Interspeech 2018
2-6 September 2018, Hyderabad

Who said that?
A comparative study of non-negative matrix factorization techniques

Teun F. Krikke[1], *Frank Broz*[1], *David Lane*[1]

[1]Heriot-Watt University, Edinburgh, United Kingdom

tfk2@hw.ac.uk, f.broz@hw.ac.uk, d.m.lane@hw.ac.uk

Abstract

In noisy environments it is difficult for a computer to understand what a person is saying, especially when there are multiple speakers. In this paper we concentrate on separating overlapping speech. Non-negative matrix factorisation (NMF) is a method of doing source separation without needing a lot of data. The choice of cost function can have a significant impact on the performance of NMF. We evaluate NMF using three different cost functions (Euclidean, Itakura-Saito and Kullback-Leibler), including modifications using sparsity, convolution or additional information in the form of the direction of arrival. We conduct this evaluation on three different speech corpora. Adding directional information to NMF in the form of non-negative tensor factorisation (NTF) gives us the best result on the map task and vocalization corpora and the Itakura-Saito cost function performs best on the acoustic-camera corpus. In this paper, we show that the Itakura-Saito cost function is the most robust cost function when the recording contains noise. We do this by applying acoustic evaluation measurements.

Index Terms: speech recognition

1. Introduction

Over the past few years home automation has become an important topic. Alongside home automation, the importance of speech recognition has increased, with it currently being used in devices such as Amazon Echo and Google Home. However, there is a big difference between speech recognition on a phone in which the microphone is close to the mouth of the speaker and speech recognition using a device that is placed somewhere in a room in which the microphone picks up background noise as well as speech. The device needs to isolate the speaker and create a sound file that is as clear as possible. It has to differentiate between different speakers that might be speaking at the same time alongside other intrusive noise sources (e.g. an extractor fan in the kitchen which is on during cooking) in order to understand what the main speaker is asking the device.

Blind source separation (BSS) is a process in which there is no prior knowledge of the location of the sources or about the sources themselves in the associated audio file. Different techniques have been applied to this problem, for example, independent component analysis (ICA) [1] and non-negative matrix factorization (NMF) [2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12]. According to Mirsamadi et al. [13], the disadvantage of ICA is that if this technique is combined with direction of arrival (DoA) it cannot be used to solve the permutation problem for high frequencies when the frequency exceeds the spatial aliasing limit.

NMF has been applied successfully on short range speech (less than 5 metres) and premixed audio files [3, 14, 15]. With NMF it is very important to choose the optimal cost function; the Kullback-Leibler cost function is the most popular one to use with NMF. Another cost function is the Itakura-Saito divergence, which has been successfully used for music analysis to separate out different instruments in an audio track [2, 3]. These cost functions are combined with different techniques, for example DoA [15] and convolution, to improve the accuracy of NMF.

By adding directionality to a technique like NMF [15], intensity information about the different sources is provided. Combining the knowledge of the possible source locations with information from multiple microphones allows the algorithm to separate the sources. However, this assumes that the location of the two sources is differentiable, which on a 2D plane is not always the case when sources move around. For example, when the sources are directly behind each other this does not show up on a 2D plane, only in a 3D environment.

NMF is not the only technique applied to BSS - an interesting alternative is deep learning [16, 17, 18]. The main disadvantage of deep learning is that it needs multiple hours of speech data in order to build a mask for the separation of speakers. However, instead of having to train it on every file, which is the case for NMF, the technique is only trained once and after that it can be used in real-time.

The novelty of this paper is that we concentrate on systematically evaluating the effectivenees of separating overlapping speech by using NMF. We chose three corpora (acoustic-camera corpus, map task corpus and vocalization corpus) for this problem. These three corpora have different recording distances (i.e. distance between speaker and microphone) to test the performance of the different NMF techniques when the recording contains noise and to see how the distance between the microphones and the speakers influences the performance of the algorithm.

In section 2, we discuss the data we use for testing the algorithms. This is followed by section 3 in which we discuss our test method and introduce the different NMF techniques that we use. The results of our experiments are presented in section 4, followed by the discussion and conclusion in section 5.

2. Corpora

We use three different corpora for testing the NMF techniques. A concise overview of the different corpora is given in Table 1.

