Reduction of Non-stationary Noise for a Robotic Living Assistant using Sparse Non-negative Matrix Factorization

Benjamin Cauchi [1] **Stefan Goetze** [1] **Simon Doclo** [1, 2]

[1] Fraunhofer IDMT, Hearing, Speech and Audio Technology, 26129 Oldenburg, Germany
firstname.surname@idmt.fraunhofer.de

[2] University of Oldenburg, 26129 Oldenburg, Germany
simon.doclo@uni-oldenburg.de

Abstract

Due to the demographic changes, support by means of assistive systems will become inevitable for home care and in nursing homes. Robot systems are promising solutions if the value is obvious to the care personnel and patients. Natural and intuitive human-machine interfaces are an essential feature to achieve acceptance of the users. Therefore, automatic speech recognition (ASR) is a promising modality for such assistive devices. However, noises produced during movement of robots can degrade the ASR performances. This work focuses on noise reduction by a non-negative matrix factorization (NMF) approach to efficiently suppress non stationary noise produced by the sensors of an assisting robot system.

1 Introduction

The amount of older people in today's societies constantly grows due to demographic changes (Commission, 2007). Technical systems become more and more common to support for routine tasks of care givers or to assist older persons living alone in their home environments (Van Den Broek et al., 2010). Various technical assistive systems have been developed recently (Lisetti et al., 2003), ranging from reminder systems (Boll et al., 2010; Goetze et al., 2010) to assisting robots (Chew et al., 2010). If robot systems are supposed to navigate autonomously they usually rely on vision sensors (Aragon-Camarasa et al., 2010) or acoustic sensors (Yamakawa et al., 2011; Youssef et al., 2010). Acoustic signals are usually picked up by microphones mounted on the robot. In real-world scenarios not only the desired signal part is picked up by these microphones as presented in Figure 1.

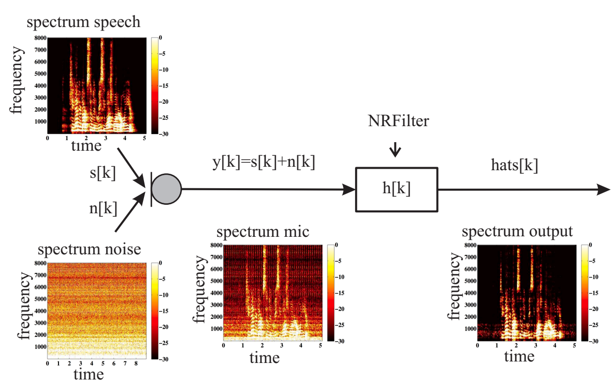

Figure 1: General denoising scheme

The desired signal part is usually superposed with disturbing noise originating from the environment or the robot system itself. This disturbance has to be removed from the microphone signal before it can be further processed, e.g. for navigation, position estimation, acoustic event detection, speaker detection or automatic speech recognition. This contribution focuses on acoustic input for a robot system and more specifically on the noise reduction preprocessing which is needed to clean up noisy sound signals.

Automatic speech recognition (Rabiner and Juang, 1993; Wölfel et al., 2009) is a convenient way to interact with robot assistants since speech is the most natural form of communication. However, to ensure acceptance of speech recognition systems a sufficiently high recognition rate has to be archieved (Pfister and Kaufmann, 2008). Todays speech recognition systems succeed in achieving this recognition rate for environments with low amount of noise and reverberation. Unfortunately, while moving, robots can produce noise degrading the reliability of the ASR.

This work focuses on a specific application, suppressing the non stationary noise produced by the ultra-sonic sensors of a robotic assistant while moving. Please note that although in theory ultrasonic sensors do not produce sound disturbances in the audible range, artefacts due to the fast activation and deactivation of the sensors are present in the audible range and are clearly perceivable as a disturbance in the picked up microphone signal as shown later in Figure 6.

