On-line NMF-based Stereo Up-Mixing of Speech Improves Perceived Reduction of Non-Stationary Noise

Christian Kirst[1], Felix Weninger[1], Cyril Joder[1], Peter Grosche[2], Jürgen Geiger[1], Björn Schuller[1]

[1]*Machine Intelligence & Signal Processing Group, Technische Universität München, Germany*

[2]*HUAWEI TECHNOLOGIES Duesseldorf GmbH, European Research Center, Germany*

Correspondence should be addressed to Christian Kirst (`christian.kirst@mytum.de`)

ABSTRACT

Speech de-noising algorithms often suffer from introduction of artifacts, either by removal of parts of the speech signal, or imperfect noise reduction causing the remaining noise to sound unnatural and disturbing. This contribution proposes to spatially distribute monaural noisy speech signals based on single-channel source separation, in order to improve the perceived speech quality. Stereo up-mixing is utilized on the estimated speech and noise sources instead of simply suppressing the noise. This paper investigates the case of non-negative matrix factorization (NMF) speech enhancement applied to high levels of non-stationary noise. NMF-based and spectral subtraction speech enhancement algorithms are evaluated in a listening test in terms of speech intelligibility, presence of interfering noises and overall quality with respect to the unprocessed signal. In the result, the listening test provides evidence for superior noise reduction by NMF, yet also a drop in perceived speech quality that is not covered by the employed set of common objective metrics. However, stereo up-mixing of NMF-separated speech and noise delivers high subjective noise reduction while preserving the perceived speech quality.

1. INTRODUCTION

The present paper deals with NMF-based single-channel speech enhancement. The algorithm is supposed to separate speech from a noisy recording, which remains a challenging problem, especially in the presence of highly non-stationary noise. Some application fields of this task can be found in telephonic communications [1], hearing aids [2], automatic speech recognition [3], speaker recognition [4] or emotion recognition [5]. NMF can be applied in various fields like image processing [6], music transcription [7], etc. In the case of audio applications one factor contains spectral features whereas the other factor contains temporal activations of these features. This property can be exploited to separate different sources, given spectral models of each source.

Many 'classic' single-channel speech enhancement algorithms such as [8] are based on an estimation of a static noise model. However, a static noise model is not capable of adapting to non-stationary noises such as transients, which occur in many applications. In contrast, the NMF approach taken in this paper is based on a fixed speech model and adaptive noise models. Because conventional noise estimators are very well suited for capturing the stationary background noise, we have proposed hybrid NMF models in [9] that combine static with adaptive noise models.

A crucial problem of speech enhancement is that it can lead to information loss in the speech signal, as is reflected by some objective metrics like Sources to Artifacts Ratio (SAR) [10]. In the present study, when considering subjective measures, this loss of speech quality becomes even more significant. Although noise is suppressed, the filtered speech might be perceived as less intelligible than the unfiltered speech – in other words, the noise suppression is achieved at the cost of producing artifacts.

To remedy this issue this paper proposes an approach which avoids the artifacts introduced by the attempt to suppress the noise. Instead, the present approach performs a spatial redistribution of noise and speech. In natural listening scenarios, the human auditory system can exploit spatial cues in order to separate speech and noise. However, when faced with single channel (mono) signals containing a mix of speech and noise, such spatial cues are not available.

In this paper, we investigate an up-mixing approach with the goal to obtain a spatial separation of speech and noise. Up-mixing by source separation has already successfully been used for musical applications [11, 12, 13], but to our knowledge, is has not been exploited for speech enhancement. Fitzgerald [11] applied up-mixing to prior separated instruments and vocals. He used different methods for the separation of drums, pitched instruments and vocals. The separation of vocals also included NMF. Each separated source might contain artifacts, but the sum of all sources is equal to the original artifact-free mono file. According to Fitzgerald, the artifacts are also not noticeable for a stereo upmixed output if the pan positions for each source are not too extreme.

