Spectrally and spatially informed noise suppression using beamforming and convolutive NMF

Benjamin Cauchi[1,4], Timo Gerkmann[2,4], Simon Doclo[1,2,4], Patrick A. Naylor[3] and Stefan Goetze[1,4]

[1]*Fraunhofer Institute for Digital Media Technology IDMT, 26129, Oldenburg, Germany*

[2]*University of Oldenburg, 26111, Oldenburg, Germany*

[3]*Imperial College London, SW7 2AZ, London, United Kingdom*

[4]*Cluster of Excellence, Oldenburg, Germany*

Correspondence should be addressed to Benjamin Cauchi (benjamin.cauchi@idmt.fraunhofer.de)

ABSTRACT

Speech enhancement in low SNR conditions or in presence of large amount of reverberation is a challenging task. However, in some applications, prior information about the interfering noise source is available and can be exploited to tackle this issue. We propose to combine a beamformer with convolutive NMF in order to estimate the PSDs of the target speech signal and of the noise to be suppressed by exploiting knowledge of the noise source location and about its spectral content. We apply the proposed system to ego-noise suppression for a robotic platform. Simulations show that the spectral information exploited using convolutive NMF is beneficial to the noise reduction performance when compared to methods based on blind estimation but that estimating the noise PSD from the output of the beamformer is beneficial mostly when no prior knowledge of the noise spectral content is available.

1. INTRODUCTION

In speech communication or automatic speech recognition (ASR) applications, the speech signal of the user is often recorded by distant microphones. Usually, the signal is corrupted by additive noise which can degrade the speech quality and intelligibility, as well as the performance of ASR systems or the accuracy of source localization. Therefore, numerous approaches have been proposed to reduce the noise present in a recorded signal. Noise suppression is a popular approach which applies a frequency-dependent spectral gain, e.g. the Wiener gain, to the short-time Fourier transform (STFT) of the input signal [1]. The computation of this spectral gain typically requires an estimate of the noise power spectral density (PSD) obtained, e.g, using estimators derived from statistical models [2, 3]. Noise suppression is often applied to the output of a beamformer aiming at suppression of the sound sources whose directions of arrival (DOA) differ from the DOA of the target speaker. This combination of spectral and spatial filtering has been shown efficient in many speech enhancement applications [4, 5]. Unfortunately, when the signal to noise ratio (SNR) in the recorded signal is low, the estimate of the noise PSD obtained using statisitical models might be inaccurate. Additionaly, the DOA of the target speaker is often unknown and errors in its estimation may lead to the beamformer cancelling the target signal. Therefore, blind speech enhancement in low SNR conditions remains a challenge.

However, in some applications, the noise source and its location are known, e.g. when recording speech on a robotic platform [6]. In such cases, the DOA of the noise source and knowledge of its spectral content can be taken advantage of when designing the speech enhancement scheme to be used. Several methods have been proposed to accomplish this task. Bases representative of the noise to be suppressed can be learned using non-negative matrix factorization (NMF) [7] or a Bayesian extension of the noise to be suppressed [8]. In [9], dictionaries of both spectral and spatial features are learned before hand and used to enhance the STFT of the input signal. In this paper, we use convolutive non-negative matrix factoriza-

tion (CNMF) [10], in order to benefit from the spectral knowledge of the noise signal to be suppressed. Contrary to the standard NMF algorithm, CNMF factorizes an input matrix using time dependent basis, therefore allowing to take into account the time dependency of the signals to be processed. We apply CNMF to estimate the PSDs of the noise and of the speech and use these estimates to compute a spectral gain to be applied in a spectral enhancement scheme. Knowledge of the DOA of the noise source is exploited by estimating the noise PSD from the output of a beamformer steered towards the interfering noise source.

The remainder of this paper is structured as follows. The proposed system is presented in Section 2 before CNMF is briefly introduced in Section 3.1 and its application to PSDs estimation in Section 3.2. The proposed system is evaluated in the context of ego-noise suppression for which the experimental framework and the obtained results are presented in Section 4 before concluding the paper in Section 5.