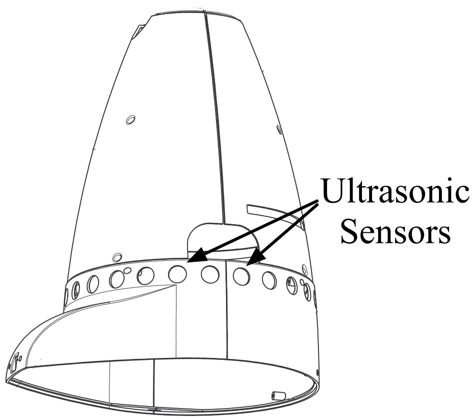

Figure 2: Lower part of the robot with ultrasonic sensors.

Non-negative Matrix Factorization (NMF) is an approach introduced by Lee & Seung (Lee and Seung, 2001) in which the data is described as the product of a set of basis and of a set of activation coefficients both being non-negative. We will apply the NMF approach to remove the disturbances caused by the ultrasonic sensors from the microphone input signal in the following. NMF and its various extensions have been proven efficient in sources separation (Cichocki et al., 2004; Virtanen, 2007), supervised detection of acoustic events (Cotton and Ellis, 2011) or to wind noise reduction (Schmidt et al., 2007). As the NMF algorithm can be fed with prior information about the content to identify, it is an handy way to suppress the non stationary noise produced by the sensors of the considered robotic assistant.

The remainder of this paper is organized as follows: The general NMF algorithm is presented in Section 2 and the proposed denoising method is described in Section 3. An experiment using the TIMIT (Zue et al., 1990) speech corpus is presented in Section 4 and finally the performances are evaluated in terms of achieved signal enhancement in Section 5 before Section 6 concludes the paper.

2 Sparse Non-negative Matrix Factorization

2.1 NMF algorithm

NMF is a low-rank approximation technique for multivariate data decomposition. Given a real valued non-negative matrix **V** of size $n \times m$ and a positive integer $r < min(n, m)$, it aims to find a factorization of **V** into an $n \times r$ real matrix **W** and an $r \times m$ real matrix **H** such that:

$$\mathbf{V} \approx \mathbf{W} \cdot \mathbf{H} \tag{1}$$

The multivariate data to decompose is stacked into **V**, whose columns represent the different observations, and whose rows represent the different variables. In the case of information extraction from audio files, **V** could be the amplitude of the spectrogram and therefore, **W** would be a basis of spectral features when **H** would represent the levels of activation of each of those features along time. The rank r of the factorization corresponds to the number of elements present in the dictionary **W**, and thereof, to the number of rows within **H**.

NMF is an iterative process that can be fed with information about the contents to extract. As an illustration of this ability, an artificial spectrogram of a mixture of two chords, C and D, has been created. Figure 3 shows the initialization of the NMF algorithm. **V** is the spectrogram of the mixture in which the two chords contain only notes' fundamentals and overlap each other. The Algorithm is fed with the spectral content of the C chord.

Figure 4 shows that during the iterative process, the elements of **W** corresponding to the C chord remained unchanged while the other elements of **W** have been updated to fit the spectral content of the D chord. The output time activations within **H** correspond to the presence of both chords within the matrix $\mathbf{V}$.

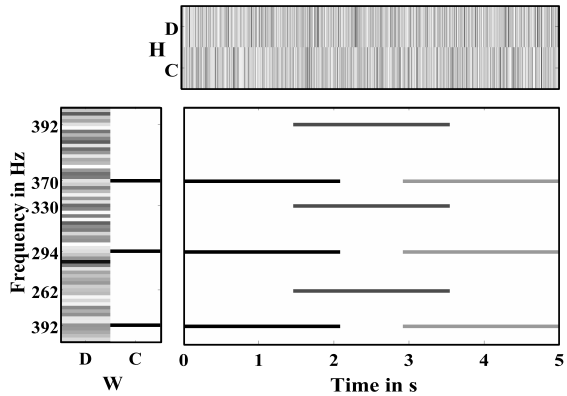

Figure 3: Illustration of the initialization of the NMF algorithm. The spectral content of the C chord is input into $\mathbf{W}$ while the other element of dictionary and activation coefficients in $\mathbf{H}$ are randomly initialized.