The first corpus that we use is the vocalization corpus[1] [8] which contains recorded telephone conversations of 120 different subjects. The speakers are asked to discuss what they would take into an emergency shelter.

The HCRC MapTask corpus [19] contains recorded speech of a person describing a route on a map. The recordings of this corpus are made using headmounted microphones. During the recordings both participants are in the same room.

Background speech of the second speaker is present in these

[1]http://www.dcs.gla.ac.uk/vincia/?p=378

10.21437/Interspeech.2018-1807

two corpora. This does not provide us with a clean ground truth. These two corpora are not scripted or transcribed.

The third corpus is a small corpus recorded with an acoustic-camera (AC). This device contains 72 microphones which are placed in a circular configuration with one camera in the middle of the circle. The AC gives us an exact location of the microphones and allows us to use beam-forming to get an approximate location of the sources. Beam-forming uses the recordings from all microphones to determine which direction the sound is coming from [20, 21]. The location information of the microphones is also used to calculate the direction of arrival of the signal for the algorithm described by Stein [15]. In a clean environment, the device is able to locate the origin of the sound, but with multiple sound sources it is not able to separate them. The room we use has noise from the air-conditioning along with reverberation due to the room size as is typical of many home and office environments. The high sensitivity of the microphones to noise and echo means that post processing is needed to create a clear approximation of the source location. The AC corpus contains noisy data, which makes it more challenging for algorithms to make a clean separation between the different sources.

The recordings made by the acoustic-camera contain one speaker each. These recordings were made in a room of 9 by 13 metres. The speakers are given a short story to read aloud, which provides us with an easy way to transcribe the speech. The speakers were instructed to stand still for two thirds of the recorded time, after which they should walk around the room keeping a minimal distance of 6 metres away from the camera.

For determing the direction of arrival of the sound, we are using the exact location of the microphones. The first two corpora do not have location information about the placement of the microphones because the recordings are made with head mounted microphones. To overcome this issue, we created it artificially by assuming that there are three microphones each placed one audio frame apart, thus the signal has a time delay of 1 audio frame. This is dependent on the frame rate of the recording. For example, when a recording is made at 16 kHz the microphones would be spaced at $\frac{\text{speed of sound}}{\text{frame rate}}$ metres or in this case $\frac{340.29}{16000}$ metres which is equal to 0.021 metres.

For testing an algorithm, we mix two individual files from each corpus to create a two speaker file giving us a ground truth to test the result against.

3. Method

NMF uses non-negative data for factorisation to separate the sources. It needs to approximate the data (X), which is the squared magnitude information of the recordings. To get this, we take the short-time Fourier transform (STFT) with a window of 30 ms and an overlap of 10 ms. The window size is slightly bigger than what is normally used (25 ms) and should pick up speech better than shorter windows. The amount of overlap is the same as what is normally used for speech recognition. The outcome of the STFT is then squared to remove the negative values. NMF approximates X by multiplying two matrices (W and H) together (see Equation 1). The W matrix is an approximation of the signal coming from the different sources (K) and the H matrix is an approximation of the gain of the different sources. When multiplied together, this gives us the approximated version of X ($\widetilde{X}$). Assuming that the size of X is frequency (F) multiplied by time (N), then the size of the matrix W is F x K and the size of H is K x N. At the end of each iteration, the difference (or cost) between X and $\widetilde{X}$ is calculated (see Equation 2) and the two matrices W and H are updated. In this paper, we concentrate on three different cost functions: the Euclidean distance (see Equation 3), Kullback-Leilbler (KL) divergence (see Equation 4) and Itakura-Saito (IS) divergence (see Equation 5). We chose these three cost functions because of their popularity.

$$X \approx \widetilde{X} = WH \tag{1}$$

$$D(X|\widetilde{X}) = \sum_{f=1}^{F}\sum_{n=1}^{N} d([X]_{fn}|[\widetilde{X}]_{fn}) \tag{2}$$

$$d_{EUC}(x|y) = \frac{1}{2}(x-y)^2 \tag{3}$$

$$d_{KL}(x|y) = xlog\frac{x}{y} - x + y \tag{4}$$

$$d_{IS}(x|y) = \frac{x}{y} - log\frac{x}{y} - 1 \tag{5}$$

We apply eight different techniques to the speaker separation problem (see Table 2). Three of these techniques use the Kullback-Leilbler (KL) divergence, while the others use the Itakura-Saito (IS) divergence or the Euclidean distance. The KL techniques are: convolution KL; sparse KL and direction of arrival (DoA) KL. The eighth technique, non-negative tensor factorisation (NTF), which is very similar to NMF, is described below.

