SUPERVISED AND SEMI-SUPERVISED SUPPRESSION OF BACKGROUND MUSIC IN MONAURAL SPEECH RECORDINGS

Felix Weninger, Jordi Feliu, Björn Schuller

Institute for Human-Machine Communication, Technische Universität München, Germany

weninger@tum.de, jfeliu2@mytum.de, schuller@tum.de

ABSTRACT

In this paper, we propose a semi-supervised algorithm based on sparse non-negative matrix factorization (NMF) to improve separation of speech from background music in monaural signals. In our approach, fixed speech basis vectors are obtained from training data whereas music bases are estimated on-the-fly to cope with spectral variability while preserving small NMF dimensionality for decreased computation effort. In a large-scale experimental evaluation with 168 speakers from the TIMIT database, we compare the semi-supervised method to supervised NMF with an explicit background music model. Our results reveal that the semi-supervised method outperforms supervised NMF at low speech-to-music ratios, and that sparsity constraints on the music spectra to enforce harmonicity can improve separation performance.

Index Terms— non-negative matrix factorization, supervised source separation, speech enhancement, sparse coding

1. INTRODUCTION

Separation of speech overlaid with music in monaural signals remains a challenging problem, especially due to the large similarity between voiced (harmonic instruments, vowels) and unvoiced (drums, consonants) parts. On the other hand, robust suppression of background music can be immediately exploited in a variety of applications, comprising speech enhancement for in-car human-machine interfaces or mobile telephony in highly noisy environments such as discotheques, speech recognition for multimedia information retrieval in TV series or on-line videos, or even lyrics transcription of rap/hip-hop music.

It is only recently that first promising results for monaural background music suppression have been obtained in [1], indicating that non-negative matrix factorization (NMF) is a promising approach to ensure intelligibility of the extracted speech signal. A more comprehensive evaluation has been carried out in [2], using an exemplar-based approach based on supervised NMF, that is, predefining a large set of base (speech and music) spectra extracted from training data whose non-negative activations in the test signal are found by an iterative algorithm; then, an estimate of the clean speech signal is obtained from the product of speech spectra and their activations. Still, it is unknown whether information about the music signal as in fully supervised NMF is required for optimal separation.

Hence, in this paper, we consider a semi-supervised variant of NMF [3] where only speech spectra are pre-defined, whereas the music bases are not characterized a priori, in order to cope with variability of the music over time, as in [1]. We compare the semi-supervised method against an 'upper bound' for the performance of (supervised) NMF where music bases are estimated from parts of the ground truth music, and evaluate the influence of sparsity constraints. Unlike [1], we enforce sparsity constraints on the NMF activations—similar to the algorithm proposed in [4]—to improve discrimination of speech and music bases, and extend this algorithm to sparse spectral bases in order to model harmonicity of music spectra. In this study, we use a rather small set of speech basis vectors for initialization that are learnt from training data by NMF, to vastly decrease computational effort compared to exemplar-based approaches such as [2]. A speaker-dependent scenario with 168 speakers from the TIMIT database is chosen for evaluation, and different music styles are investigated. Methods for semi-supervised sparse NMF are outlined in Section 2; details on experimental setup and results are given in Section 3 before concluding in Section 4.

The research leading to these results has been partly funded by the Federal Republic of Germany through the German Research Foundation (DFG) under grant No. SCHU 2508/2-1.

2. NMF-BASED METHODS FOR MUSIC SUPPRESSION

2.1. Signal Model

Our approach for music suppression in monaural speech recordings is based on the assumption that speech is 'corrupted' by addition of background music:

$$\mathbf{V} = \mathbf{V}^{(s)} + \mathbf{V}^{(m)},$$

where $\mathbf{V} \in \mathbb{R}_+^{M \times N}$ is an observed magnitude spectrogram of speech overlaid by music, $\mathbf{V}^{(s)}$ is the (true) spectrogram of the speech signal, and $\mathbf{V}^{(m)}$ is the (true) music spectrogram. Furthermore, we assume that both, the speech and noise spectrograms can be approximated as linear combinations of base spectra (dictionaries) $\mathbf{w}_j^{(s)} \in \mathbb{R}_+^M$, $j = 1, \ldots, R^{(s)}$, respectively $\mathbf{w}_j^{(m)}$, $j = 1, \ldots, R^{(m)}$, with non-negative coefficients (activations) $\mathbf{H}^{(s)} \in \mathbb{R}_+^{R^{(s)} \times N}$, $\mathbf{H}^{(m)} \in \mathbb{R}_+^{R^{(n)} \times N}$. Defining