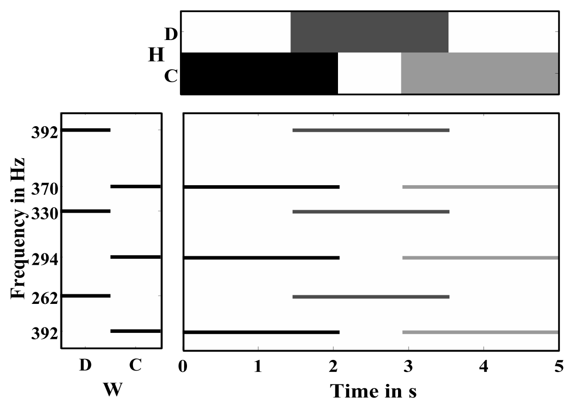

Figure 4: Illustration of the output of the NMF algorithm. The spectral content of the D chord has been learned while the updated $\mathbf{H}$ corresponds to the activations of the chords C and D along time.

2.2 Sparseness Constraint

The very definition of sparseness (or sparsity) is that a vector is sparse when most of its elements are zero. In its application to NMF, the addition of a sparseness constraint λ permits to trade off between the fitness of the factorization and the sparseness of $\mathbf{H}$.

At each iteration, the process aims at reducing a cost function $\mathcal{C}$. In this paper, a generalized version of the Kullback Leibler divergence is used as cost function:

$$\mathcal{D}(\mathbf{V}, \mathbf{WH}) = \left\| \mathbf{V} \otimes \log \frac{\mathbf{V}}{\mathbf{W} \cdot \mathbf{H}} - \mathbf{V} + \mathbf{W} \cdot \mathbf{H} \right\| \quad (2)$$

In 2 the multiplication $\otimes$ and the division are element-wise. The sparseness constraint results in the new cost function:

$$\mathcal{C}(\mathbf{V}, \mathbf{WH}) = \mathcal{D}(\mathbf{V}, \mathbf{WH}) + \lambda \sum_{ij} \mathbf{H}_{ij} \quad (3)$$

The norm of each of the objects within $\mathbf{W}$ is fixed to unity.

3 Supervised NMF denoising

3.1 Method overview

The method is supervised in the sense that it uses a noise dictionary $\mathbf{W}_n$ built from a recording of the known noise to be reduced. The noise spectrogram Φ_n, *i.e.* the short-term fourier transform (STFT), is computed using a hamming window of 32ms and a 50% overlap. The magnitude $\mathbf{V}_n$ of Φ_n is input to the NMF algorithm with a sparseness constraint λ and an order r_n, providing the noise dictionary of r_n spectral vectors. The spectrogram $\mathbf{V}_s$ of the noisy speech is then input to the NMF algorithm along with $\mathbf{W}_n$ in order to obtain the denoised speech spectrogram.

3.2 Separation of the speech signal

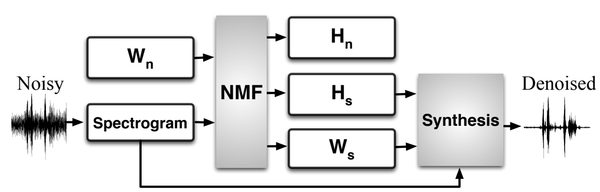

Figure 5: Overview of the NMF based denoising.

The denoising is summarized in Figure 5. The spectrum Φ_s of the noisy speech and its amplitude $\mathbf{V}_s$ are computed as in Section 3.1. $\mathbf{V}_s$ is input to the NMF algorithm along with $\mathbf{W}_n$. The order of factorization r is equal to $r_n + r_s$, r_s being the number of spectral vector used in the speech dictionary $\mathbf{W}_s$. Different sparseness constraint λ_n and λ_s can be applied to the activation matrices $\mathbf{H}_n$ and $\mathbf{H}_s$.

