Detection and Correction of Phase Errors in Audio Signals for Application in Blind Source Separation

Max BLAESER[1], Julian BECKER[1]

[1]Institute of Communications Engineering, RWTH Aachen University, 52056 Aachen, Germany

blaeser@ient.rwth-aachen.de, becker@ient.rwth-aachen.de

Abstract. *The Nonnegative Matrix Factorization is a powerful and widely used tool in blind audio source separation applications. However, it operates only on the magnitude spectrogram of the Short-Time-Fourier-Transform and disregards the phase information which can lead to audible artifacts in the reconstructed sources after separation. This paper presents a postprocessing method to detect and correct potential phase errors within the Short-Time-Fourier-Transform of a signal, based on autoregressive modeling and linear prediction. Both magnitude and phase of the distorted signal are evaluated and interpolated independently.*

Keywords

BSS, NMF, phase errors, instantaneous frequency

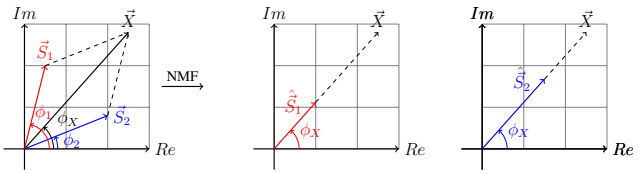
1. Introduction

The problem of Blind Source Separation (BSS) is highly under-determined as there are numerous possibilites of separating a mixture audio signal into its source components. The Short-Time-Fourier-Transform (STFT) provides a useful representation of a signal's frequency contents over time. Musical sources often exhibit regular vertical and horizontal structures in the magnitude spectrogram, which motivates the use of Nonnegative Matrix Factorization (NMF). The NMF iteratively factors the real valued magnitude spectrogram into two sparse matrices that represent activation and frequency base vectors. Recombination of base vectors may yield the magnitude spectra of individual notes played by an instrument or other distinct sound events. By clustering these isolated sounds, instruments, voices and other meaningful contexts can be recovered. The last step in a BSS scenario is performed by the inverse Short Time Fourier Transform (ISTFT) of the separated magnitude spectra. Since the STFT is complex valued, a phase information is needed for transforming the magnitude spectra back into the time domain. The only phase information available however, is the mixture phase of the original signal. This mixture phase is used in many implementations of separation via NMF [1][2] and considered to be a better approximation than phase estimates based on iterative methods that generate a completely new phase from the magnitude such as in *Griffin & Lim's* algorithm or *Real-Time Iterative Spectrogram Inversion* (RTISI) [6][7].
It is valid to use the mixture phase for the ISTFT under certain assumptions:

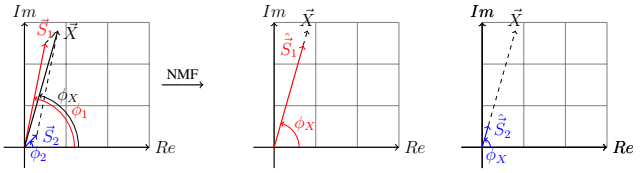
- The sources are separated in both temporal and spectral direction.
- If the sources have temporal and spectral overlap, only one active source contains relevant energy compared to the other active sources.

For typical musical signals, these assumptions are frequently not met. Especially music that is composed of instruments with harmonic and percussive character, identifiable by horizontal and vertical structures respectively, have rich spectra where horizontal and vertical structures overlap. As can be seen in Figure 1a, the mixing process of two sources in the STFT domain is equivalent to the vector addition of two complex numbers (Fourier-coefficients), involving both the magnitude and phase component. Separation via NMF is equal to allocating a portion of the magnitude to the potential sources while retaining the mixture phase (Figure 1b). Audio signals with strong harmonic character (string instruments, aerophones or artifical electronic sounds) show a distinct linear phase behaviour in temporal direction, meaning that the phase increment between STFT bins is nearly constant whenever the instrument is active. Using the mixture phase for ISTFT induces transients in the otherwise smooth and linear phase of harmonic signals. Percussive signals show a similar behaviour of linear phase increment, along the spectral direction but much less pronounced than signals with harmonic character. To eliminate the disturbing glitches present in the separated audio signals and ensure high fidelity, it would be desirable to safely identify phase errors in the STFT and interpolate the distorted Fourier-coefficients.