2. PROBLEM STATEMENT

When using M microphones to record speech in an encosure containing a noise source, the reverberant and noisy m-th microphone signal $z_m(n)$ at time index n is given by

$$z_m(n) = x(n) * h_m^s(n) + y(n) * h_m^v(n) \quad (1)$$
$$= s_m(n) + v_m(n), \text{ for } m = 1, \cdots, M, \quad (2)$$

with $x(n)$ and $y(n)$ denoting the anechoic speech and noise signals, respectively, $h_m^s(n)$ denoting the room impulse response (RIR) between the speech source and the m-th microphone, $h_m^v(n)$ denoting the RIR between the noise source and the m-th microphone, and $s_m(n)$ and $v_m(n)$ denoting the reverberant speech component and the additive noise component in the m-th microphone signal, respectively. In the remainder of this paper, the STFT representations of $z_m(n)$, $s_m(n)$ and $v_m(n)$ are denoted by $z_m(k,\ell)$, $s_m(k,\ell)$ and $v_m(k,\ell)$, respectively, with k and ℓ denoting the discrete frequency bin and frame indices, respectively.

The proposed system aims at obtaining an estimate $\hat{s}_\rho(n)$, with $\hat{\cdot}$ denoting estimated quantities, of the reverberant speech signal $s_\rho(n)$ in the arbitrarly chosen reference channel of index ρ, from the microphone signals, $z_m(n)$.

In the STFT domain, (2) can be rewritten as

$$z_m(k,\ell) = s_m(k,\ell) + v_m(k,\ell), \text{ for } m = 1, \cdots, M. \quad (3)$$

The proposed system consists in applying a real-valued spectral gain $g(k,\ell)$ to the STFT coefficients of the signal recorded in the reference channel, i.e.

$$\hat{s}_\rho(k,\ell) = g(k,\ell) z_\rho(k,\ell), \quad (4)$$

with $\hat{s}_\rho(k,\ell)$ denoting the estimated STFT of the target speech signal from which $\hat{s}_\rho(n)$ is obtained. The spectral gain $g(k,\ell)$ is typically computed using estimates of the speech and of the noise PSDs, e.g. when using the Wiener gain,

$$g(k,\ell) = \frac{\hat{\sigma}_s^2(k,\ell)}{\hat{\sigma}_s^2(k,\ell) + \hat{\sigma}_v^2(k,\ell)} \quad (5)$$

with $\hat{\sigma}_s^2(k,\ell)$ and $\hat{\sigma}_v^2(k,\ell)$ denoting estimates of $\sigma_s^2(k,\ell) = \mathrm{E}\left\{|s_\rho(k,\ell)|^2\right\}$ and $\sigma_v^2(k,\ell) = \mathrm{E}\left\{|v_\rho(k,\ell)|^2\right\}$, respectively, and with $\mathrm{E}\{\cdot\}$ denoting the expectation operator. Contrary to many speech enhancement applications that assume that the DOA of the target speaker is known or use knowlege of the speech spectral content [8], the proposed system does not use any prior information about the target speech. However, it is assumed that the DOA, θ_v, of the noise source is known and that a recording of a previous realization of the noise is available.

The proposed system can be summarized as follows. First, a beamformer, aiming at reducing the speech by suppressing the sound sources not arriving from the DOA of the noise source, is applied by filtering and summing the microphone signals as

$$\tilde{z}(k,\ell) = W_{\theta_v}^{\mathrm{H}}(k) Z(k,\ell) \quad (6)$$

with

$$Z(k,\ell) = [z_1(k,\ell)\ z_2(k,\ell)\ \ldots\ z_M(k,\ell)]^{\mathrm{T}}, \quad (7)$$

denoting the M-dimensional stacked vector of the received microphone signals and with $W_{\theta_v}(k)$ and $\tilde{z}(k,\ell)$ denoting the stacked filter coefficient vector of the beamformer steered towards the angle θ_v and the STFT of the output of the beamformer, respectively. The estimate $\hat{\sigma}_v^2(k,\ell)$ of the noise PSD is then estimated from $\tilde{z}(k,\ell)$ and the estimate $\hat{\sigma}_s^2(k,\ell)$ of the speech PSD is estimated from $z_\rho(k,\ell)$. It can be noted that computing

$\hat{\sigma}_v^2(k,\ell)$ from the output of the beamformer can prevent the speech power from leaking into the estimate of the noise PSD but might lead to an underestimation.

The influence of this beamforming step on the performance of the proposed system is examined in Section 4.2 and the computation of both $\hat{\sigma}_v^2(k,\ell)$ and $\hat{\sigma}_s^2(k,\ell)$, using CNMF, is described in the next section.

