Adaptive Time Frequency Resolution for Blind Source Separation

ALEXANDRA CRACIUN[1], *MARTIN SPIERTZ*[1]

[1]Institut für Nachrichtentechnik, RWTH Aachen, Melatener Str. 23, Aachen, Germany

alexandra.craciun@rwth-aachen.de, spiertz@ient.rwth-aachen.de

Abstract. *In this article, we investigate the influence of adaptive time-frequency resolution schemes on a monaural blind source separation algorithm. The goal is to show that the capability of separating the original signals from the mixture is increased if we adapt the time-frequency resolution of the short-time Fourier transform to a certain mixture's characteristics. We will therefore implement different adaptive time-frequency schemes based on the usage of analysis windows of various lengths. The rules that define the mixing schemes rely on two measures: a transient detection measure by means of phase deviations and an energy concentration measure. Both algorithms are evaluated on a large test set, improvements being shown by use of objective quality measures.*

Keywords

Monaural blind source separation, adaptive time-frequency resolution.

1. Introduction

Blind source separation (BSS) is a method used to recover audio signals from a given mixture without having any a priori information available on the sources themselves, therefore the term "blind ". Such a separation scenario is quite similar to the cocktail party effect, which describes the ability of the human auditory system to concentrate on only one speaker in the presence of different interfering sources such as other speakers or background noise [1]. Initially, this issue was examined by Colin Cherry in 1953 and ever since, assiduous research has been carried out in an attempt to better understand the hidden mechanisms of human perception.

Nevertheless, solving the BSS task is more complex for a computer. In the current scenario, we consider separating a monaural mixture of two sources, which could be done by using the spectrogram of the mixture. However, time-frequency representations of signal mixtures such as the spectrogram often suffer from energy smearing effects [4]. This makes it more difficult for a human observer to identify in the spectrogram structures pertaining to separate instruments. If this is unfavorable to the classification of instruments directly by the human eye, we will show that it will also have a negative effect on an automatic BSS scheme which involves a transformation of the mixture to frequency domain. We will therefore develop schemes that adapt the time-frequency resolution of the investigated mixture for improving the separation of the sources.

In Section 2, we will introduce the basic theory behind the BSS scheme step by step. In Section 3, the two adaptive time-frequency schemes and the window mixing algorithms behind them are explained in detail and in Section 4, the experimental results are shown. Last, in Section 5, some conclusions and future work ideas are included.

2. Fundamentals

2.1. Blind Source Separation Algorithm

The used algorithm for blind source separation was first introduced in [2] and a basic scheme of it can be seen in Figure 1. Initially, a monaural instantaneous mixture $x[n]$ is created from two sources s_1 and s_2, of the same loudness level and the same duration. Once the mixture is created, the algorithm transforms it into frequency domain by use of the short-time Fourier transform (STFT). The STFT involves a multiplication of the mixture by an analysis window, followed by a Fourier transform of the resulting product:

$$X[f,n] = \sum_{k=0}^{M-1} x[k]w[k-n]e^{-\frac{j2\pi fk}{M}}. \qquad (1)$$

The next step is the Non-Negative Matrix Factorization (NMF). This is required in order to factorize from the mixture spectrogram the two matrices corresponding to the single note events. Thus $F \times N$ matrix $\mathbf{X}$ will be split into $F \times I$ matrix $\mathbf{B}$ and $I \times N$ matrix $\mathbf{G}$, where dimension I is user-defined. The first matrix contains the frequency basis vectors, while the second matrix contains the envelopes of single acoustic events. One column of $\mathbf{B}$ multiplied by one row of $\mathbf{G}$ corresponds to the spectrogram $\mathbf{C}_i$ of the i-th channel.

Because each instrument plays a whole melody and not a single note, we need to group the notes into melodies. Thus

the NMF is followed by a clustering step that outputs the frequency-domain estimated source signals $\tilde{\mathbf{S}}_1$ and $\tilde{\mathbf{S}}_2$ — for more details, see [2]. The last step, also called synthesis, transforms the signals back into time domain. It is important to include this step in order to obtain an objective measurement of the source separation quality. Due to the fact that the spectrogram resolution changes, it is not possible to perform this measurement in frequency domain. Therefore, the results are compared using the time domain separated sources. For this scope, we calculate the improvement signal-to-noise ratio (ISNR) [3], which is simply the difference between input SNR and output SNR.