Experiments using the separated magnitude via NMF and a perfect phase (belonging to the original harmonic source before the mixing process, which in reality is never available) for reconstruction showed an increase of approx. 3 dB–4 dB in separation quality compared to the usual



(a) Mixing of two sources and BSS with comparable energies.



(b) Mixing of two sources and BSS with dissimilar energy content.

Fig. 1. Mixing model and source separation of two signals $\hat{S}_1, \hat{S}_2$ in the complex domain using the mixture phase ϕ_X.

method of using the mixture phase in the ISTFT process. The results prove that correct phase information is of great importance and can significantly improve separation quality. The paper is organized as followed: Section 2 provides an overview of the concepts of phase and instantaneous frequeny and how they can be interpreted in regard to musical signals. In Section 3, the algorithm for the proposed method of phase error detection and correction for harmonic signals is presented. Experimental results are given in Section 4, followed by a conclusion of the investigations in Section 5.

2. Phase and Instantaneous Frequency

The phase of audio signals is often considered to be of lesser importance than the magnitude as it is known that the human hearing apparatus is insensitive to the *relative* phase position of harmonic oscillations. Sudden *changes* in phase, demonstrated in [4], are clearly noticeable for durations as small as 6 ms–8 ms. The STFT is typically computed for a window length of 2^N samples, where N ranges from 9 to 11. Considering that most high fidelity music is sampled at 44100 Hz, a single coefficient of the STFT spans 11.6 ms–34.8 ms. Any phase error present in the STFT domain therefore translates to transients in the time domain. Another reason for the neglectance of phase is its seemingly difficult visual interpretation compared to the easily available magnitude as it is depicted in the spectrogram. A method to circumvent this problem involves the temporal deviation of phase, called *instantaneous frequency* ω_{IF}, which is frequently used in techniques such as the *Phase Vocoder* [3] and *Spectrogram Reassignment* [5].

$$\omega_{\mathrm{IF}}(t) = \frac{\delta}{\delta t}\phi(t) \tag{1}$$

If a simple harmonic oscillation $s(t) = A \cdot \cos(\omega_0 t + \theta)$ with amplitude A, fixed frequeny ω_0 and phase offset θ is considered, its analytical form is given by $s_a(t) = A \cdot \mathrm{e}^{\mathrm{j}\phi(t)}$. It is obvious that the instantaneous frequency of $s_a(t)$ is equal to the fixed frequency ω_0 of the harmonic oscillation. For any mono-component signal $s(t)$ with a varying frequency (such as in frequency modulated signals) the instantaneous frequency is valid but becomes ambiguous if more than one harmonic component is present in the signal at a single time instant. Musical signals rarely contain just a single dominant frequency, which necessitates a different approach to instantaneous frequency. The STFT can be interpreted as applying a filter bank of equally spaced band passes to the signal, depending on the number of frequency bins used for calculating the STFT. Transferring the concept of instantaneous frequency to each frequency bin (or frequency band) of the STFT results in the so called *channelized instantaneous frequency* (CIF) ω_{CIF},

$$\omega_{\mathrm{CIF}}(\omega, t) = \frac{1}{2\pi} \cdot \frac{\delta}{\delta t}\arg(\mathrm{STFT}(s(t))) \tag{2}$$

where $\arg()$ denotes the argument function of the complex valued STFT signal [5]. To avoid periodic discontinuities within the CIF, it is further necessary to *unwrap* the phase by addition of $\pm 2\pi$ prior to derivation. For a sufficiently high number of frequency bands it can now be assumed that each band of the CIF contains only a single dominant frequency component. Figure 3 visualizes the result of phase derivation for a single frequency band in comparison to the usual spectrogram (Figure 2). When the instrument is active from 3.3 s–5.1 s, the high temporal correlation of phase differences for neighboring STFT bins is noticeable. For inactive regions however, the phase differences appear randomly distributed. The pulse-shaped phase errors in the active region, caused by the percussive instrument are also cleary visible. The algorithm specified in the following section is focused on the detection and correction of phase errors for harmonic signals only. An extension to percussive instruments by derivation in the spectral direction can be implemented using similar approaches.

