsymbol{G}_i^{\text{train}}$ accordingly. All the matrices $\boldsymbol{B}_i$ and $\boldsymbol{G}_i^{\text{train}}$ are initialized by positive random noise. For each source i there is a corresponding trained basis matrix $\boldsymbol{B}_i$ and a gains matrix $\boldsymbol{G}_i^{\text{train}}$. The trained basis matrices are only used in the separation process as we explain in the next sections.

3.2. Decomposing the mixed signal

After observing the mixed signal $x(t)$, NMF is used to decompose the magnitude spectrogram matrix $\boldsymbol{X}$ but with a fixed concatenated bases matrix as follows:

$$\boldsymbol{X} \approx \boldsymbol{B}\boldsymbol{G}, \quad \text{or} \quad \boldsymbol{X} \approx [\boldsymbol{B}_1, .., \boldsymbol{B}_i, .., \boldsymbol{B}_N] \begin{bmatrix} \boldsymbol{G}_1 \\ . \\ . \\ \boldsymbol{G}_i \\ . \\ . \\ \boldsymbol{G}_N \end{bmatrix}, \tag{9}$$

where the matrices $\boldsymbol{B}_1, ..., \boldsymbol{B}_N$ are the N trained basis matrices corresponding to N source signals that are found from solving equation (8). The only unknown in equation (9) is the gains matrix $\boldsymbol{G}$, which is a combination of submatrices as shown in equation (9). The gains matrix can be found using the gain multiplicative update rule in equation (7). $\boldsymbol{G}$ is initialized by positive random noise. The estimate of the magnitude spectrogram of source i is found by multiplying its corresponding basis matrix $\boldsymbol{B}_i$ with its corresponding gains submatrix $\boldsymbol{G}_i$ in the gains matrix $\boldsymbol{G}$ in equation (9) as follows:

$$\tilde{\boldsymbol{S}}_i = \boldsymbol{B}_i \boldsymbol{G}_i. \tag{10}$$

4. SOURCE SIGNALS RECONSTRUCTION AND SMOOTHED MASKS

Instead of finding the source signal estimates using equation (10) as usually used in literature, we have proposed a different method to find the estimates of the source signals [7]. The solution of equation (9) is used to build a spectral mask for source i as follows:

$$\boldsymbol{A}_i = \frac{(\boldsymbol{B}_i \boldsymbol{G}_i)^p}{\sum_{j=1}^{N} (\boldsymbol{B}_j \boldsymbol{G}_j)^p}, \tag{11}$$

where $p > 0$ is a parameter, $(.)^p$, and the division are element wise operations. Notice that, elements of $\boldsymbol{A}_i \in [0, 1]$, and using different p values leads to different kinds of masks. These masks will scale every entry of the mixed signal magnitude spectrogram with a ratio that explains how much each source contributes in the mixed signal as follows:

$$\hat{\boldsymbol{S}}_i = \boldsymbol{A}_i \otimes \boldsymbol{X}, \tag{12}$$

where $\hat{\boldsymbol{S}}_i$ is the final estimate of the magnitude spectrogram of source i, and $\otimes$ is element-wise multiplication. As shown in [7, 8], changing the value of p may improve the performance of the separation results. When $p = 2$, the mask can be considered as a Wiener filter, and when $p = \infty$ we get a binary mask.

Typically, in the literature [1], the continuity and smoothness between the estimated consequent frames are enforced in the solution of the matrix $\boldsymbol{G}$ in equation (9). In this work, we enforce smoothness by applying different smoothing filters to the spectral mask $\boldsymbol{A}_i$. We deal with the mask as a 2-D image, and we apply the smoothing filter in two different ways using three different type of filters for each way. The first way of applying the smoothing filter to the spectral mask is as follows:

$$\boldsymbol{H}_i = \xi\left(\frac{(\boldsymbol{B}_i\boldsymbol{G}_i)^p}{\sum_{j=1}^{N}(\boldsymbol{B}_j\boldsymbol{G}_j)^p}\right), \tag{13}$$

where $\xi\,(.)$ is a smoothing filter. The second way of applying the smoothing filter to the spectral mask is as follows:

$$\boldsymbol{H}_i = \frac{(\boldsymbol{B}_i\xi\,(\boldsymbol{G}_i))^p}{\sum_{j=1}^{N}(\boldsymbol{B}_j\xi\,(\boldsymbol{G}_j))^p}, \tag{14}$$

which means we apply the smoothing filter on the gains matrices only in the spectral mask formula.