$$\mathbf{W}^{(s)} = [\mathbf{w}_1^{(s)} \ \cdots \ \mathbf{w}_{R^{(s)}}^{(s)})]$$

and $\mathbf{W}^{(m)}$ analogously, this signal model can be written in matrix notation, where $\mathbf{\Lambda}$, $\mathbf{\Lambda}^{(s)}$ and $\mathbf{\Lambda}^{(m)}$ denote approximations of $\mathbf{V}$, $\mathbf{V}^{(s)}$ and $\mathbf{V}^{(m)}$, respectively:

$$\mathbf{\Lambda} = \mathbf{\Lambda}^{(s)} + \mathbf{\Lambda}^{(m)} = \mathbf{W}^{(s)}\mathbf{H}^{(s)} + \mathbf{W}^{(m)}\mathbf{H}^{(m)},$$

or $\mathbf{\Lambda} = \mathbf{W}\mathbf{H}$ for $\mathbf{W} := [\mathbf{W}^{(s)}\mathbf{W}^{(m)}]$, $\mathbf{H} := \begin{bmatrix} \mathbf{H}^{(s)} \\ \mathbf{H}^{(m)} \end{bmatrix}$.

2.2. Supervised and Sparse Semi-Supervised NMF

In the remainder of this paper, we assume that the speech 'basis' $\mathbf{W}^{(s)}$ is fixed after estimation from training data. More precisely, it is computed by reducing a set of training utterances through NMF, as proposed, e. g., in [5]. A fully supervised NMF approach for background music suppression, in analogy to [3, 5], is obtained when a similar procedure is followed for the music basis $\mathbf{W}^{(m)}$ as well. In that case, the speech enhancement problem is reduced to finding non-negative coefficients (activations) $\mathbf{H}^{(s)}$ and $\mathbf{H}^{(n)}$. In the proposed *semi-supervised* approach however, $\mathbf{W}^{(m)}$ is estimated along with $\mathbf{H}^{(s)}$ and $\mathbf{H}^{(n)}$ such that the following cost function is minimized:

$$\begin{aligned} c(\mathbf{W}^{(m)}, \mathbf{H}) &= c_r(\mathbf{W}^{(m)}, \mathbf{H}) \\ &+ \lambda\, c_s^{\mathbf{H}}(\mathbf{H}^{(m)}) + \mu c_s^{\mathbf{W}}(\mathbf{W}^{(m)}) \end{aligned} \quad (1)$$

where c_r corresponds to the reconstruction error as β-divergence $d_\beta(\mathbf{V}|\mathbf{\Lambda})$ for $\beta = 1$ (Kullback-Leibler divergence) and

$$\begin{aligned} c_s(\mathbf{H}^{(m)}) &= \sum_{j=1}^{R^{(m)}} \frac{1}{\sigma(\mathbf{H}^{(m)}_{j,:})} \sum_{t=1}^{N} \mathbf{H}^{(m)}_{j,t} \\ c_s(\mathbf{W}^{(m)}) &= \sum_{j=1}^{R^{(m)}} \frac{1}{\sigma(\mathbf{W}^{(m)}_{:,j})} \sum_{k=1}^{M} \mathbf{W}^{(m)}_{k,j}. \end{aligned}$$

λ and μ are free parameters ($0 \leq \lambda, \mu \ll 1$), and $\sigma(\mathbf{W}^{(m)}_{:,j})$ and $\sigma(\mathbf{H}^{(m)}_{j,:})$ are standard deviation estimates for the j-th column of $\mathbf{W}^{(m)}$ and the j-th row of $\mathbf{H}^{(m)}$, respectively that are introduced to avoid dependency on the scaling of the matrices, following [4].

Informally, c_s is a sparsity constraint that is only enforced on the music part: The purpose of imposing sparsity on $\mathbf{H}^{(m)}$ is to mitigate the fact that the algorithm can 'mis-use' the bases designated to isolate the music for modeling the speech parts; additionally, sparsity on $\mathbf{W}^{(m)}$ is imposed to increase the discrimination between speech and music, as the latter is arguably characterized by higher harmonicity compared to speech. Unlike in recent studies that exploit NMF for speech recognition directly such as [6], the purpose of sparsity is not to force that only a few basis vectors can be active at a given time: As the base estimation from training data is entirely unsupervised, they may largely overlap in their spectral and/or phonetic content.