3. Detection and Correction Algorithm

The main approach behind the detection of possible errors is to detect overlapping regions in the STFT domain of the mixture signal that represent potential phase errors. Two approaches are being used to achieve this:

- the STFT magnitudes of already separated signals are used to find overlapping areas by cross comparison,
- the phase of the mixture signal is examined for discontinuities using a discrete form of the CIF and edge detection.

Both results are subsequently combined and a thresholding is performed to generate a binary mask, indicating the position

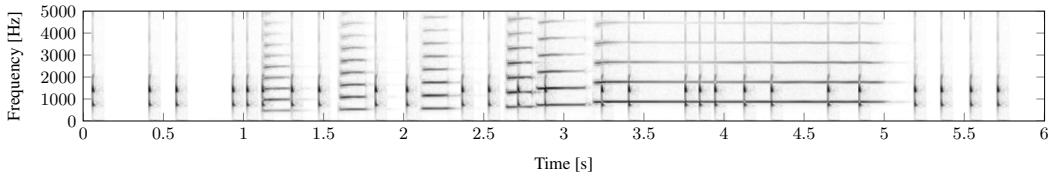


Fig. 2. Spectrogramm of a 5s musical sample containing a trumpet melody and claves.

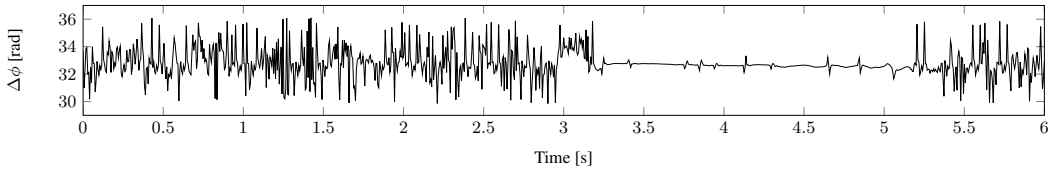


Fig. 3. CIF of the above signal for an STFT band centered at 947 Hz.

of potential phase errors within the spectrum of the harmonic signal. This mask is used in the subsequent correction part of the scheme. As harmonic signals in the STFT domain tend to show a rather smooth local evolution over time (meaning high correlation both in magnitude and phase) it is plausible to use an interpolation method based on autoregressive modeling and linear prediction. However, other interpolation methods have been investigated as well, such as median-filtering and Savitzky-Golay-filtering. The correction step is performed independently on the magnitude and CIF of the distorted Fourier-coefficients.

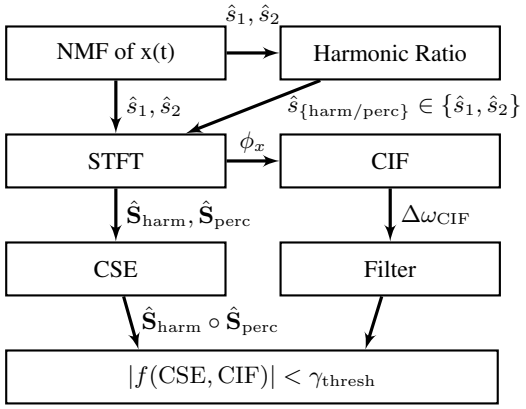


Fig. 4. Block diagram of the steps involed in the detection algorithm.

3.1 Detection of Phase Errors

The detection of phase errors involves multiple steps as depicted in Figure 4. First, it is assumed that the digital musical sample $x[n]$ consists of just two components, a harmonic part s_{harm} and a percussive part s_{perc}. The mixture signal is separated into two estimates $\hat{s}_1, \hat{s}_2$ using NMF and an ideal, non-blind clustering based on methods inspired by [1][2], to avoid any clustering related errors. After monaural BSS it is unknown which of the separated signals has more harmonic or percussive character respectively. For classification, the *Harmonic Ratio* (HR) is used, which is based on a frame-wise computation of autocorrelation values of the respective signal [10]. Large HR values typically indicate the presence of a continuous harmonic component.