In its simplest form, upmixing of a mono signal leads to a stereo signal with spatial (left/right) information. A more sophisticated spatial simulation of sound sources can be found in the work of Gerzon [14]. In our approach we use straightforward amplitude panning; nevertheless, the subjective measures show a massive improvement of perceptual speech quality. Let us now outline the employed de-noising algorithms before turning to the evaluation and discussion.

2. METHODOLOGY

2.1. Non-Negative Matrix Factorization

NMF calculates two non-negative factors $W \in \mathbb{R}_+^{m\times r}$ and $H \in \mathbb{R}_+^{r\times n}$ such that their product $\Lambda = WH$ approximates a given matrix $V \in \mathbb{R}_+^{m\times n}$. The method is based on minimizing an error function $D(V,WH)$.

In the case of speech separation the columns of W represent characteristic spectra of speech, such as phonemes, or spectra of noise sounds. The rows of H represent corresponding activations of the spectra contained in W.

In general NMF algorithms use a modified version of gradient descent; in order to maintain non-negativity the gradients are split into positive and negative parts and multiplicative updates are used.

The error function used in this paper is the generalized Kullback-Leibler divergence. It has shown already good results in previous work [15]:

$$D_{\mathrm{KL}}(X,Y) = \sum_{i,j} x_{i,j} \log \frac{x_{i,j}}{y_{i,j}} - x_{i,j} + y_{i,j} \qquad (1)$$

According to [6] and [16] update rules are then given by:

$$W \leftarrow \frac{\frac{V}{WH} H^T}{1 W^T} \qquad (2)$$

$$H \leftarrow \frac{W^T \frac{V}{WH}}{W^T 1} \qquad (3)$$

1 is an all-one matrix. These rules are applied for K iterations.

2.2. Semi-supervised NMF for speech separation

In order to separate the speech from a noisy recording we use semi-supervised NMF as in [1, 9]. The dictionary matrix W is represented as concatenation of speech components $W^{(s)}$ and noise components $W^{(n)}$. In our work, the speech dictionary is learned from clean speech data by performing NMF and keeping the first factor (spectra). During the separation the speech components are fixed and only the activations are calculated.

The NMF signal model Λ is equal to the sum of the speech approximation $\Lambda^{(s)} = W^{(s)}H^{(s)}$ and the noise approximation $\Lambda^{(n)} = W^{(n)}H^{(n)}$:

$$\Lambda = \Lambda^{(s)} + \Lambda^{(n)} \qquad (4)$$

The semi-supervised NMF used in this paper updates the noise components during the separation using multiplicative rules, thus fitting the noise model to the observations of noisy speech. A Wiener filter for the original spectrogram V is constructed from the speech and noise estimates $\Lambda^{(s)}$ and $\Lambda^{(n)}$.

For real-time applications like speech enhancement in mobile phones NMF can be applied *online*. The online NMF used for this paper considers a few time frames from the past as temporal context in a ‘sliding window’ approach. Noise models are adapted over time to fit the incoming observations of noisy speech. This approach is presented in [1]. Further improvements of the NMF algorithm can be achieved by enforcing sparsity during the learning phase [17].

2.3. Hybrid semi-supervised NMF

The hybrid approach combines conventional noise estimators with NMF. It was proposed in previous work [9]. A noise estimate B is computed by an unsupervised method such as minimum statistics [8]. It can be incorporated such that the noise estimate is given by:

$$\Lambda^{(n)} = W^{(n)}H^{(n)} + (1h^{(b)}) \otimes B \tag{5}$$

where 1 is an all-one column vector of dimension m and $h^{(b)} \in \mathbb{R}_{+}^{1\times n}$ is a row vector. $\otimes$ is the Hadamard product. Intuitively the first part of the sum models non-stationary parts of the noise whereas the second summand models stationary (background) noises. Previous work has shown that constant activations $h^{(b)}$ can be reasonably assumed [9]. For this paper the components of $h^{(b)}$ are set a-priori to 1, corresponding to 'full confidence' into the static noise model B.