The cost function (1) is minized by applying component-wise multiplicative updates to $\mathbf{W}^{(m)}$, $\mathbf{H}^{(s)}$ and $\mathbf{H}^{(m)}$ based on the algorithm proposed in [4]. We straightforwardly extend the algorithm to the semi-supervised case, including the sparsity constraint for the spectra $\mathbf{W}^{(m)}$ which was not considered in [4], yielding the following update rule for $\mathbf{W}^{(m)}$ ($\otimes$ denotes the Hadamard product):

$$\mathbf{W}^{(m)} \leftarrow \mathbf{W}^{(m)} \otimes \frac{\nabla c^-(\mathbf{W}^{(m)}, \mathbf{H})}{\nabla c^+(\mathbf{W}^{(m)}, \mathbf{H})}$$

where ∇^+ and ∇^- indicate the positive and negative parts of the gradient, respectively, which in turn are determined by $\nabla c_r^+(\mathbf{W}^{(m)})$ and $\nabla c_r^-(\mathbf{W}^{(m)})$ as laid out in [4] and

$$\begin{aligned} [\nabla c_s^{\mathbf{W}^+}(\mathbf{W}^{(m)})]_{i,j} &= \frac{\sqrt{M}}{\sqrt{\sum_{k=1}^{M} \mathbf{W}^{(m)2}_{k,j}}} \\ [\nabla c_s^{\mathbf{W}^-}(\mathbf{W}^{(m)})]_{i,j} &= \mathbf{W}^{(m)}_{i,j} \frac{\sqrt{M} \sum_{k=1}^{M} \mathbf{W}^{(m)}_{k,j}}{(\sum_{k=1}^{M} \mathbf{W}^{(m)2}_{k,j})^{3/2}} \end{aligned}$$

The update rules are applied for 100 iterations starting from a (Gaussian) random solution. Finally, the estimated clean speech spectrogram $\widehat{\mathbf{V}}^{(s)}$ is obtained by filtering the observed spectrogram $\mathbf{V}$:

$$\widehat{\mathbf{V}}^{(s)} = \frac{\mathbf{\Lambda}^{(s)}}{\mathbf{\Lambda}} \otimes \mathbf{V}.$$

Note that the asymptotic complexity of this algorithm is polynomial ($O(RMN)$), and linear in each of $R := R^{(s)} + R^{(m)}$, M and N. This means that especially for applications with real-time constraints, it is desirable to keep the number of components R as low as possible at a reasonable separation quality. All experiments for this paper are based on the NMF implementations found in our open-source toolkit openBliSSART [7] to enforce reproducibility of our results.

3. EXPERIMENTAL SETUP AND RESULTS

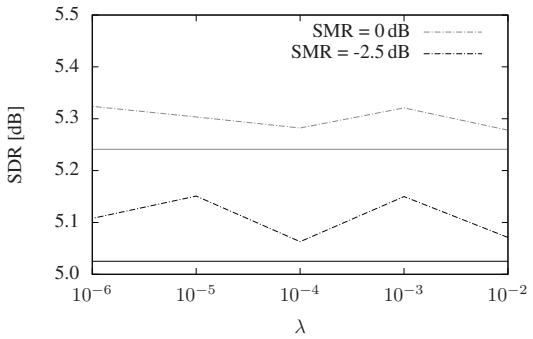
The aim of our experiments is to evaluate the extraction of the speech from mixed speech and music audio signals, as well as to determine the influence of sparsity weights, Discrete Fourier Transform (DFT) window size and music style.