An STFT is performed after the two estimated signals have been associated with class labels. The magntiudes of both complex matrices are multiplied using the Hadamard-product, resulting in a *Cross-Spectral-Energy-Density* (CSE) $\hat{\mathbf{S}}_{\mathrm{CSE}} = \hat{\mathbf{S}}_{\mathrm{harm}} \circ \hat{\mathbf{S}}_{\mathrm{perc}}$. Regions in both matrices $\hat{\mathbf{S}}_{\mathrm{harm}}$ and $\hat{\mathbf{S}}_{\mathrm{perc}}$, which are temporally and spectrally overlapping are amplified according to their respective energies, while isolated horizontal or vertical structures are filtered out. This rough estimate can be further refined by detecting discontinuities within the mixture CIF and combining it with the CSE. For computation of CIF, finite differences are used, implemented by subtracting neighboring frames of the STFT phase. Multiple approaches have been investigated to safely detect edges within the phase and a 1D-Laplacian filter $\Delta^2\phi_{\mathrm{CIF}}$ has proven to be a simple and effective solution.
The phase error estimates using CSE and CIF are normalized and combined by a weighting function γ, resulting in the final estimate $\hat{\mathbf{S}}_{\mathrm{CSE,CIF}}$ of the location of potential phase errors:

$$\hat{\mathbf{S}}_{\mathrm{CSE,CIF}} = \hat{\mathbf{S}}_{\mathrm{CSE,CIF}} \circ \gamma\left(\Delta^2\phi_{\mathrm{CIF}}\right) \tag{3}$$

For the following correction step, a hard decision is needed, whether regions found in $\hat{\mathbf{S}}_{\mathrm{CSE,CIF}}$ are considered to be relevant phase errors. A simple classification based on maximizing the between-class variance is used, constrained by an additional absolute threshold, derived from the energy-density floor of the mixture signal's STFT. Potential phase errors with energy close to this lower bound are eliminated.

3.2 Correction of Phase Errors

The correction of phase errors is based on the assumption that harmonic signals in the STFT domain show a smooth local temporal evolution within each frequency band. Of course, this assumption is violated for the *onsets* of an instrument and other sudden changes (loudness, intonation) that are intentional and therefore a source for false positives in the detection. A scheme of onset protection is also implemented, detecting the beginning of a note or tone and removing any phase error candidates in the temporal vicinity [11]. Figure 5 shows an overview of the steps involved in

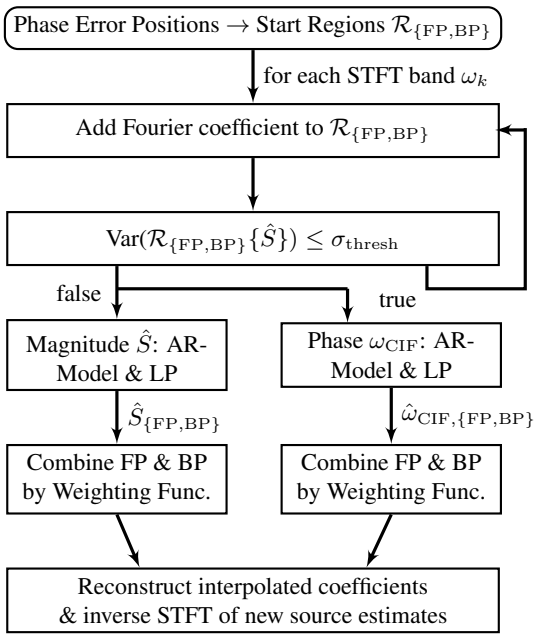


Fig. 5. Block diagram of the steps involed in the correction algorithm.