3. EXPERIMENTS

3.1. Experimental Setup

The methods are evaluated on mixtures of speech and noise from publicly available corpora. Spontaneous speech from the Buckeye corpus [18] is used to reflect use cases such as speech enhancement in wideband telephone channels or multimedia retrieval in web videos. In order to simulate realistic noise, we consider the CHiME 2011 Challenge [19] background noise corpus. This corpus contains sounds recorded in a domestic environment with stationary and non-stationary noises.

Because the focus of this paper lies on a systematic and comparative subjective evaluation of different enhancement algorithms, only six (randomly chosen) recordings of lengths between 10 and 20 seconds are considered. These were mixed with noise at SNRs between -9 dB and 12 dB. More extensive objective evaluations were already done in previous work [17], [9].

NMF is applied to magnitude spectrograms computed using Hamming windows of 32 ms length with 50% overlap. Speaker-dependent speech dictionaries are learned from a 1-minute set of (clean) utterances that is disjoint from the set of test utterances. A generic speech dictionary is trained using a subset of the 40 speakers included in the Buckeye corpus. Each speech dictionary consists of 25 components.

The evaluation compares the original noisy speech with speech separated by different methods:

- spectral subtraction based on minimum statistics [8]
- online semi-supervised hybrid NMF (generic dictionary matrix, $K = 1$ iteration per frame)
- online semi-supervised hybrid NMF (speaker-dependent dictionary matrix, $K = 2$ iterations per frame)

The algorithms uses 2 noise components and a sliding window length of 336 ms, which is an appropriate setting according to previous evaluations of hybrid NMF [9]; it has been shown that hybrid NMF needs less iterations than conventional semi-supervised NMF to achieve good separation results. In informal listening tests, the NMF version using only 1 iteration and a generic speaker model turned out to be a 'softer' version that produces less artifacts, probably due to less overfitting to the observations.

3.2. Objective Evaluation

The objective evaluation compares different measures. The standard source separation metrics are *Sources to Distortion Ratio (SDR)*, *Sources to Interferences Ratio (SIR)* and *Sources to Artifacts Ratio (SAR)* [10]. Since these metrics are based on simple energy ratios and do not take into account how the speech distortions are perceived by the human hearing, [20] introduced new metrics for the perceptual quality assessment of audio source separation. They performed listening tests where the listeners were instructed to rate the enhanced speech signal on a scale from 1 to 5 (1=bad, 2=poor, 3=fair, 4=good, 5=excellent) according to:

- the speech signal distortion (Csig)
- the intrusiveness of the background noise (Cbak)
- the overall quality (Covl)

Based on their listening test they built a regression model to predict these scores based on several objective measures.

Another predicted mean opinion score (MOS) is the *Perceptual Evaluation of Speech Quality (PESQ)* score, which is the most widespread measure for perceptual speech quality assessment. However, PESQ has been designed for the evaluation of speech codecs and is not directly aimed at other applications (in particular, speech denoising).

The individual scores for each sample are shown in figure 2 where as figures 3, 4 display the average values of these objective measures. Each corner of the net plot corresponds to a sample.

The plots show that NMF achieves the best ratings for nearly all measures. Figure 2 confirms this result for the individual samples. Only the SAR is rather low. Especially figure 4 shows that NMF achieves a better separation of speech than spectral subtraction, but at the cost of producing more artifacts. Although this tendency is not depicted by the predicted MOS values, it will be confirmed by the subjective results of the next section.

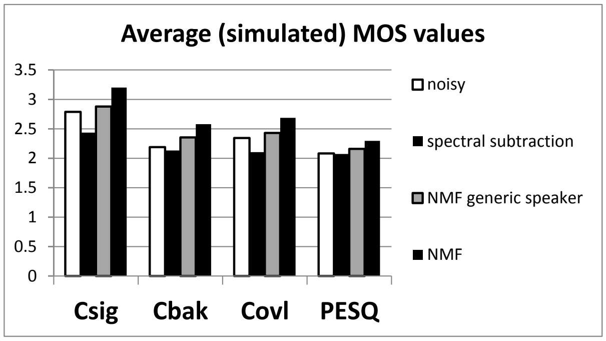

Fig. 3: Average mean opinion score predictors - speaker-dependent NMF achieves the highest scores in all categories. Notably, NMF outperforms the traditional spectral subtraction for these measures.