We use a MATLAB library called NMFlib for the different KL and Euclidean techniques[2]. For the IS techniques, we use the implementation given by [2][3] and the implementation for both DoA techniques comes from [15][4]. The three additions to the cost function (sparsity, convolution, DoA) are chosen because of their performance on speech. Combined, they will allow us to compare the performance of the cost function and the performance of the additional functions.

As mentioned, NMF works by updating the W and H matrices. Some techniques, for example sparse KL, modify these update functions to improve the accuracy of the NMF algorithm. The addition of sparsity or convolution provide a different way of separating the sources. The convolution should be able to convolve the different frequencies together to form the original matrix (see Equation 6). Sparsity ensures that the H matrix does not converge towards a solution, instead the H matrix is constantly being slightly modified. Despite the fact that it never converges fully, it still separates the sources. The update rules are given in Equations 8 and 9. The β parameter in these two equations defines which cost function is used. The values for β are given in Table 2 [2]. The λ parameter is added to the update function of the H matrix to ensure the sparsity of this matrix (Equation 9) - sparsity is only enforced when $\lambda > 0$ [22, 23]. We chose the sparsity parameter empirically by running experiments on the corpora with sparsity values between 0.9 and 0.001.

DoA NMF works by changing only the update rule for W and multiplying W with a direction of arrival matrix (D_W, see Equation 7). This last matrix is the same size as W ($F \times K$) and its value are calculated using the least squares method.

$$W = W\frac{((WH)^{\beta-2}\dot{V})\overrightarrow{H}^t}{(WH)^{\beta-1}\overrightarrow{H}^t} \tag{6}$$

[2]https://github.com/audiofilter/nmflib

[3]https://www.irit.fr/ Cedric.Fevotte/extras/neco09/code.zip

[4]https://arxiv.org/src/1411.5010v2/anc

corpus	# subjects	# mics	# files	file length (mm:ss)	ground truth	noise	mic. to source	location	F_s	transcripts
vocalization corpus	120 (63 women, 57 men)	1 (per file)	2763	0:10	No	No	$<$ 1 metre	Lab setting	16kHz	No
MapTask corpus	64 (32 women, 32 men)	1 (per file)	191	5:00	No	No	$<$ 1 metre	Lab setting	16kHz	No
acoustic-camera	16 (12 men, 4 women)	72 (per file)	7	1:30	Yes	Yes	$>$ 6 metres	Empty room	192kHz	Yes

Table 1: *Overview of the different corpora.*

Technique	Cost function	Parameters	
		λ	β
Sparse Euclidean	Euclidean	0.0001	2
Convolution Euclidean	Euclidean	0	2
IS	Itakura-Saito	0	0
Sparse IS	Itakura-Saito	0.0001	0
Convolution IS	Itakura-Saito	0	0
Sparse KL	Kullback-Leibler	0.0001	1
Convolution KL	Kullback-Leibler	0	1
DoA	Kullback-Leibler	0	1

Table 2: *Overview of the different NMF techniques.*

$$W = D_W W \frac{((WH)^{\beta-2}\dot{V})H^T}{(WH)^{\beta-1}H^T} \tag{7}$$

$$W = W \frac{((WH)^{\beta-2}\dot{V})H^T}{(WH)^{\beta-1}H^T} \tag{8}$$

$$H = H \frac{W^T((WH)^{\beta-2}\dot{V})}{W^T(WH)^{\beta-1} + \lambda} \tag{9}$$

Lastly, DoA NTF factorises 3 matrices instead of 2. The third matrix is made up of the information from the DoA [15]. This extra matrix (D_x) describes the direction of the sound [15] (see Equation 10). D_x contains the direction of arrival for the different frequencies over time, giving D_x the size F x N. This changes NMF into non-negative tensor factorisation (NTF). This additional information should improve performance when it is available.

$$X \approx \widetilde{X} = D_x W H \tag{10}$$

For testing the different techniques, we apply 3 objective measurements introduced in [10] namely: signal-to-distortion ratio (SDR); signal-to-interference ratio (SIR) and signal-to-artefact ratio (SAR). Positive values indicate better performance for all measurements. We use the vocalization corpus and MapTask corpus to determine how well each technique performs the separation task. With our own corpus, we measure the performance of the different techniques when there is noise and reverberation in the recording. To deal with this, we apply some preprocessing techniques in the form of noise reduction and a multi-band compressor for reverb reduction. This gives us four different sets of files; one without both reverb and noise, one with only reverb, one with only noise and the original recording containing both.