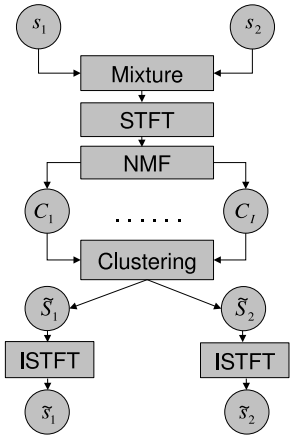

Fig. 1. Scheme of the used blind source separation algorithm

2.2. Energy Smearing in the Time-Frequency Plane

The output of the STFT can be visualized as a 2D-image, also known as *spectrogram*, where the dimensions correspond to time and frequency. The loss of precision in either dimension is characterized by smearing effects in the spectrogram. On one hand, smearing in time causes artifacts such as pre- or post-echoes around transients which can for example lead to an incorrect detection of transient events. On the other hand, smearing of signal energy in frequency domain makes harmonics look thicker, thus preventing the detection of closely spaced ones [4]. In this paper, we will use different types of analysis windows in order to achieve less energy smearing in both time and frequency domain. The adaptive time-frequency schemes used in this purpose are going to be introduced in Chapter 3.

3. Adaptive Time-Frequency Resolution

This chapter begins by describing the main characteristics of the adaptive time-frequency schemes used in this paper: the one with constant hop size and the one with variable hop size. This is followed by a presentation of the window mixing schemes. For these, two ideas have been used. The first idea, introduced in [4], consists on one hand of increasing time resolution in transient regions in order to reduce pre-echoes. On the other hand, frequency resolution is increased in stationary regions, allowing to differentiate between closely spaced harmonics. We will therefore use short analysis windows in transient regions and long analysis windows in stationary regions as our first window mixing scheme. In order to be able to distinguish between the two types of regions, we are going to use a transient detection measure based on phase deviations. The second window mixing scheme is quite different from the first one and it is based on viewing energy concentration as opposite effect to energy smearing. The scheme chooses columnwise different analysis windows, according to which of them maximizes the energy concentration measure [5].

3.1. Adaptive Time-Frequency Scheme with Constant Hop Size

The adaptive scheme with constant hop size is motivated by the paper of Lukin and Todd [4]. It uses long analysis windows of type hann (4096 samples long) and short analysis windows of type $\sqrt{\text{hann}}$ (2048 samples long). The hop size is constant between both long and short analysis windows and has a length of 1024 samples. This results in 75% overlap between the long windows and 50% overlap between the short ones.

Figure 2 shows an example of such a scheme, where the long window in the middle of the left image has been replaced by a short one in the right image. Additionally, we observe that the short window has been zero-padded to the left and right such that it reaches the same length as the long window. This operation is necessary due to the NMF step, which requires a constant time-frequency support.

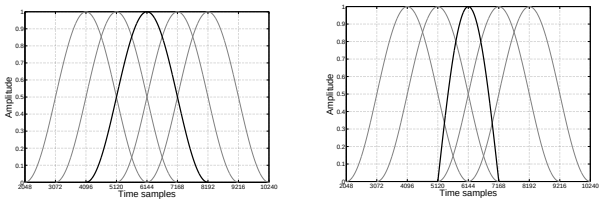

Fig. 2. Substitution of a long analysis window with a short analysis one using constant hop size.

For the signal synthesis, the combination of short and long windows of this scheme is problematic. Synthesis requires that the product of analysis and synthesis windows adds up to a constant value. However, the combinations of short and long hann windows do not add to a constant value, see Figure 3. Therefore, it is necessary to construct a window-sum buffer, which contains the overlap-add of the analysis windows multiplied with the synthesis windows. Afterwards, the reconstructed time domain signals of the es-

timated sources will be divided by this window-sum buffer. In the following we will call this scheme *constant ATF*.

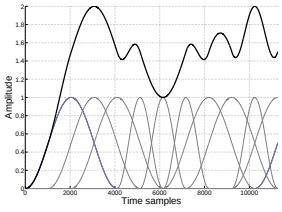

Fig. 3. Combination of long and short analysis windows. The thick line depicts the content of the window-sum buffer.

3.2. Adaptive Time-Frequency Scheme with Variable Hop Size

The adaptive scheme with variable hop size is explained in [7]. While being similar to the one with constant hop size, it also has some important differences. The first difference consists in having more choices for the type of short analysis window. That is, we can choose between short windows of 2048, 1024, 512 or 256 samples long. Additionally, both windows are now of type $\sqrt{\text{hann}}$. The hop size varies according to the analysis window — half of the long window size between long windows and half of the short window size between short windows. This results in a 50% overlap between all analysis windows. Another difference is that two new window types are introduced, which make the transition between long and short windows. Thus, we will have a start window, which is of type long in the first half and type short in the second half and a stop window, which is the opposite of the start window. Again, zero padding is required for the short windows.