3.1. Evaluation Data Set

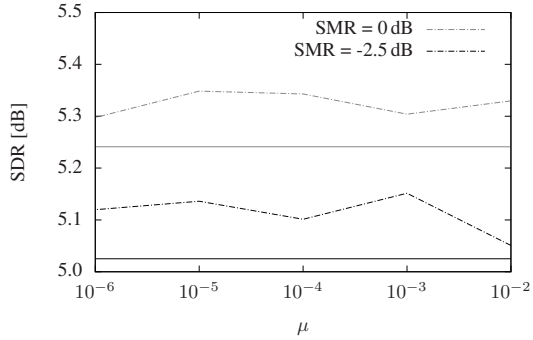
Our evaluation set is formed by 1 680 audio signals (sentences) spoken by 168 different subjects (56 females and 112 males) from the TIMIT database test set, i. e., there are 10 sentences of typically 2–3 seconds length for each speaker. We chose the TIMIT database for its rich phonetic content. Each of the TIMIT test sentences is artificially mixed with a random cut of music of the same length at various speech-to-music ratios (SMR) from -7.5 dB to +5 dB in intervals of 2.5 dB. These SMRs correspond to typical application scenarios as indicated in the introduction. To demonstrate performance of our method on a variety of music styles, the experiment was first carried out with the 136 Viennese Waltzes from the Ballroom Dance (BRD) Database [8] as an example for classical music, then it was repeated with 136 pieces of each of the latin, jazz and rock genres. Note that these frequently contain segments with sung vocals, which makes separation of speech particularly challenging.

3.2. Experimental Setup

In all experiments we evaluate by means of a speaker-dependent cross-validation, i. e., for each TIMIT test sentence all other sentences of the same speaker are concatenated, and their spectrogram is reduced to a NMF speech basis $\mathbf{W}^{(s)}$. As this results in aproximately 20–30 seconds training material for separation of each test instance, this methodology represents realistic use-cases where the user adapts the system to his/her voice. For supervised NMF, music bases $\mathbf{W}^{(m)}$ are computed from a disjoint random section of 25 seconds of the same music track used for mixing the test file—this yields an upper benchmark on the performance of supervised NMF assuming the exact characteristics of the music are known during separation. The NMF dimensionality parameters $R^{(m)}$ and $R^{(s)}$ were set to 10 and 20, respectively. 20 speech components have been found to represent a good compromise between separation quality and computational complexity [5], and the ratio 2/1 for the speech and music bases was chosen empirically in preliminary experiments. The experiment was repeated for different Hann window sizes from 8 ms to 512 ms whereas the overlap between consecutive DFT frames in the time domain remained fixed at 75 %.

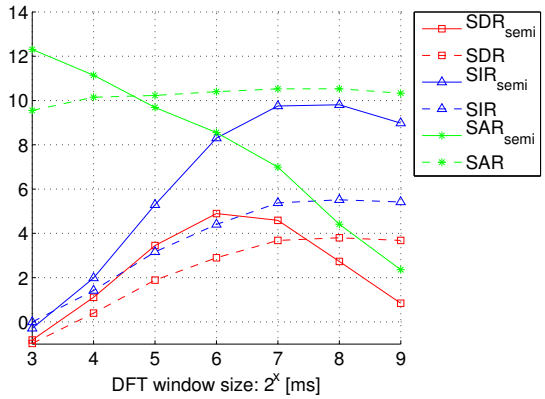


(a) Sparsity on $\mathbf{H}^{(m)}$

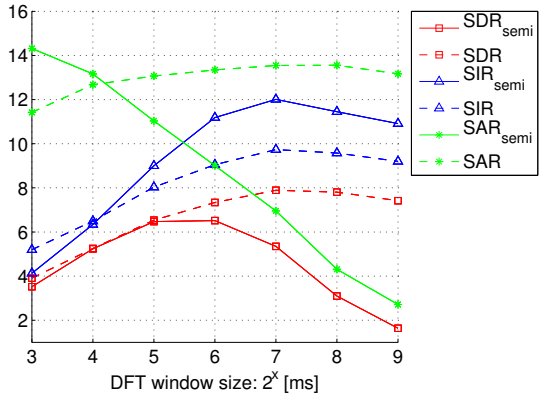


(b) Sparsity on $\mathbf{W}^{(m)}$

Fig. 1: Average signal-to-distortion ratio (SDR) on TIMIT test set (1 680 sentences overlaid with waltzes from BRD database) by sparse semi-supervised NMF (dashed-dotted lines) for SMRs of -2.5 and 0 dB. Variation of the sparsity weights λ for $\mu = 0$ (a), and of μ for $\lambda = 0$ (b). Continuous lines: semi-supervised NMF with $\lambda = \mu = 0$, i. e., no sparsity constraints. DFT window size 128 ms.



(a) SMR = -5 dB



(b) SMR = 0 dB

Fig. 2: Average separation performance on TIMIT test set (1 680 sentences) overlaid with waltzes from the BRD database. Effect of DFT window size on non-sparse semi-supervised NMF separation (dashed lines) compared to supervised NMF (continuous lines).