3. PSDS ESTIMATION USING CNMF

3.1. Convolutive NMF

CNMF is an extension of the well known NMF which aims at computing a factorization of a non-negative matrix $\mathbf{M}$ into a set of activation coefficients $\mathbf{A}$ and a set of bases $\mathbf{D}$. While in NMF, each basis in $\mathbf{D}$ is a vector, in CNMF, each basis is a matrix, which, in audio signal processing applications, allows to take into account the time dependency of the input signal.

In this paper, the input data to be factorized, i.e. $\mathbf{M}$, is the periodogram of an audio signal. Therefore, in the remainder of this paper, $\mathbf{M} \in \mathbb{R}^{\geq 0, K\times L}$, with K and L denoting the number of frequency bins and the number of frames present in the factorized periodogram, respectively. The resulting factorization can be expressed as

$$\mathbf{M} \approx \hat{\mathbf{M}} = \sum_{t=0}^{T} \mathbf{D}(t) \cdot \overset{t\rightarrow}{\mathbf{A}}, \tag{8}$$

with R denoting the number of bases to be extracted, $\mathbf{A} \in \mathbb{R}^{\geq 0, R\times L}$ and with $\mathbf{D}(t) \in \mathbb{R}^{\geq 0, K\times T}$, with T denoting the number of frames in each extracted basis. The i-th column of $\mathbf{D}(t)$ denotes the i-th time frame of the t-th basis. Finally, the operator $\overset{t\rightarrow}{\cdot}$ denotes a shift of t columns to the right while $\overset{\leftarrow t}{\cdot}$ denotes a shift of t columns to the left.

The application of CNMF, as presented in [10], consists in iteratively updating $\mathbf{A}$ and $\mathbf{D}$ in order to minimize the cost function $\mathscr{D}(\mathbf{M}\|\hat{\mathbf{M}})$ defined as

$$\mathscr{D}(\mathbf{M}\|\hat{\mathbf{M}}) = \left\| \mathbf{M} \otimes \log\frac{\mathbf{M}}{\hat{\mathbf{M}}} - \mathbf{M} + \hat{\mathbf{M}} \right\|. \tag{9}$$

The update rule minimizing this cost function can be expressed as

$$\mathbf{A} = \mathbf{A} \otimes \frac{\mathbf{D}(t)^{\mathrm{T}} \cdot \overset{\leftarrow t}{\left[\frac{\mathbf{M}}{\hat{\mathbf{M}}}\right]}}{\mathbf{D}(t)^{\mathrm{T}} \cdot \mathbf{1}}, \quad \mathbf{D}(t) = \mathbf{D}(t) \otimes \frac{\frac{\mathbf{M}}{\hat{\mathbf{M}}} \cdot \overset{t\rightarrow^{\mathrm{T}}}{\mathbf{A}}}{\mathbf{1} \cdot \overset{t\rightarrow^{\mathrm{T}}}{\mathbf{A}}}, \tag{10}$$

with $\mathbf{A}$ being averaged over all t after each update. The updating process can be stopped either based on a stopping criterion, depending on the value of $\mathscr{D}(\mathbf{M}\|\hat{\mathbf{M}})$, or simply once a fixed number of N iterations has been computed. The CNMF factorization can be applied blindly, i.e. randomly initializing both $\mathbf{D}$ and $\mathbf{A}$ or in a supervised way, i.e, initializing $\mathbf{D}$ using knowledge of the content to factorize.

3.2. Application to PSDs estimation

In this paper, CNMF is first applied blindly to the matrix $\mathbf{M}_{\overline{v}}$ defined as

$$\mathbf{M}_{\overline{v}} = \begin{bmatrix} |\overline{v}(0,0)|^2 & |\overline{v}(0,1)|^2 & \cdots & |\overline{v}(0,L)|^2 \\ |\overline{v}(1,0)|^2 & |\overline{v}(1,1)|^2 & \cdots & |\overline{v}(1,L)|^2 \\ \vdots & \vdots & \ddots & \vdots \\ |\overline{v}(K,0)|^2 & |\overline{v}(K,1)|^2 & \cdots & |\overline{v}(K,L)|^2 \end{bmatrix}, \tag{11}$$

with $|\overline{v}(k,\ell)|^2$ denoting the periodogram of the signal $\overline{v}(n)$ which consists of a recorded realization of the noise sound source to be suppressed. Randomly initializing both basis and activation matrix and applying N iterations of the updates described in (10) produces a matrix $\mathbf{D}_{\overline{v}}$ containing $R_{\overline{v}}$ bases representative of the considered noise source. It can be noted that the number of bases $R_{\overline{v}}$ has to be set by the user and that the optimal value will depend on the type of noise to be suppressed.