the phase error correction algorithm. For every phase error candidate, two regions located temporally ahead and subsequent are assigned. These regions are considered to contain undistorted STFT coefficients and are being used to calculate the linear prediction coefficients. As the accuracy of any auto-regressive (AR) model also depends on the number of previous observed values, the collocated regions are iteratively expanded. New coefficients that are located further away from the phase error candidate are included and added to each region, as long as the variance over the STFT magnitudes does not exceed a predefined threshold. This ensures that only the local behavior is used for the autoregressive modeling, as can be seen in Figure 6. After region growing, the model order of the underlying AR-process is estimated by the partial autocorrelation function (PACF) [9] and the prediction coefficients are calculated using Burg's method. If the phase errors span more than one distorted Fourier-coefficient, a multiple-step prediction is performed by feeding the last predicted coefficient back into the prediction loop. Since two predictions from both temporal directions $\hat{X}_{\mathrm{FP}}$ and $\hat{X}_{\mathrm{BP}}$ are available, a weighting function is used to determine the final prediction of magnitude and phase of the distorted STFT coefficients. The weighting function favors those predictions that are temporally closer to the respective region they emanate from.

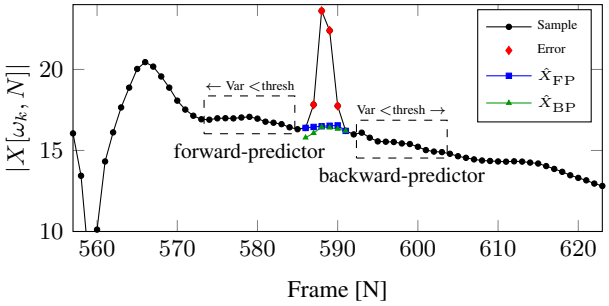


Fig. 6. Examplary interpolation of phase errors for the STFT magnitude by adaptive linear prediction based on autoregressive models of the local neighborhood.

4. Experimental Results

The algorithms were evaluated using audio samples of 5 s–10 s duration from the QUASI test set [12]. The test set contained eleven audio samples that were considered to be of harmonic and seven to be of percussive character. Two audio samples of each class were mixed, resulting in a total of 77 combinations. After blind source separation, the signal-to-distortion-ratio (SDR) [8] was measured based on the original audio sample and the estimation by BSS. Then, the phase error detection and correction was performed. For the entire test set, a gain of 0.138 dB in SDR was observed. However, not every combination of harmonic and percussive signals could be corrected and losses in SDR were also present. Table 1 lists the top five and bottom five signal combinations and the respective gain / loss in separation quality after correction.

Harmonic signal	Percussive signal	Δ SDR [dB]
11. Trumpet	03. Wood Claves	+3.089
02. Harp	04. Snare Drum	+2.162
03. Electronic Pads	01. Hi-Hat	+1.903
08. Acoustic Guitar	05. Kick Drum 1	+1.901
10. Sus Guitar?	03. Wood Claves	+1.850
...	...	...
02. Harp	06. Kick Drum 2	-2.398
03. Electronic Pads	02. Tamburine	-2.502
08. Acoustic Guitar	03. Wood Claves	-2.973
02. Harp	02. Tamburine	-3.422
11. Trumpet	06.Kick Drum 2	-4.210

Tab. 1. Top five and bottom five results of SDR gain / loss for certain instrument combinations.

Especially the phase errors caused by percussive instruments with high frequency content and slow decay, resulting in large overlapping areas in the STFT, such as the tamburine, were deemed difficult to remove from the

spectrum of the harmonic signal. The correction of phase for more than ten distorted Fourier coefficients may also introduce a small relative phase shift in the time domain compared to the original signal. While this relative phase shift is not noticeable to the human listener, the drop in measured SDR can be substantial due to a cumulative error that appears when the SDR is calculated over time. Also, signals that had very few overlapping regions and therefore already high separation quality were more easily affected by this effect.

For harmonic signals with low separation quality in the range of 10 dB–15 dB, which were mixed with short percussive sounds, increases in SDR after correction of around 1.5 dB–3 dB were consistently observed. Glitch and transient artifacts, that closely model the ideal pulse shape of percussive instruments were almost completely removed from the spectrum while retaining the intonation and peculiarity of the harmonic instrument. Also, harmonic signals with few modulation effects, such as vibrato and tremolo, were more easily interpolated by autoregressive modeling.