Figure 4 illustrates an example, where the long window in the middle of the left image is replaced by three short windows of 2048 samples in the image on the right. We also notice that the two windows adjacent to the long one that was exchanged have been transformed into a start and a stop window. In the following we will call this scheme *variable ATF*.

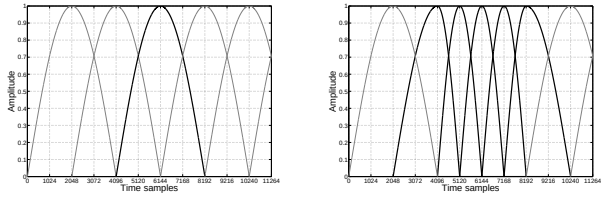

Fig. 4. Substitution of a long analysis window with short analysis ones using variable hop size.

3.3. Window Mixing Scheme Based on Phase Deviations

The transient detection measure based on phase deviations uses the fact that instantaneous frequency is well defined for stationary regions, but not for transient ones [6]. For a local stationary sinusoid, instantaneous frequency should be approximately constant over neighboring windows, which means that

$$\phi_k(n) - \phi_k(n-1) \simeq \phi_k(n-1) - \phi_k(n-2) \qquad (2)$$

where $\phi_k(n)$ represents the 2π-unwrapped phase of the STFT coefficient $X_k(n)$. This allows us to define the phase deviation $\Delta\phi_k(n)$ as the difference of instantaneous frequencies in adjacent windows:

$$\begin{aligned}\Delta\varphi_k(n) = [\varphi_k(n) - \varphi_k(n-1)] - \\ [\varphi_k(n-1) - \varphi_k(n-2)]\end{aligned} \qquad (3)$$

From Equation 2, it results that the phase deviation should be approximately zero in the case of stationary regions. However, the situation is different for transient regions, where the instantaneous frequency is not well defined and as a result $|\Delta\phi_k(n)|$ tends to be large. Therefore, we can use the phase deviation to distinguish between stationary and transient regions in a spectrogram. We derive from it a measure of the distribution of phase deviations across frequency domain that works as a detection function for transients:

$$\zeta(n) = \frac{1}{M}\sum_{k=1}^{M} |\Delta\varphi_k(n)|. \qquad (4)$$

The decision whether a stationary or a transient region is detected will be done in the following manner:

$$\text{If} \begin{cases} \zeta(n) < \text{Th}, & \text{STFT column n is } \textit{stationary} \\ \zeta(n) \geq \text{Th}, & \text{STFT column n is } \textit{transient} \end{cases} \qquad (5)$$

where M = FFT length and Th is a given threshold.

One of the main disadvantages of the transient detection algorithm is the manual setting of the detection threshold Th. Due to the fact that the current algorithm is used for blind source separation, manual intervention must be avoided as much as possible. Though a completely automatic manner of setting Th is not yet possible, we have developed a semi-automatic threshold setting based on the *68-95-99.7 rule* of the pdf [8]. Assuming Gaussian distribution, we rewrite Th in the following manner:

$$\text{Th} = \mu + c_{\text{sel}}\sigma \qquad (6)$$

where μ and σ are the mean, respectively standard deviation of $\zeta(n)$ and c_{sel} is a user-defined selectivity constant. c_{sel} will be varied between low values such as 0.5 and high values such as 2.5.

3.4. Window Mixing Scheme Based on Energy Concentration

The idea behind the energy concentration measure is to avoid as much overlap or interference from neighbouring components as possible and achieve compact time-frequency representations with less smearing effects. Similar to the concept of kurtosis in statistics, the measure acts as an index of how peaky or flat a distribution is. The energy concentration measure is first introduced in [5]. The authors argument their choice on the need of a method that copes with large/fast data variations and that does not require continuous user intervention to match the window size with the data.