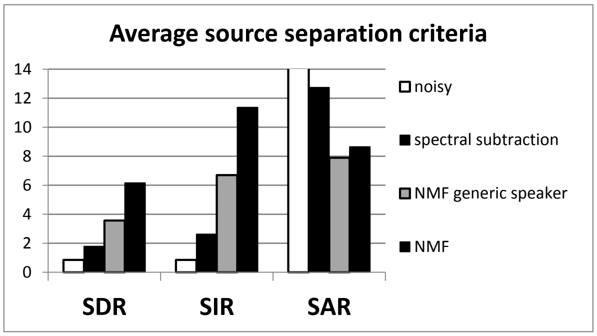

Fig. 4: Average source separation criteria - the plot shows that speaker-dependent NMF achieves the best separation, but it may not provide the best speech quality (in terms of SAR).

3.3. Subjective Evaluation

For a user-centered assessment of the quality of a speech separation algorithm, subjective tests were conducted. Three different properties of the obtained audio signals were evaluated: the sound quality, the intelligibility of speech and the amount of interfering sounds. Three corresponding subjective hearing scores were then defined accordingly.

In the conducted listening tests, the subjects listened to each sound individually, in a blind fashion (without knowing which sound they heard) and in a random order. To assess the consistency of the ratings, half of the sounds occurred twice in the test (subjects were not informed of this fact). The listening was performed through headphones and took place in a quiet room. The subjects were then asked the following questions:

- Which is the grade of speech quality in this sample? (similar to Csig)
- Which is the grade of speech intelligibility in this sample?
- Which is the grade of interfering sounds in this sample? Interfering sounds are meant to be all the sounds not belonging to the main speaker. (similar to Cbak)

The subjects could listen to the sounds as many times as they wanted, before and between these questions. The grades were given as integers in a scale from -3 to +3, -3 meaning the worst and +3 the best subjectiv quality. This gives rise to the *Subjective Speech Quality score (SSQ)*, the *Subjective Speech Intelligibility score (SSI)* and the *Subjective Separation of Noise score (SSN)*.

15 subjects participated in the evaluation (12 male, 3 female, mean age = 29.7, standard deviation = 5.5). 13 of them have at least a master's degree in electrical engineering or related fields and have background in audio signal processing. All participants are non-native, yet very experienced English speakers. Participants were not paid for their service. None of them reported hearing loss.

In order to reduce the burden on the subjects all samples were cut to 10 seconds (starting from the beginning). Each sample appears in different versions: the speech enhanced versions that were already evaluated in the last section plus a stereo version of online hybrid NMF. This version is produced using the sox tool [1]. A spatial situation is simulated where the speaker stands in front of the listener and the background noise comes from somewhere on the left; this is achieved by straightfoward amplitude panning,