In order to estimate the noise PSD, the extracted basis matrix $\mathbf{D}_{\overline{v}}$ is concatenated with a matrix $\mathbf{D}_{\overline{s}}$ containing $R_{\overline{s}}$ randomly initialized bases. The resulting concatenated matrix is used as initialization for the updates described in (10) which are applied for N iterations using the matrix $\mathbf{M}_{\tilde{z}}$ as input, with $\mathbf{M}_{\tilde{z}}$ being constructed from the coefficients $\tilde{z}(k,\ell)$, similarly as in (11). The extracted matrix of activation coefficients, $\mathbf{A}_v$, will be used to estimate the noise PSD. The same process, using $\mathbf{M}_{z_\rho}$ constructed from the coefficients $|z_\rho(k,\ell)|^2$, is repeated in order to extract $\mathbf{D}_s$ and $\mathbf{A}_s$, which denote the matrices of bases and of activation, respectively, which will be used to estimate the speech PSD.

Finally, both $\sigma_s^2(k,\ell)$ and $\sigma_v^2(k,\ell)$ are estimated as

$$\hat{\sigma}_v^2(k,\ell) = \sum_{t=0}^{T} \mathbf{D}_{\overline{v}}(t) \cdot \overset{t\rightarrow}{\mathbf{A}}_v \tag{12}$$

$$\hat{\sigma}_s^2(k,\ell) = \sum_{t=0}^{T} \mathbf{D}_s(t) \cdot \overset{t\rightarrow}{\mathbf{A}}_s, \tag{13}$$

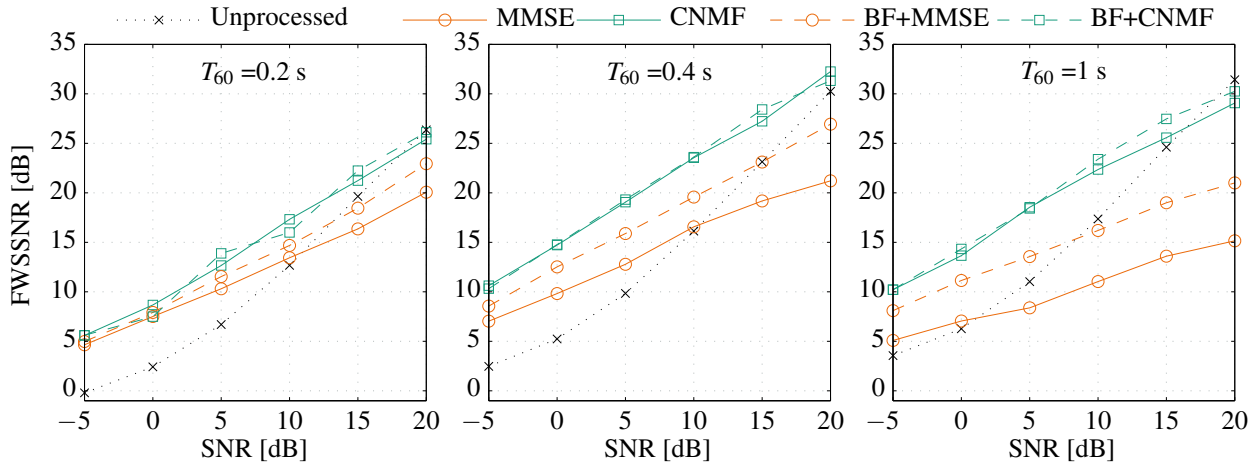

Fig. 1: Achieved FWSSNR as a function of the SNR for different T_{60}.

and these estimates are used to compute the gain described in (5) and estimate the clean speech STFT as in (4).