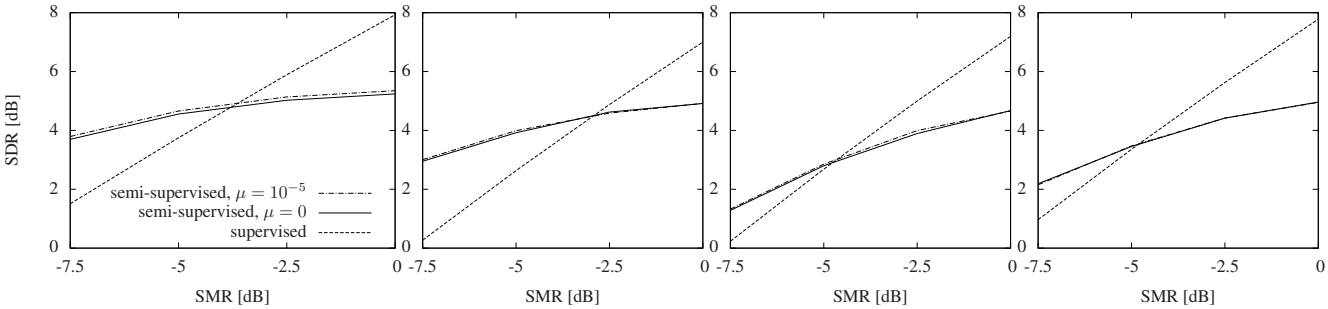
To assess the characteristics of semi-supervised and supervised speech and music separation in detail, we employ signal-to-distortion ratio (SDR) as a measure of overall separation quality, source-to-interference ratio (SIR) to quantify suppression of the undesired music source (which may however lead to information loss in the speech signal due to spectral overlap), and source-to-artifact ratio (SAR) to evaluate degradation of speech quality by the separation. Measurements were carried out using the open-source BSS_Eval toolkit [9].

3.3. Results

Figure 1 shows the performance in terms of SDR for semi-supervised NMF when varying the sparsity weight λ for the music activations $\mathbf{H}^{(m)}$ while keeping the sparsity weight μ for the music spectra $\mathbf{W}^{(m)}$ constant at zero (1a), and vice versa (1b). In both evaluation scenarios, sparsity constraints slightly—yet consistently—increase performance by over 0.1 dB absolute at all SMRs. Overall, the harmonicity constraint (sparsity of $\mathbf{W}^{(m)}$) is most promising. Across SMRs, best results are obtained at $\mu = 10^{-5}$. We found that using both $\lambda, \mu > 0$ could not further improve results. Evidently, this sparsity parameter is highly 'non-aggressive'—larger values of μ would probably force a reduction of the $\mathbf{W}^{(m)}$ to single harmonics which is not desirable in the case of complex music.

Next, Figures 2a and 2b show the results in terms of SDR (indicated by squares), SIR (triangles) and SAR (asterisks) for both supervised and (non-sparse) semi-supervised NMF at SMRs of -5 dB and 0 dB, respectively for varying window sizes (8, 16, 32, 64, 128, 256 and 512 ms), corresponding to DFT sizes of 128–8192 points. For both supervised and semi-supervised NMF, the best suppression (SIR) is achieved at a window size of 128 ms while small window sizes ($<$ 32 ms) do not enable robust suppression in general. For a SMR of -5 dB, the semi-supervised method improves SIR by more than 4 dB compared to the supervised case, boosting the average SIR to almost 10 dB. At 0 dB SMR the improvement by semi-supervised NMF is smaller but still clearly visible, achieving over 12 dB average SIR. In terms of overall quality (SDR), at -5 dB semi-supervised NMF delivers equal or higher SDR for up to 128 ms window size; at 0 dB, larger window sizes ($>$ 64 ms) decrease the SAR—and hence the SDR—of the separated speech for semi-supervised NMF. We conclude that semi-supervised NMF seems to be prone to lose some speech information at higher SMRs. Since this effect does not occur for supervised NMF, we argue that larger window sizes help to create a more precise model of the true music in supervised NMF.