[1] http://sox.sourceforge.net/

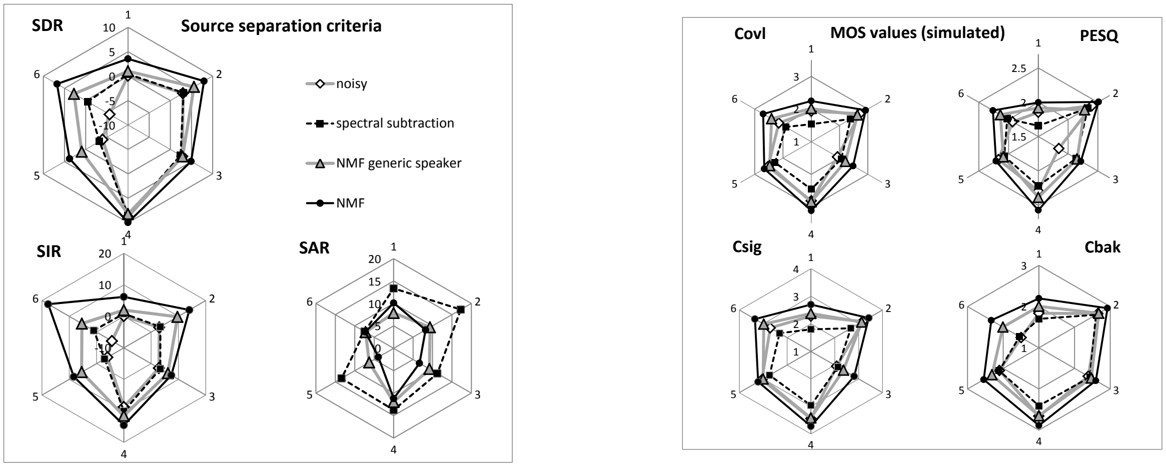

Fig. 2: Netplots of the objective measures for each individual sample. The plots show that the ratings are pretty consistent over the depicted samples. The speaker-dependent NMF enhancement (as opposed to NMF with a generic speaker model) achieves the best scores for all measures but SAR, which means that it produces good speech-noise separations, but the resulting speech signal contains serious artifacts.

i.e., mixing 90 percent of the noise on the left output, 10 percent of the noise on the right output, 50 percent of the speech on the left output and 50 percent of the speech on the right output.

Since half of the stimuli were randomly selected to appear twice, we can measure the reliability of the score values by the *intra-rater agreement*. It is defined as the mean rank (Spearman's) correlation coefficient of the first and second rating of the control stimuli that were played twice. Another measure of interest is the *inter-rater agreement*. It shows the consistency between the ratings by different participants. It is calculated as the rank correlation of individual ratings with the mean of all other ratings. Note that in this case, ratings are averaged for the stimuli played twice. The agreement values are shown in table 1. The low mean and high standard deviation values for SSN hint that the grade of interfering sounds seems hard to measure and the participants seem to have different opinions about it. Noises can be perceived different depending on the listener. Looking at the inter-rater agreement values for SSQ and SSI there is a reasonable consistency between the ratings of different participants.

In order to compare the rating results for different separation algorithms mean ratings are calculated. These values are shown in figure 5. To avoid the problem that some raters might give all stimuli a higher or lower rating than others, also normalized ratings are considered (figures 6 and 7). The normalization is done by calculating the difference between a rating and the mean rating for

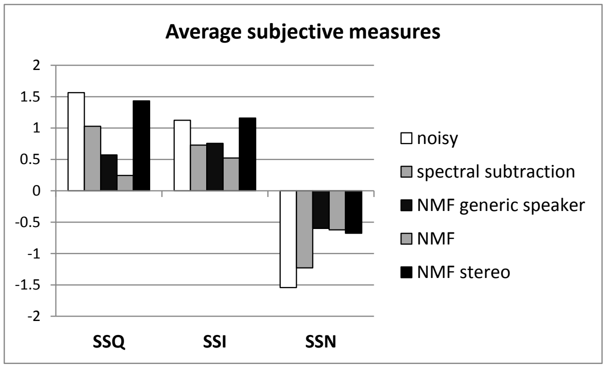

Fig. 5: Average subjective measures. The original noisy speech is compared to separated speech and stereo mixed noise and speech. The stereo stimuli achieved nearly the same perceptual speech quality. Although nothing is left out for the stereo stimuli, they achieved a much higher noise separation score (SSN).

the participant and scaling it with the reciprocal standard deviation. However, comparing both plots there are no qualitative differences.

Considering the speech quality (SSQ) and speech intelligibility (SSI), the stereo method significantly outperforms the mono separated speech, according to a two-tailed paired t-test ($p \ll 0.001$). An interesting result is that the NMF using a generic speaker model and only 1 iter-

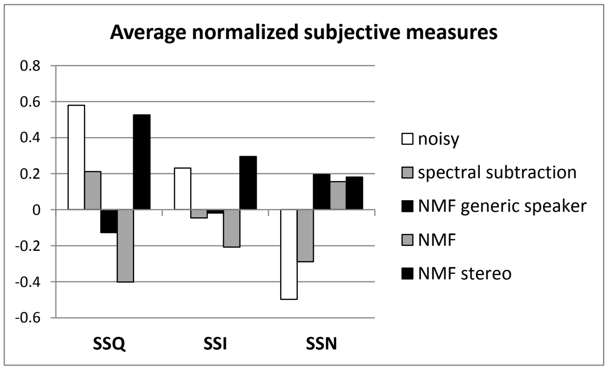

Fig. 6: Average subjective measures normalized (for each participant) by subtracting the mean and dividing by the standard deviation.