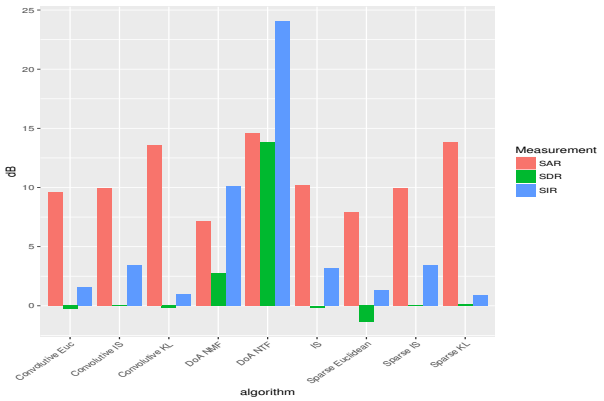

Figure 1: *A comparison between different NTF and NMF techniques on the vocalization corpus.*

4. Results

As mentioned in Section 3, we used the output of the STFT as input for the NMF algorithms. These algorithms have two fixed parameters (F and K), while parameter N depends on the length of the file. For F, we used 513 frequency bins (this value is empirically chosen) and set K to be the desired number of speakers, in our case 2. We stopped the algorithms after 1000 iterations, by which time the cost function has converged.

The results show that all techniques have a good SAR ratio on the vocalization corpus and the MapTask corpus, meaning not many artefacts are introduced (see Figures 1 and 2). The SDR and SIR ratios are poor for all techniques except for NTF. This shows that NTF is able to remove both speakers from a single file more clearly than the other techniques. However, the combined speech files have no noise or reverb which in the real world is rarely the case, except from telephone conversations.

We applied the different NMF techniques to compare the performance with and without noise and reverb on the acoustic camera corpus (see Figure 3). The IS cost function performs the best when noise is removed (see Figure 3B). The same applies when we remove only the reverb (see Figure 3C) and when we compare to the original (non post-processed) files (see Figure 3D). On removing both noise and reverb (see Figure 3A) the sound gets distorted to an extent that the KL and Euclidean versions of NMF out perform the IS and DoA versions.

Comparing the results of all techniques on the different versions of the acoustic camera corpus, when we remove the re-

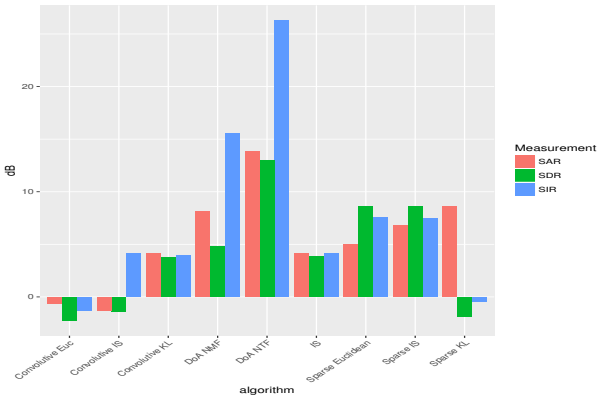

Figure 2: *A comparison between different NTF and NMF techniques on the MapTask corpus.*

verb all techniques show an improvement in the SAR and SDR values but get lower SIR values. The Euclidean cost function has the greatest improvement compared to the rest of the techniques. However, over all three corpora, the Euclidean cost function has the lowest SAR of all the techniques and is therefore the worst performing technique on our corpora. Both the sparse and convolution techniques work better on the noisy version of the acoustic camera corpus. With this version, we see more positive values. However, all the algorithms are out performed by the IS cost function without the use of sparsity or convolution on the reverberant speech.