The decision on the window length for each column of the STFT is made by maximizing the following measure:

$$\mathrm{Conc}[t,p] = \frac{\sum_{n=1}^{N}\sum_{f=1}^{F}|D_p[f,n]z[n-\tau]|^4}{\left(\sum_{n=1}^{N}\sum_{f=1}^{F}|D_p[f,n]z[n-\tau]|^2\right)^2} \quad (7)$$

where p is the length of analysis window used in the construction of the STFT, $D_p[f,n]$ is equivalent to $\mathrm{STFT}_p[f,n]$ and $z[n]$ is a weighting window centered at $n = 0$, n and f corresponding to the STFT frame index, respectively frequency bin index. Here, $z[n]$ is used as a localization weighting window that has the largest amplitude at its center and decays monotonically to the right and left, thus allowing only neighbouring components to influence the concentration measure. In our tests, we used a Gaussian function of the form $z[n] = ae^{\frac{-(n-\mu)^2}{2\sigma^2}}$, where $\mu = 0$, $a = 1$, $\sigma^2 = 0.1$ and n is a vector of length equal to the number of columns in the STFT matrix.

The decision on the optimal analysis window length p for each column n of the STFT follows from maximizing the concentration measure. Using the notation $Q[n] = \mathrm{Conc}[n,p_2] - \mathrm{Conc}[n,p_1]$, the condition becomes:

$$\text{If}\begin{cases}0 \geq Q[n], & \text{choose window length } p_1\\ \text{else}, & \text{choose window length } p_2\end{cases} \quad (8)$$

Nevertheless, from experimental results, the decision in Equation 8 is too strict and therefore, we derived a more relaxed condition, where the user decides on a selectivity degree:

$$\text{If}\begin{cases}\mu_n - c_{\text{sel}}\sigma_n \geq Q[n], & \text{choose window length } p_1\\ \text{else}, & \text{choose window length } p_2\end{cases} \quad (9)$$

where μ_n and σ_n stand for the mean, respectively the standard deviation of $Q[n]$ and c_{sel} represents the selectivity constant.

4. Results

In this chapter we will present some of the results obtained by using the two adaptive schemes presented in Chapters 3.1 and 3.2. Before showing the results, there are a few important observations that need to be made. First of all, the scheme with variable hop size results in spectrograms of different dimensions, which means that calculating the energy concentration measure directly on these spectrograms is not possible. In addition, the used implementation also works with time-domain detected transients, while the two window mixing measures result in frequency-domain detected transients. Our solution consisted in using the same window mixing scheme as for the adaptive scheme with constant hop size, with the only difference that the detected STFT columns were scaled back into time samples.

Constant ATF scheme	**non-adaptive (long windows)**	**7.368 dB**
	phase deviation c_{sel}=2.2	**7.354 dB**
	energy concentration c_{sel}=2.2	**7.378 dB**
Variable ATF scheme	**non-adaptive (long windows)**	**7.436 dB**
	phase deviation c_{sel}=2.2	**7.618 dB**
	energy concentration c_{sel}=2	**7.559 dB**

Tab. 1. Mean ISNR values over complete test set of 780 mixtures.

Table 1 summarizes the mean ISNR results obtained by the two adaptive schemes with a fixed selectivity constant c_{sel} over the entire test set of 780 mixtures. Our investigations showed that a non-adaptive scheme with long windows performs better than one with short windows, which resulted in choosing the non-adaptive scheme with long windows as reference. Moreover, the best results for a fixed selectivity parameter were obtained for $c_{\text{sel}} = 2.2$. Looking at Table 1, we observe that for the constant ATF scheme, only the window mixing scheme based on the energy concentration measure outperforms the non-adaptive scheme. Nevertheless, for the variable ATF scheme, both phase deviations and energy concentration measures lead to larger mean ISNR values than in the case of the non-adaptive scheme. The 7.618 dB and 7.559 dB values in Table 1 represent the maximum mean ISNR that can be obtained with a variable ATF scheme using different settings for the selectivity constant c_{sel} and the short analysis window length. In the case of phase deviations, the maximum mean ISNR value was obtained for $c_{\text{sel}} = 2.2$ and short analysis windows of 2048 samples and in the case of energy concentration, for $c_{\text{sel}} = 2$ and short analysis windows of 512 samples.

Since the variable ATF scheme showed the most promising results, we will concentrate only on this scheme from now on. We will continue to apply the same short analysis window settings that resulted in the maximum mean ISNR, but we will use optimal threshold settings rather than a fixed selectivity constant c_{sel}. This means that for $c_{\text{sel}} \in \{0.5, 1, 1.5, 2, 2.2, 2.3, 2.5\}$, we create 7 test sets and choose

the maximum mean ISNR for each mixture of the test set for a varying value of c_{sel}.