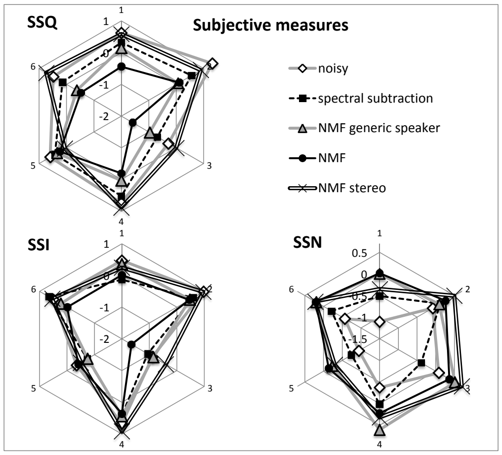

Fig. 7: Average normalized subjective measures for each sample - each corner corresponds to an individual sample.

ation provides better subjective ratings than the speaker dependent version. This result is not at all covered by the objective measures. In fact, in contrast to the objective evaluation the on-line NMF algorithms based on speaker dependent models do not provide the best ratings in any category; this may be due to overfitting to the observations.

Looking at the SSN ratings, the stereo stimuli containing speech and noise seem to contain the same grade of interfering noise as the mono stimuli containing the estimated speech. The differences in noise separation scores (SSN) between NMF online hybrid mono/stereo and the original noisy speech are also significant according to a t-test ($p \ll 0.001$). This is a very interesting result, since in terms of energy, the stereo stimuli contain exactly as much noise as the original noisy speech. This can be attributed to the fact that the spatial distribution of speech and noise facilitates human source separation.

Table 1: The intra-rater and inter-rater agreements (rank correlation coefficients) as a measure of rating consistency.

	SSQ	SSI	SSN
mean intra-rater agreement ± std. dev.	0.51 ±0.13	0.50 ±0.18	0.40 ±0.30
mean inter-rater agreement ± std. dev.	0.60 ±0.12	0.61 ±0.12	0.34 ±0.25

4. CONCLUSIONS

The evaluations show that NMF speech separation suppresses noise at the cost of producing artifacts. These artifacts are shown by the objective measure SAR, but they seem not as significant as the loss of speech quality in the subjective evaluation. Overall the comparison between objective measures and a subjective evaluation reveals that some tendencies of the subjective evaluation are not covered by any objective metric. We believe that this is due to 'unnaturally' sounding speech and noise increasing the cognitive load of the listeners because the sounds cannot be explained by human experience. Such 'semantic' issues are not captured by simple energy or other acoustic feature based objective measures. For this reason, the subjective intelligibility and speech quality of the original noisy signal is often rated better than the enhanced signal, in contrast to the objective measures. The stereo-up-mixing restores the original subjective speech quality and intelligibility while significantly reducing the perceived grade of interfering sounds – which, again, cannot be explained by any objective measure we know of, since the sum of the noise components remains unchanged by enhancement. Future work will hence concentrate on noise source clustering and advanced auditory scene rendering and on improved objective measures to evaluate the rendering quality for the up-mixing case.

5. REFERENCES

[1] C. Joder, F. Weninger, F. Eyben, D. Virette, and B. Schuller, "Real-time speech separation by semi-supervised nonnegative matrix factorization," in

Proc. LVA ICA, Special Session "Real-world constraints and opportunities in audio source separation". Tel Aviv, Israel: Springer, Mar. 2012, pp. 322–329.