The first type of filters that are used in this work is the median filter, which replaces the entry values of the mask by the median of all entries in the neighborhood. The second filter is the moving average low pass filter. The 1-D moving average low pass filter coefficients c_n are defined as $c_n = \frac{1}{b},\ \ n = \{1, 2, ..., b\}$, where b is the filter length. The third filter is the Hamming windowed moving average filter “Hamming filter” for short with 1-D coefficients c_n defined as $c_n = \frac{1}{c}w_n,\ \ n = \{1, 2, ..., b\}$, where c is chosen such that $\sum_n c_n = 1$, and w is the Hamming window with length b. The direction of the smoothing filter is usually in the time axis, which is the horizontal axis of the spectral mask. As we elaborate in the next sections, it is important to note that, both methods of applying the smoothing filters on the spectral mask are neither similar to applying the same smoothing filter to the gains matrix $\boldsymbol{G}$ without mask, nor applying the same smoothing filter to the estimated magnitude spectra of the source signals.

After finding suitable estimates of the magnitude spectrograms of the source signals. The estimated source $\hat{s}_i(t)$ can be found by using inverse STFT to the estimated source magnitude spectrogram $\hat{\boldsymbol{S}}_i$ with the phase angle of the mixed signal.

5. EXPERIMENTS AND DISCUSSION

We applied the proposed algorithm to separate a speech signal from a background piano music signal. Our main goal was to get a clean speech signal from a mixture of speech and piano music. We simulated our algorithm on a collection of speech and piano music data at 16kHz sampling rate. For training speech data, we used 540 short utterances from a single speaker, we used other 20 utterances for testing. For music data, we downloaded piano music from piano society web site [9]. We used 38 pieces from different composers but from a single artist for training and left out one piece for testing. The magnitude spectra of the training speech and music data were calculated by using the STFT: A Hamming window with 480 points length and 60% overlap was used and the FFT was taken at 512 points, the first 257 FFT points only were used since the conjugate of the 255 remaining points are involved in the first FFT points. The test data was formed by adding random portions of the test music file to the 20 speech utterance files at different speech-to-music ratio (SMR) values in dB. The audio power levels of each file were found using the ”audio voltmeter” program from the G.191 ITU-T STL software suite [10]. For each SMR value, we obtained 20 test utterances this way.

We trained 128 basis vectors for each source in equation (8), which makes the size of each trained basis matrix $\boldsymbol{B}_{\text{speech}}$ and $\boldsymbol{B}_{\text{music}}$ to be 257×128, and we fixed the parameter $p = 3$ in equation (11). Those choices gave good results on the same data set in [7].

Performance measurement of the separation algorithms was done using the signal to noise ratio (SNR).

For comparison with our proposed algorithm, we applied the continuity prior algorithm in [1] on our training and testing data set. In [1], the solution of $\boldsymbol{G}$ in equation (9) was found by solving the following regularized Kullback-Leibler divergence cost function:

$$C\,(\boldsymbol{B}_d, \boldsymbol{G}) = C_r\,(\boldsymbol{B}_d, \boldsymbol{G}) + \alpha C_t\,(\boldsymbol{G})\,. \tag{15}$$

Table 1. SNR in dB for the estimated speech signal using only NMF and with using regularized NMF in [1].

SMR dB	Just NMF No Mask No priors	regularized NMF $\alpha_s = 10^{-5}$ $\alpha_m = 10^{-5}$	regularized NMF $\alpha_s = 10^{-5}$ $\alpha_m = 10^{-3}$
-5	**6.17**	6.13	3.53
0	9.15	**9.16**	7.37
5	10.81	**10.81**	10.18
10	12.81	12.81	**14.58**
15	14.02	14.03	**17.60**
20	14.67	14.66	**20.37**

Table 2. SNR in dB for the estimated speech signal using spectral mask without and with smoothing filter, with different filter types and different filter size $a \times b$.