5. Discussion and Conclusion

In this paper we have compared different NMF techniques to determine which performs best in a natural environment (where there is noise and reverb). When comparing the results, we see that NTF works best on the different corpora. Therefore, when we show NTF where the source is most likely coming from, the technique is better at separating the sources compared to when either NMF needs to determine this information by convolution or the DoA information is not used at all. When we look at the IS cost function, we see a smaller decrease in performance when the recordings contain more noise. This improvement is clearly visible when we compare the result to a version that does not contain noise or reverb. In general, all techniques improve when the noise is removed from the acoustic camera recordings. Comparing different versions of the corpus also shows that removing reverb and noise is not always good for separating sources. The performance of the IS cost function decreases when both reverb and noise are removed.

The poor performance of the DoA technique on the AC corpus could be caused by the difference in microphone distance. Stein [15] assumes a distance of one sample between the microphones, whereas we have an exact distance and are not using this relative distance for the calculation of the angles as it is described in [15]. This could explain why the algorithm performs worse on our data.

6. Acknowledgements

This research is supported by the EPSRC, as part of the CDT in RAS at Heriot-Watt University and The University of Edinburgh. Grant reference EP/L016834/1.

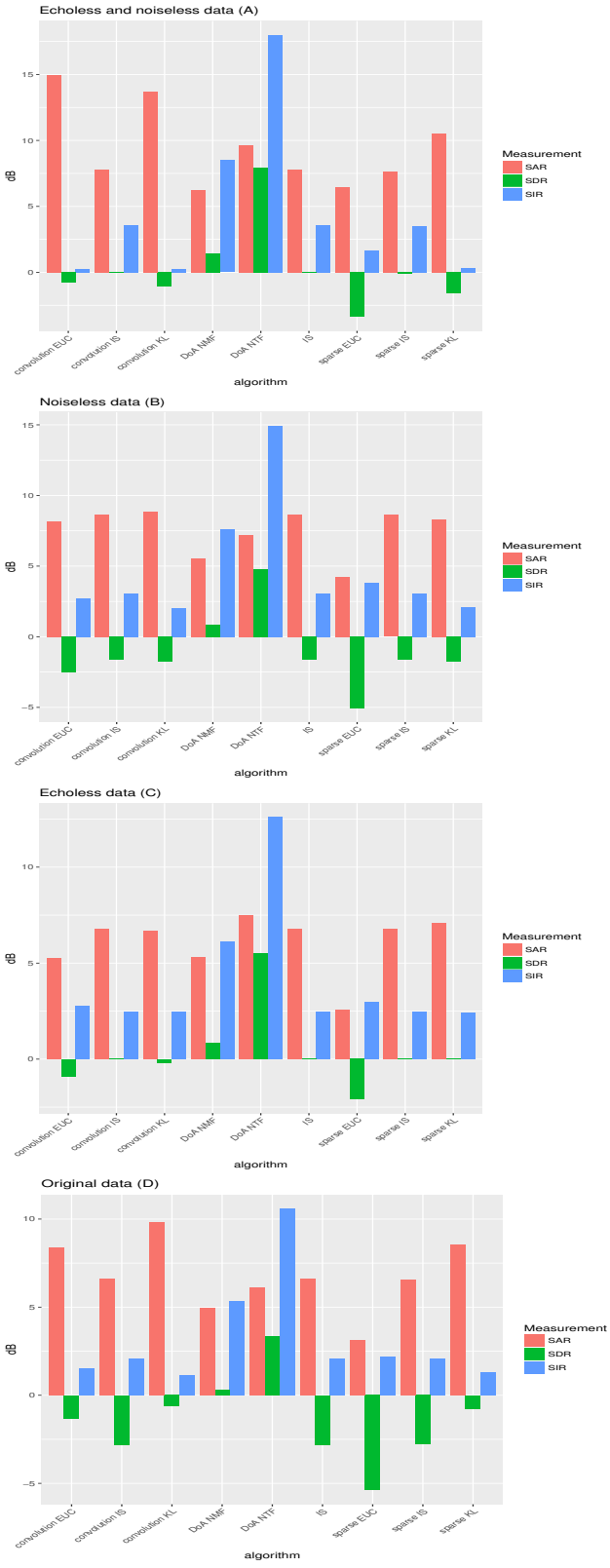

Figure 3: *A comparison between different NMF techniques on the noiseless and echoless (A), reverberant (B), noisy (C) and original (D) AC corpus recordings.*

7. References

[1] M. E. Davies and C. J. James, "Source separation using single channel ica," *Signal Processing*, vol. 87, no. 8, pp. 1819–1832, 2007.