4. EXPERIMENT

4.1. Application to ego-noise suppression

In robotic applications, the speech from a user is typically recorded using an array of microphones mounted on a robot. As the robot generates noise while in action, and as the noise sources can be close to the microphones, the SNR of the recorded signal is often low, making the suppression of the so-called ego-noise a challenging task. However, the location of the noise source is usually known and a realization of the noise can be recorded before hand. Therefore, ego-noise suppression is an example of application in which the proposed system could be beneficial. We apply the proposed system, combining beamforming and CNMF to the suppression of motor noise generated by a robot platform Scitos A5 [11] during deplacements of the robot. Two realizations of the noise have been recorded by a microphone placed nearby the motor. One noise recording has been used to build the basis $\mathbf{D}_{\overline{v}}$ while the other has been used to simulate the noisy input data. The recorded signals have been simulated by convolving noise and speech signals with RIRs generated using the image method [12] in a room of dimension $5\times6\times4$ m, assuming that the motor is located 1 m under the center of a circular microphone array of $M = 8$ microphones and of 20 cm of diameter and that the speech source is located 2 m away from the center of the array. The speech signal consists of 10 minutes of speech built by concatenating utterances from the TIMIT database [13].

The proposed system, combining a beamformer steered towards the noise source with the spectral suppression scheme based on CNMF is denoted by BF+CNMF. This system is compared with the single channel case, i.e. $\tilde{z}(k,\ell) = z_1(k,\ell)$, denoted by CNMF, in order to evaluate the impact of the beamforming stage. In order to evaluate the benefit of estimating the speech and noise PSDs using CNMF, the proposed system is compared to a blind approach combining an minimum mean square error (MMSE) estimator of the noise PSD [3] with the MMSE estimator of the speech amplitude presented in [14]. This blind approach will be refered to as BF+MMSE in the multichannel case and as MMSE in the single-channel case.

The STFTs have been computed using a 32 ms Hann window with 50 % overlap. The RIRs have been generated for different levels of reverberation, with T_{60} ranging from 0.2 s to 1 s. The noise signal $v_m(n)$ has been added to the speech signal $s_m(n)$ at SNRs, measured in the 1-st channel, ranging from -5 dB to 20 dB. The applied beamformer consists of a delay and sum beamformer. All the bases extracted when applying CNMF contain $T = 6$ frames. The number of extracted basis has been set to $R_{\overline{v}} = R_s = 6$ and the number of iterations set to $N = 25$.

4.2. Results

The performance of the considered noise suppression systems is evaluated in terms of frequency-weighted seg-

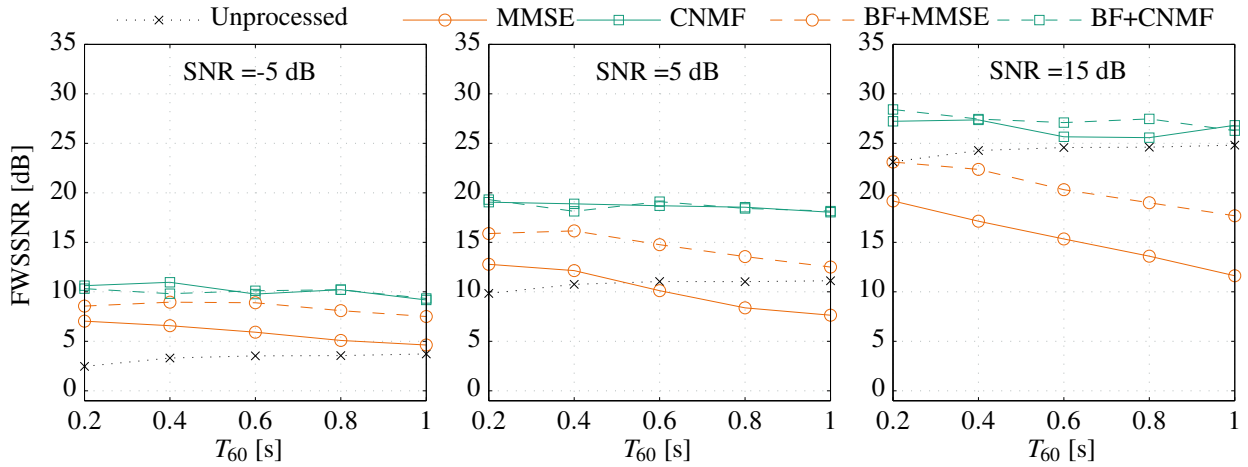

Fig. 2: Achieved FWSSNR as a function of the T60 for different input SNR.