SMR dB	Just Using Mask	Median Filter					Moving Average Filter					Hamming Filter				
		$a=1$	$a=1$	$a=1$	$a=1$	$a=2$	$a=1$	$a=1$	$a=1$	$a=1$	$a=2$	$a=1$	$a=1$	$a=1$	$a=1$	$a=2$
		$b=3$	$b=5$	$b=7$	$b=9$	$b=3$	$b=2$	$b=3$	$b=5$	$b=7$	$b=3$	$b=3$	$b=5$	$b=7$	$b=9$	$b=3$
-5	7.05	7.26	7.44	**7.45**	7.30	7.04	7.18	7.34	**7.38**	7.32	6.84	7.17	7.39	**7.43**	7.42	6.72
0	10.37	10.69	**10.86**	10.82	10.71	10.47	10.56	10.72	**10.74**	10.57	10.13	10.51	10.76	**10.80**	10.75	10.01
5	12.46	12.80	**12.95**	12.92	12.73	12.31	12.60	**12.77**	12.72	12.44	11.87	12.59	**12.81**	**12.81**	12.70	11.78
10	15.23	15.83	**16.03**	15.97	15.78	15.40	15.44	**15.65**	15.53	15.13	14.67	15.40	**15.68**	15.66	15.50	14.54
15	17.05	17.81	**17.98**	17.91	17.72	17.54	17.34	**17.52**	17.32	16.81	16.56	17.24	**17.55**	17.50	17.28	16.43
20	18.40	19.37	19.56	**19.58**	19.41	19.11	18.74	**18.87**	18.63	18.07	17.87	18.60	**18.91**	18.84	18.60	17.75

Table 3. SNR in dB for the estimated speech signal using spectral mask after smoothing the matrix $\boldsymbol{G}$ in the mask, with different filter types and different filter size $a \times b$.

SMR dB	Median Filter			Moving Average Filter					Hamming Filter					
	$a=1$	$a=1$	$a=1$	$a=1$	$a=1$	$a=1$	$a=1$	$a=1$	$a=1$	$a=1$	$a=1$	$a=1$	$a=1$	$a=1$
	$b=3$	$b=5$	$b=7$	$b=3$	$b=5$	$b=7$	$b=9$	$b=11$	$b=3$	$b=5$	$b=7$	$b=9$	$b=11$	$b=13$
-5	7.16	**7.17**	7.15	7.56	7.79	**7.85**	7.82	7.74	7.21	7.60	7.76	7.85	7.88	**7.89**
0	10.46	**10.48**	10.41	10.95	11.16	**11.18**	11.12	10.99	10.56	10.97	11.13	11.20	**11.22**	11.20
5	12.57	**12.69**	12.57	13.12	13.40	**13.48**	13.44	13.31	12.67	13.15	13.35	13.46	13.51	**13.51**
10	15.57	**15.59**	15.54	16.19	16.49	**16.58**	16.56	16.48	15.53	16.20	16.43	16.55	16.60	**16.61**
15	**17.57**	17.54	17.29	18.25	18.60	18.73	**18.75**	18.70	17.44	18.26	18.53	18.68	18.76	**18.79**
20	19.00	**19.06**	18.89	19.85	20.33	20.56	**20.67**	20.59	18.86	19.87	20.24	20.46	**20.68**	20.67

Where $\boldsymbol{B}_d = \left[\boldsymbol{B}_{\text{speech}}, \boldsymbol{B}_{\text{music}}\right]$, C_r is the generalized Kullback-Leibler divergence cost function in (5), α is the regularization parameter, and C_t is the continuity penalty term that was defined as follow:

$$C_t\left(\boldsymbol{G}\right) = \sum_{k=1}^{K} \frac{1}{\sigma_k^2} \sum_{t=2}^{T} \left(g_{k,t} - g_{k,t-1}\right)^2, \tag{16}$$

where k, t are the row and column index of the gains matrix $\boldsymbol{G}$, and $\sigma_k = \sqrt{\left(\frac{1}{T}\right)\sum_{t=1}^{T} g_{k,t}^2}$. In our experiment, we chose different values for the regularization parameter for each source signal. α_s is the regularization parameter for the speech continuity prior and α_m is for the music continuity prior.