$$\begin{array}{lll} \text{Given} & \mathbf{V} \in \mathbb{R}_+^{n \times m}, r \in \mathbb{N}^* \; s.t. \; r < min(n,m) & \\ \text{minimize} & \mathcal{C}(\mathbf{V}, \mathbf{WH}) \; w.r.t. \; \mathbf{W}, \mathbf{H} & (4) \\ \text{subject to} & \mathbf{W} \in \mathbb{R}_+^{n \times r}, \; \mathbf{H} \in \mathbb{R}_+^{r \times m} & \end{array}$$

The update rules on $\mathbf{W}$ and $\mathbf{H}$ can be expressed as multiplicative updates:

$$\mathbf{W}_s \leftarrow \mathbf{W}_s \otimes \frac{\frac{\mathbf{V}}{\mathbf{W}_s\mathbf{H}_s} \cdot \mathbf{H}_s^T}{\mathbf{1} \cdot \mathbf{H}_s^T} \qquad \mathbf{H} \leftarrow \mathbf{H} \otimes \frac{\mathbf{W}^T \cdot \frac{\mathbf{V}}{\mathbf{WH}}}{\mathbf{W}^T \cdot \mathbf{1}} \qquad (5)$$

The NMF algorithm provides thereof $\mathbf{W}_s$ and $\mathbf{H}_s$ to be used to approximate the spectrogram of the denoised speech. Therefore, $\times$ being the matrix product:

$$\tilde{\mathbf{V}}_s = \mathbf{W}_s \times \mathbf{H}_s \qquad \tilde{\mathbf{V}}_n = \mathbf{W}_n \times \mathbf{H}_n$$
$$\tilde{\mathbf{S}}_s = \Phi_s \otimes \frac{\tilde{\mathbf{V}}_s}{\tilde{\mathbf{V}}_s + \tilde{\mathbf{V}}_n} \qquad (6)$$

The denoised speech signal is finally obtained by applying ISTFT on the spectrogram $\tilde{\mathbf{S}}_s$. The interested reader is referred to (O'Grady and Pearlmutter, 2006) for a more detailed discussion of the needed derivations for Eqs. (5)-(6).

4 Experiment

4.1 Context

The robot platform Scitos A5[1] can be used as a robotic assistant for elderly care. Its built-in microphones allow to interact with the robot using if their signal is analysed by an ASR system. However, while in motion, the robot uses ultrasonic sensors (c.f. Figure 2) to detect potential obstacles. Their constant activation and deactivation produces artifacts that can sever the ASR reliability. The following experiment aims to evaluate the efficiency of the denoising method proposed in Section 3 on speech signals corrupted by this specific sensors noise. The Figure 6 examplarily presents the spectrogram of a corrupted speech signal.

4.2 Protocol

The noise produced by the sensors and the room impulse response (RIR) have been recorded in a quiet office room using the robot's microphone. The test data has been built from the test portion of the TIMIT corpus (Zue et al., 1990). The clean speech files have been built conctenating a silent period of 0.5 seconds in their beginning, to allow for comparison with VAD based methods, and convolving it with the measured RIR. From those prepared clean files, noisy corpora have been built by adding the recorded sensors noise with a SNR set to 10, 5, 0 and -5 dB.

When applying the NMF algorithm the cost function (3) has been used but no stop criterion has been set and a fixed number of 25 iterations has been run. $\mathbf{W}_n$ has been built by applying the NMF algorithm with $r_n = 64$ and $\lambda = 0$ to a 10 seconds noise recording. When applying the algorithm to the speech samples denoising, r has been set to 128. A