mental SNR (FWSSNR) [15], for which the signal $s_1(n)$ has been used as a reference. Fig. 1 illustrates the achieved FWSSNR as a function of the SNR of the input signal. It can be observed that both CNMF and BF+CNMF achieve similar performance in the 3 different reverberant scenarios, suggesting that the accuracy of the these PSDs estimators does not depend on the T_{60}. Addionally, it seems that little to no benefit is obtained by using BF+CNMF instead of CNMF, suggesting that the knowledge of the noise spectral content taken into account in these methods is more beneficial than the application of the beamformer. More importantly, the benefit of CNMF is illustrated by the fact that both CNMF and BF+CNMF perform better than MMSE and BF+MMSE in all considered conditions. On the other hand, by observing the performance as a function of the T_{60} depicted in Fig. 2, it appears that the application of the beamformer is beneficial in the case of a blind estimation of the PSDs, as illustrated by the larger improvement in FWSSNR obtained by BF+MMSE compared to MMSE. Addionally, it appears that the advantage procured by the use of the beamformer is greater in conditions with a larger amount of reverberation.

5. CONCLUSION

This paper proposes a speech enhancement system to be applied when the DOA of the noise source is known and when a recording of a previous realization of the noise signal is available. The proposed system combines a beamformer with CNMF in order to estimate the PSDs of the speech and of the noise to be suppressed and use the resulting estimates to compute a spectral gain to be applied to the STFT of the input signal. The evaluation has been done by applying the proposed system to ego-noise suppression for a robotic platform. It appeared that by using knowledge of the noise to be suppressed, i.e. estimating PSDs using CNMF, the proposed system performed better than a blind spectral enhancement scheme. The application of the beamformer provided little improvement compared to the single-channel application of CNMF. However, the application of a beamformer steered towards the noise source seemed advantageous when combined with a blind estimator, particularly in conditions with larger amount of reverberation.

6. ACKNOWLEDGEMENTS

The research leading to these results has received funding from the EU Seventh Framework Programme projects DREAMS and EcoShopping (grant agreements ITN-GA-2012-316969 and 609180).

7. REFERENCES

[1] E. A. P. Habets, "Speech dereverberation using spectral enhancement," in *Speech Dereverberation*, P. A. Naylor and N. D. Gaubitch, Eds. Springer, 2010.

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

[3] Timo Gerkmann and Richard C. Hendriks, "Unbiased MMSE-based noise power estimation with low complexity and low tracking delay," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 4, pp. 1383 –1393, May 2012.

[4] B. Cauchi, I. Kodrasi, R. Rehr, S. Gerlach, A. Jukić, T. Gerkmann, S. Doclo, and S. Goetze, "Joint dereverberation and noise reduction using beamforming and a single-channel speech-enhancement scheme," in *Proc. REVERB challenge workshop*, Florence, Italy, May 2014.

[5] S. Braun and E.A.P. Habets, "Dereverberation in noisy environments using reference signals and a maximum likelihood estimator," in *Proc. European Signal Processing Conference (EUSIPCO)*, Marrakech, Morocco, Sep 2013.

[6] H.W. Löllmann, H. Barfuss, A. Deleforge, and W. Kellerman, "Challenges in acoustic signal enhancement for human-robot communication," in *Proc ITG Conf on Speech Communication*, Erlangen, Germany, Sept. 2014.

[7] B. Cauchi, , S. Goetze, and S. Doclo, "Reduction of non-stationary noise for a robotic living assistant using sparse non-negative matrix factorization," in *Speech and Multimodal Interaction in Assistive Environments Workshop (SMIAE)*, Jeju, Republic of Korea, July 2012.

[8] N. Mohammadiha, P. Smaragdis, and A. Leijon, "Supervised and unsupervised speech enhancement using nonnegative matrix factorization," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 10, pp. 2140–2151, 2010.

[9] A. Deleforge and W. Kellerman, "Phase-optimized k-svd for signal extraction from underdetermined multichannel sparse mixtures," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Brisbane, Australia, Apr. 2015, vol. 1.

[10] P.A. O'Grady and Pearlmutter B.A., "discovering speech phones using convolutive non-negative matrix factorization with a sparseness constraint," *Neurocomputing*, vol. 72, pp. 88–101, 2008.

[11] Metralabs, "http://www.metralabs.com," 2015.

[12] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Am.*, vol. 65, no. 4, pp. 943–950, Apr. 1979.

[13] C. Jankowski, A. Kalyanswamy, S. Basson, and J. Spitz, "NTIMIT: a phonetically balanced, continuous speech, telephone bandwidth speech database," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, White Plains, NY, USA, 1990, pp. 109–112.

[14] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 32, no. 6, pp. 1109–1121, Dec. 1984.

[15] Y. Hu and P.C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 1, pp. 229–238, 2008.