4.1. Variable ATF Scheme with Optimal Threshold Settings

This section shows some of the results obtained by the ATF scheme with variable hop size and optimal threshold settings. The values that are displayed in Table 2 represent mean values (dB or %) for types of mixtures. As an example, the mixtures of electronic and percussion instruments achieved the maximum dB improvement for the adaptive scheme: +3.8 dB (phase deviation measure) and +3.16 dB (energy concentration measure). Some of the worst results

	Phase deviation measure	Energy concentration measure
Max. improvement over non-adaptive scheme	+3.8 dB	+3.16 dB
Min. improvement over non-adaptive scheme	$\simeq$ +0.5 dB	$\simeq$ +0.4 dB
% of mixtures that prefer the adaptive scheme	$\simeq$ 80-90%	$\simeq$ 90-100%
dB improvement of the adaptive scheme	$\simeq$ [+1,+2] dB	$\simeq$ [+0.5,+1.5] dB

Tab. 2. Variable ATF scheme with optimal threshold settings.

were obtained for mixtures containing noise, which achieved the minimum dB improvement: roughly +0.5 dB, +0.4 dB for phase deviation, respectively energy concentration measure. This suggests that such an adaptive scheme does not perform so well for such mixture types. Overall, however, the results are quite good. For more than 80% of the mixtures types, the variable ATF scheme using the phase deviation measure outperformed the non-adaptive one, reaching even higher percentages for the energy concentration measure (90-100%). Though the adaptive scheme based on energy concentration measure achieved the highest percentages, the dB improvements are slightly lower than when using the phase deviation measure. In general, values between roughly +1 and +2 dB have been obtained for the phase deviation measure, while the energy concentration measure resulted in dB improvements between roughly +0.5 and +1.5 dB.

5. Conclusion

In this paper, we have introduced different adaptive time-frequency resolution schemes and analyzed their capability of improving the BSS of a monaural mixture. The adaptive schemes were based on combining analysis windows of different lengths, by using either a phase deviation or an energy concentration method. Since both measures require manual threshold setting, one of the most challenging tasks was of determining a semi-automatic way to set the threshold. This was done by using a selectivity variable defined by the user. The simulations performed on both constant and variable ATF schemes showed that for a fixed selectivity threshold, the value of $c_{\text{sel}} = 2.2$ achieves the largest improvements. In addition, the variable ATF scheme showed the most promising results. We believe that finding a manner of directly applying the two detection methods would improve even more the source separation.

Acknowledgements

The research described in this paper was supervised by Prof. J. R. Ohm, IENT, RWTH Aachen. The authors would also like to thank the team at IENT for assistance and help.

References

[1] CHEN, Z. *An odyssey of the cocktail party problem*. Technical Report. Adaptive Systems Lab, McMaster University, Hamilton, Ontario, CA, 2003.

[2] SPIERTZ, M., GNANN, V. *Source-filter based clustering for monaural blind source separation*. In *Proc. of the 12th Int. Conference on Digital Audio Effects*, September 2009.

[3] MOLLA, K.I., HIROSE, K. *Single-mixture audio source separation by subspace decomposition of Hilbert spectrum*. In *IEEE Transactions on Audio, Speech and language Processing*, March 2007, vol. 15, no.3, p. 893-900.

[4] LUKIN, A., TODD, J. *Adaptive time-frequency resolution for analysis and processing of audio*. In *120th AES Convention*, May 2006.

[5] VARELA, P., SILVA, A., MANSO, M., ASDEX UPGRADE TEAM. *Adaptive window calculation for automatic spectrogram analysis of broadband reflectometry data*. In *Proc. of the 12th Int. Conference on Digital Audio Effects*, September 2009.

[6] BELLO, J.P., DAUDET, L., ABDALLAH, S., DUXBURY, C., DAVIES, M., SANDLER, M.B. *A tutorial on onset detection in music signals*. In *IEEE Transactions on Signal Processing*, September 2005, vol. 13, no. 5, p. 1035-1047.

[7] SEYMOUR, S. *The modulated lapped transform, its time-varying forms, and its applications to audio coding standards*. In *IEEE Transactions on Speech and Audio Processing*, July 1997, vol. 5, no. 4, p. 359-366.

[8] GERSTMAN, B.B. *Basic biostatistics: statistics for public health practice*. Jones and Bartlett Publishers, Inc., 2008.

About Authors...

A. CRACIUN was born in 1985 and is currently studying a Master program in Communications Engineering at RWTH Aachen. She finished her Bachelor degree in Electrical Engineering and Computer Science at Jacobs University Bremen.

M. SPIERTZ was born in 1980. He received his diploma degree in Electrical Engineering from RWTH Aachen. He is currently working towards his PhD degree at IENT, RWTH Aachen.