[2] C. Févotte, N. Bertin, and J.-L. Durrieu, "Nonnegative matrix factorization with the itakura-saito divergence: With application to music analysis," *Neural computation*, vol. 21, no. 3, pp. 793–830, 2009.

[3] C. Févotte, E. Vincent, and A. Ozerov, "Single-channel audio source separation with nmf: divergences, constraints and algorithms," in *Audio Source Separation*. Springer, 2018, pp. 1–24.

[4] E. M. Grais and H. Erdogan, "Single channel speech music separation using nonnegative matrix factorization and spectral masks," in *17th International Conference on Digital Signal Processing (DSP), 2011.* IEEE, 2011, pp. 1–6.

[5] T. G. Kang, K. Kwon, J. W. Shin, and N. S. Kim, "NMF-based target source separation using deep neural network," *IEEE Signal Processing Letters*, vol. 22, no. 2, pp. 229–233, 2015.

[6] A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 3, pp. 550–563, 2010.

[7] P. Parathai, W. L. Woo, S. Dlay, and B. Gao, "Single-channel blind separation using l 1-sparse complex non-negative matrix factorization for acoustic signals," *The Journal of the Acoustical Society of America*, vol. 137, no. 1, pp. EL124–EL129, 2015.

[8] H. Salamin, A. Polychroniou, and A. Vinciarelli, "Automatic detection of laughter and fillers in spontaneous mobile phone conversations," in *IEEE International Conference on Systems, Man, and Cybernetics (SMC), 2013.* IEEE, 2013, pp. 4282–4287.

[9] J. Traa, P. Smaragdis, N. D. Stein, and D. Wingate, "Directional NMF for joint source localization and separation," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), 2015.* IEEE, 2015, pp. 1–5.

[10] 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.

[11] Z. Wang and F. Sha, "Discriminative non-negative matrix factorization for single-channel speech separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2014.* IEEE, 2014, pp. 3749–3753.

[12] F. Weninger, J. Le Roux, J. R. Hershey, and S. Watanabe, "Discriminative NMF and its application to single-channel source separation." in *INTERSPEECH*, 2014, pp. 865–869.

[13] S. Mirsamadi and J. H. Hansen, "Multichannel speech dereverberation based on convolutive nonnegative tensor factorization for asr applications." in *INTERSPEECH*, 2014, pp. 2828–2832.

[14] J. Nikunen and T. Virtanen, "Direction of arrival based spatial covariance model for blind sound source separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 3, pp. 727–739, 2014.

[15] N. D. Stein, "Nonnegative tensor factorization for directional blind audio source separation," *stat*, vol. 1050, p. 18, 2014.

[16] G.-X. Wang, C.-C. Hsu, and J.-T. Chien, "Discriminative deep recurrent neural networks for monaural speech separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2016.* IEEE, 2016, pp. 2544–2548.

[17] Y. Yu, W. Wang, J. Luo, and P. Feng, "Localization based stereo speech separation using deep networks," in *IEEE International Conference on Digital Signal Processing (DSP), 2015.* IEEE, 2015, pp. 153–157.

[18] P.-S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Deep learning for monaural speech separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2014.* IEEE, 2014, pp. 1562–1566.

[19] A. H. Anderson, M. Bader, E. G. Bard, E. Boyle, G. Doherty, S. Garrod, S. Isard, J. Kowtko, J. McAllister, J. Miller *et al.*, "The hcrc map task corpus," *Language and speech*, vol. 34, no. 4, pp. 351–366, 1991.

[20] D. Döbler and G. Heilmann, "Time-domain beamforming using zero-padding," in *Berlin Beamforming Conference (BeBeC)*, 2008.

[21] R. Schröder and O. Jaeckel, "Evaluation of beamforming systems," in *Proceedings of the 4th Berlin Beamforming Conference*, 2012, pp. 22–23.

[22] M. N. Schmidt, "Speech separation using non-negative features and sparse non-negative matrix factorization," *Elsevier*, 2007.

[23] J. Eggert and E. Korner, "Sparse coding and NMF," in *IEEE International Joint Conference on Neural Networks, 2004.*, vol. 4. IEEE, 2004, pp. 2529–2533.