Table 1, shows the signal to noise ratio results of the estimated speech signal. We chose the best results according to different values of the parameters α_s and α_m. We also show the separation results using only NMF without any continuity prior or any spectral masks. As we can see from the table, using NMF with continuity prior remarkably improves the separation results at SMR higher than 5dB, but it does not improve the results at low SMR ratio.

Table 2, shows the signal to noise ratio results of the estimated speech signal using spectral mask without and with smoothing filter as in equation (13). In this table, we show the results for different types of filters and different filter size $a \times b$. Where a is the size of the filter in the vertical direction, which is the frequency direction of the spectral mask, and b is the size of the filter in the horizontal direction, which is the time direction of the spectral mask. If $a > 1$ then the filter is smoothing in the frequency direction. If $b > 1$, the filter is smoothing in the time direction, which is equivalent to temporal smoothness. As we can see from the table, using the median filter gives better improvement in the results than using other filters. Also, we can see that, using smoothed spectral mask gives better results than using only the spectral mask. Smoothing the mask in frequency direction as shown in the table for $a > 1$ cases, does not improve the results but it degrades the performance.

Table 3 shows the signal to noise ratio of applying the smoothing filter only on the matrix $\boldsymbol{G}$ in the mask as shown in equation (14). In this table, we got the best SNR results by using the Hamming filter.

It is important to note that, finding the estimates of the sources by smoothing $\boldsymbol{G}$ in the mask formula is different than finding the estimate by smoothing $\boldsymbol{G}$ without mask. Finding the final estimate of the source signal magnitude spectrogram by smoothing $\boldsymbol{G}$ without mask degrades the separation performance as we can see from Table 4. In Table 4, we found the final estimate of the speech magnitude spectrogram as follows:

$$\hat{\boldsymbol{S}}_{\text{speech}} = \boldsymbol{B}_{\text{speech}}\xi(\boldsymbol{G}_{\text{speech}}), \tag{17}$$

where $\boldsymbol{B}_{\text{speech}}$ is the trained basis matrix for the training speech signal, $\boldsymbol{G}_{\text{speech}}$ is the speech gains submatrix in the gains matrix $\boldsymbol{G}$ in equation (9). The smoothed $\boldsymbol{G}$ in (17) is not a minimum of $D\left(\boldsymbol{X}||\boldsymbol{BG}\right)$, and it does not guarantee the sum of the two estimated sources to be equal to the mixed signal. Smoothing $\boldsymbol{G}$ inside the spectral mask in equation

(14) guarantees the sum of the two estimated sources to be equal to the mixed signal. This explains the better results in Table 3 comparing to the results in Table 4.

Table 5 shows the differences between applying the smoothing filter to the spectral mask as in Table 2, and applying the smoothing filter directly to the estimated magnitude spectrogram. In Table 5, we estimated the speech magnitude spectrogram as follows:

$$\hat{\boldsymbol{S}}_{\text{speech}} = \xi\left(\boldsymbol{A}_{\text{speech}} \otimes \boldsymbol{X}\right). \tag{18}$$

This means, we applied the mask on the mixed signal magnitude spectrogram and then we smoothed the result. The effect of the smoothing filter on the widely changing term $\boldsymbol{A}_{\text{speech}} \otimes \boldsymbol{X}$ is different than the effect of the smoothing filter on the mask $\boldsymbol{A}_{\text{speech}} \in [0, 1]$ in equation (11). As we can see from Tables 2 and 5, smoothing the spectral mask in (13) gives better results than the smoothing in equation (18).

In Tables 4 and 5, we showed the results for $b = 3$ only. Since using $b = 3$ did not yield better results than the proposed approaches, we did not continue for larger b.