The rather slight effect of using sparsity constraints on the music bases deserves some further investigation. Figures 3a through 3d show how the music genre affects the separation for four different styles: classical, jazz, latin and rock. The experiments are done using



(a) waltz (b) jazz (c) latin (d) rock

Fig. 3: Separation of TIMIT test set (1 680 sentences): Average SDR results for speech overlaid by different music styles, for sparse semi-supervised ($\mu = 10^{-5}, \lambda = 0$), semi-supervised ($\mu = \lambda = 0$) and supervised NMF. DFT window size 128 ms.

a DFT window size equal to 128 ms with sparse ($\mu = 10^{-5}, \lambda = 0$) and non-sparse ($\mu = \lambda = 0$) semi-supervised as well as supervised NMF. The results reveal that the improvement by using sparse instead of non-sparse semi-supervised NMF—i. e., enforcing harmonicity in the music spectra—is mostly visible for waltz music with its arguably high degree of harmonicity, while for jazz and rock music no gains can be observed.

Corroborating the results obtained for speech recognition accuracy in [2] on the acoustic level, all methods exhibit highest performance for waltz music while results consistently downgrade for the other styles: On average a loss of almost 2 dB SDR is observed for jazz compared to waltz music, which can be attributed to the complex harmonic and rhythmic structure of the jazz genre which also decreases the SDR obtained by supervised NMF; more music components could probably increase performance of the latter.

4. CONCLUSIONS

We have presented a large-scale study on performance of semi-supervised and supervised NMF algorithms for compensation of background music in speech and have demonstrated the effectiveness of semi-supervised NMF particularly in highly noisy environments. Comparing the semi-supervised method to an upper benchmark for supervised NMF assuming the characteristics of the music are known, it is notable that in highly 'noisy' conditions, the semi-supervised method suppresses the music to a larger extent than the supervised benchmark, in terms of SIR; however, at higher speech-to-music ratios a decrease in overall quality (SDR) has to be accepted. We attribute this to the relative simple modeling of speech by predefined spectral vectors, which can cause information loss in the reconstructed speech signal since subtleties of speech not modeled by the predefined basis are captured by the iteratively updated basis vectors designated to contain the background music. By enforcing sparsity constraints on the spectra and on the activations, the performance of semi-supervised NMF could be improved slightly, but the gain strongly depends on the genre and according complexity of the music signal.

Future work could focus on integration of automatic genre recognition on the separated music signal in a two-stage separation algorithm: This enables use of optimal NMF parameterizations for different genres. Furthermore, the relatively simple speech modeling and factorization constraints in this study could be extended with more advanced techniques such as temporal dependencies, as in [10].

5. REFERENCES

[1] A. Ozerov, C. Févotte, and M. Charbit, "Factorial scaled hidden Markov model for polyphonic audio representation and source separation," in *Proc. of WASPAA*, Mohonk, NY, United States, 2009, pp. 121–124.

[2] B. Raj, T. Virtanen, S. Chaudhuri, and R. Singh, "Non-negative matrix factorization based compensation of music for automatic speech recognition," in *Proc. of Interspeech*, Makuhari, Japan, 2010.

[3] P. Smaragdis, B. Raj, and M. Shashanka, "Supervised and semi-supervised separation of sounds from single-channel mixtures," in *Proc. of ICA*, Berlin, Heidelberg, 2007, pp. 414–421, Springer-Verlag.

[4] T. Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 3, pp. 1066–1074, March 2007.

[5] P. Smaragdis, "Convolutive speech bases and their application to supervised speech separation," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 1, pp. 1–14, 2007.

[6] J.F. Gemmeke, T. Virtanen, and A. Hurmalainen, "Exemplar-based sparse representations for noise robust automatic speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2067–2080, 2011.

[7] F. Weninger, A. Lehmann, and B. Schuller, "openBliSSART: Design and Evaluation of a Research Toolkit for Blind Source Separation in Audio Recognition Tasks," in *Proc. of ICASSP*, Prague, Czech Republic, 2011.

[8] B. Schuller, F. Eyben, and G. Rigoll, "Tango or Waltz?—Putting Ballroom Dance Style into Tempo Detection," *EURASIP Journal on Audio, Speech, and Music Processing (JASMP), Special Issue on "Intelligent Audio, Speech, and Music Processing Applications"*, 2008, Article ID 846135, 12 pages.

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

[10] G. J. Mysore and P. Smaragdis, "A Non-Negative Approach to Semi-Supervised Separation of Speech from Noise with the Use of Temporal Dynamics," in *Proc. of ICASSP*, Prague, Czech Republic, 2011.