[2] N. Madhu, A. Spriet, S. Jansen, R. Koning, and J. Wouters, "The potential for speech intelligibility improvement using the ideal binary mask and the ideal wiener filter in single channel noise reduction systems: Application to auditory prostheses," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 21, no. 1, pp. 63–72, Jan. 2013.

[3] F. Weninger, J. Geiger, M. Wöllmer, B. Schuller, and G. Rigoll, "The Munich 2011 CHiME Challenge contribution: NMF-BLSTM speech enhancement and recognition for reverberated multisource environments," in *Proc. Intern. Workshop on Machine Listening in Multisource Environments (CHiME)*, Florence, Italy, Sep. 2011, pp. 24–29.

[4] J. Ming, T. J. Hazen, J. R. Glass, and D. A. Reynolds, "Robust speaker recognition in noisy conditions," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 5, pp. 1711–1723, Jun. 2007.

[5] F. Weninger, B. Schuller, A. Batliner, S. Steidl, and D. Seppi, "Recognition of nonprototypical emotions in reverberated and noisy speech by nonnegative matrix factorization," *Journal on Advances in Signal Processing, Special Issue on Emotion and Mental State Recognition from Speech*, vol. 2011, 2011, article ID 838790, 16 pages.

[6] D. D. Lee, H. Seung *et al.*, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[7] F. Weninger, C. Kirst, B. Schuller, and H.-J. Bungartz, "A discriminative approach to polyphonic piano note transcription using supervised non-negative matrix factorization," in *Proc. of ICASSP*, Vancouver, Canada, May 2013, pp. 6–10.

[8] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *Speech and Audio Processing, IEEE Transactions on*, vol. 9, no. 5, pp. 504–512, July 2001.

[9] C. Joder, F. Weninger, V. David, and B. Schuller, "Integrating noise estimation and factorization-based speech spearation: A novel hybrid approach," in *Proc. of ICASSP*, Vancouver, Canada, May 2013, pp. 131–135.

[10] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, no. 4, pp. 1462 – 1469, July 2006.

[11] D. Fitzgerald, "Upmixing from mono: a source separation approach," in *17th International Conference on Digital Signal Processing*, Corfu, Greece, July 2011.

[12] R. Orban, "A rational technique for synthesizing pseudo-stereo from monophonic sources," *J. Audio Eng. Soc*, vol. 18, no. 2, pp. 157–164, 1970.

[13] M. Lagrange, L. G. Martins, and G. Tzanetakis, "Semi-automatic mono to stereo up-mixing using sound source formation," in *Audio Engineering Society Convention 122*, May 2007.

[14] M. A. Gerzon, "Signal processing for simulating realistic stereo images," in *Audio Engineering Society Convention 93*, Oct 1992.

[15] F. Weninger and B. Schuller, "Optimization and parallelization of monaural source separation algorithms in the openBliSSART toolkit," *Journal of Signal Processing Systems*, vol. 69, pp. 267–277, 2012.

[16] Z. Duan, G. J. Mysore, and P. Smaragdis, "Speech enhancement by online non-negative spectrogram decomposition in non-stationary noise environments," in *Proc. of Interspeech*, Portland, USA, 2012.

[17] C. Joder, F. Weninger, V. David, and B. Schuller, "A comparative study on sparsity penalties for NMF-based speech separation: Beyond LP-norms," in *Proc. of ICASSP*, Vancouver, Canada, May 2013, pp. 858–862.

[18] M. A. Pitt, L. Dilley, K. Johnson, S. Kiesling, W. Raymond, E. Hume, and E. Fosler-Lussier, *Buckeye Corpus of Conversational Speech (2nd release)*. Columbus, OH, USA: Department of Psychology, Ohio State University (Distributor), 2007, [www.buckeyecorpus.osu.edu].

[19] H. Christensen, J. Barker, N. Ma, and P. Green, "The CHiME corpus: a resource and a challenge for computational hearing in multisource environments," in *Proc. of Interspeech*, Makuhari, Japan, Sep. 2010, pp. 1918–1921.

[20] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 16, no. 1, pp. 229–238, Jan. 2008.