Table 4. SNR in dB for the estimated speech signal with smoothing $\boldsymbol{G}$ without using mask with different filters with $a = 1, b = 3$.

SMR dB	Median Filter	Moving Average Filter	Hamming Filter
-5	5.29	5.89	6.18
0	7.17	8.52	9.11
5	7.99	9.83	10.70
10	8.97	11.34	12.62
15	9.46	12.18	13.78
20	9.71	12.59	14.38

Table 5. SNR in dB for the estimated speech signal with smoothing the estimated magnitude spectrogram of speech signal with different filters with $a = 1, b = 3$.

SMR dB	Median Filter	Moving Average Filter	Hamming Filter
-5	6.96	7.05	7.18
0	9.86	10.06	10.49
5	11.49	11.69	12.54
10	13.49	13.68	15.25
15	14.59	14.80	17.03
20	15.27	15.48	18.30

Comparing the results of enforcing temporal smoothness in the spectral mask as shown in Tables 2 and 3, with the results of using regularized NMF in Table 1, we can see that using smoothed masks give better results for all SMR values. We got the best results as shown in Table 3 by using Hamming filter to smooth the mask using equation (14). Smoothing the mask using equation (14) is the only method in this paper that guarantees the sum of the estimated source signals to be equal to the observed mixed signal.

Comparing our results in Tables 2 and 3, with the results of using only NMF without using the smoothed masks as shown in the first column in Table 1, we can see that, our proposed method improves the results by **6dB** in some cases.

Audio demonstrations of our experiments are available at: http://students.sabanciuniv.edu/grais/speech/nmfwssm/

6. CONCLUSION

In this work, we studied new methods to enforce smoothness on NMF solution rather than using regularized NMF. The new methods are based on post smoothing the NMF decomposition results. The NMF was used to decompose the magnitude spectra of the mixed signal as a nonnegative weighted linear combination of the trained basis vectors. The decomposition results are used to build spectral masks, then the masks were smoothed. The smoothed masks were used to find an estimate for each source in the mixed signal.

7. ACKNOWLEDGEMENTS

This research is partially supported by Turk-Telekom Group Research and Development, project entitled “Single-Channel Source Separation,” project year 2012.

8. REFERENCES

[1] T. Virtanen, “Monaural sound source separation by non-negative matrix factorization with temporal continuity and sparseness criteria,” *IEEE Trans. Audio, Speech, Lang. Process*, vol. 15, pp. 1066–1074, Mar. 2007.

[2] Nancy Bertin, Roland Badeau, and Emmanual Vincent, “Enforcing harmonicity and smoothness in bayesian nonnegative matrix factorization applied to polyphonic music transcription,” *IEEE Trans. Audio, Speech, Lang. Process*, vol. 18, no. 3, pp. 538–549, 2010.

[3] C. Fevotte, N. Bertin, and J.-L Durrieu, “Nonnegative matrix factorization with the itakura-saito divergence. with application to music analysis,” *Neural Computation*, vol. 21, no. 3, pp. 793–830, 2009.

[4] Hakan Erdogan and Emad M. Grais, “Semi-blind speech-music separation using sparsity and continuity priors,” in *ICPR*, 2010.

[5] D. D. Lee and H. S. Seung, “Algorithms for non-negative matrix factorization,” *Advances in Neural Information Processing Systems*, vol. 13, pp. 556–562, 2001.

[6] Mikkel N. Schmidt and Rasmus K. Olsson, “Single-channel speech separation using sparse non-negative matrix factorization,” in *InterSpeech*, 2006.

[7] Emad M. Grais and Hakan Erdogan, “Single channel speech music separation using nonnegative matrix factorization and spectral masks,” in *17th International Conference on Digital Signal Processing*, 2011.

[8] Emad M. Grais and Hakan Erdogan, “Single channel speech-music separation using matching pursuit and spectral masks,” in *19th IEEE Conference on Signal Processing and Communications Applications*, 2011.

[9] URL, “http://pianosociety.com,” 2009.

[10] URL, “http://www.itu.int/rec/T-REC-G.191/en,” 2009.