5. Conclusion

A new postprocessing method for the removal of phase errors in the STFT of harmonic signals after Blind Source Separation was presented. Both magnitude and phase of the Fourier-coefficients are included in the detection method and interpolated independently using autoregressive modeling. The presented method can improve the separation quality for certain combinations of harmonic and percussive signals. Those signals with low separation quality can benefit from the phase error correction, while signals with high separation quality were rarely affected to a noticeable or disturbing degree. More research is needed to adapt the detection and correction parameters based on the signals present in the mixture. Also, the method of linear prediction could be potentially improved by consideration of the global temporal and spectral correlation found in natural signals. Especially musical signals contain repeating patterns and have spectral correlation that could be used to improve the interpolation of distorted STFT coefficients. Lastly, the potential gain of using a combined approach of phase correction before iterative phase estimation according to *Griffin & Lim* should be investigated.

References

[1] SPIERTZ, M. *Underdetermined Blind Source Separation for Audio Signals.* 2012, vol. 10 of Aachen Series on Multimedia and Communications Engineering. Aachen: Shaker Verlag.

[2] VIRTANEN, T. Monaural Sound Source Separation by Nonnegative matrix Factorization With Temporal Continuity and Sparseness Criteria. *IEEE Transactions on Audio, Speech, and Language Processing*, 2000, vol. 15, no. 3, pp. 1066-1074.

[3] LAROCHE, J., DOLSON, M. New phase-vocoder techniques for pitch-shifting, harmonizing and other exotic effects, *Applications of Signal Processing to Audio and Acoustics, IEEE*, 1999, pp. 91–94.

[4] MOORE, B. *Hearing (Handbook of Perception and Cognition)*, Academic Press, Second Edition, 1995.

[5] FULOP, S.,FITZ, K. Algorithms for computing the time-corrected instantaneous frequency (reassigned) spectrogram, with applications. *The Journal of the Acoustical Society of America*, 2006, vol. 119, no. 1, pp. 360–371.

[6] GRIFFIN, D., LIM, J. Signal estimation from modified short-time fourier transform. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 1984, vol. 32, no. 2, pp. 236–243.

[7] GNANN, V., SPIERTZ, M. Inversion of short-time fourier transform magnitude spectrograms with adaptive window lengths. *ICASSP*, 2009, pp. 325–328.

[8] VINCENT, E., GRIBONVAL, R.,FEVOTTE, C. Performance measurement in blind audio source separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 2006, 14(4):1462-1469.

[9] HAMILTON, J. *Time-series analysis*, Princeton Univerity Press, 1994, 1 ed., Jan.

[10] KIM, H., MOREAU, N., SIKORA, T. *MPEG-7 Audio and Beyond. Audio Content Indexing and Retrieval*, 2005, Wiley.

[11] BELLO, J., SANDLER, M. Phase-based note onset detection for music signals. *Applications of Signal Processing to Audio and Acoustics, 2003 IEEE Workshop on*, 2003, vol.5, pp. 441–444.

[12] *Quasi database – A musical audio signal database for source separation*, http://www.tsi.telecom-paristech.fr/aao/en/2012/03/12/quasi/

[13] GUNAWAN, D., SEN, D. Iterative phase estimation for the synthesis of separated sources from single-channel mixtures. *Signal Processing Letters, IEEE*, 2010, vol. 17, no. 5, pp. 421–424.

About Authors ...

Max BLAESER was born in Oberhausen, Germany in 1987. He received the Dipl.-Ing. degree in Electrical Engineering, Information Technology and Communication Engineering from RWTH Aachen University, Aachen, Germany, in 2013 and is currently working as a Ph.D. student at the Institute of Communications Engineering, RWTH Aachen University. His focus is on video coding.

Julian BECKER was born in Bad Kreuznach, Germany in 1985. He received the Dipl.-Ing. degree in Electrical Engineering, Information Technology and Communication Engineering from RWTH Aachen University, Aachen, Germany, in 2010 and is currently working as a Ph.D. student at the Institute of Communications Engineering, RWTH Aachen University. His focus is on audio signal processing.