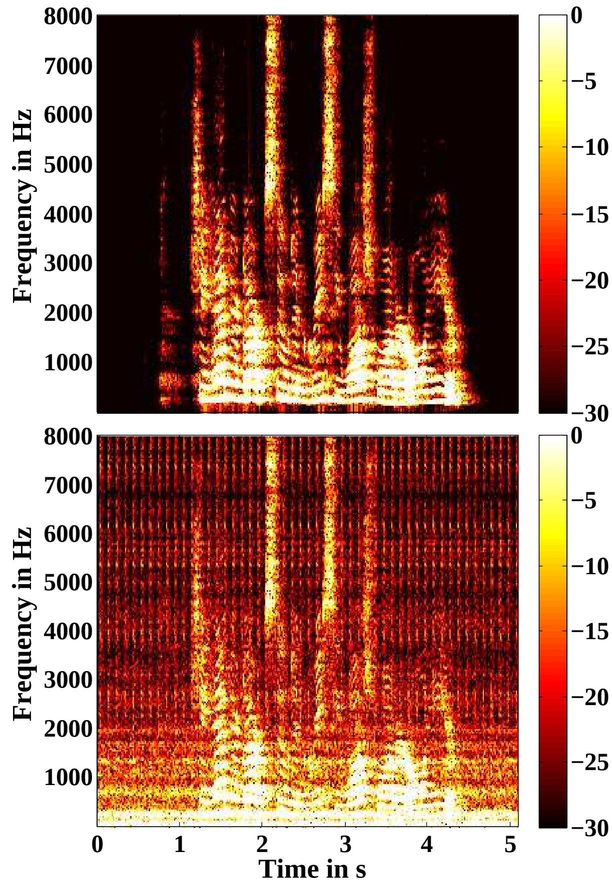

Figure 6: Spectrogram of a speech sentence from the TIMIT corpora: «*She had your dark suit in greasy wash water all year.*», clean (top) and with added sensors noise at SNR=10dB.

[1] http://metralabs.com/

different sparseness constraint has been applied to $\mathbf{H}_n$ and $\mathbf{H}_s$ with $\lambda_n = 0$ and $\lambda_s = 0.2$.

As a reference, the noisy sound samples have as well been processed using a state-of-the-art single-channel noise reduction scheme, i.e. the decision-directed approach according to (Ephraim and Malah, 1985) based on two different noise estimation schemes, i.e. the minimum statistics approach (MS) as described in (Martin, 2001) and the minimum mean square error (MMSE) approach according to (Gerkmann and Hendriks, 2011).

5 Results

The achieved denoising is evaluated with the SNR of the denoised samples and with the noise reduction (NR) as described in (Loizou, 2007). For both scores, the presented values are the mean of the achieved scores on all tested speech samples and the standard deviation along the corpus. The results are presented in Figure 7 for varying input SNR and spectrograms of a denoised speech sample using the three methods is shown in Figure 8. It appears that the NMF based method provides better results, both in term of signal enhancement and of reliability.

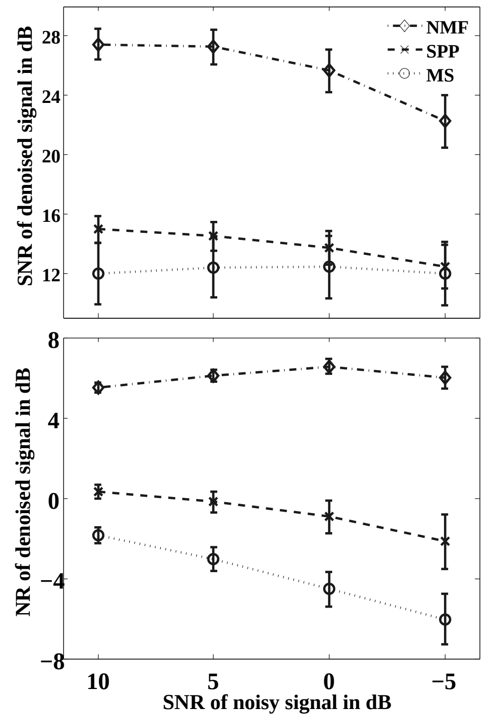

Figure 7: Mean and standard deviation of the achieved SNR and NR for the three tested methods and for different noise levels (SNR).

6 Conclusion

A NMF based method to enhance speech signal when provided with spectral knowledge of the noise has been presented. This method has been applied to the reduction of the non stationary noise produced by the sensors of a robotic assistant. When tested on a corpus of speech signals, the proposed method achieved better performances than well known VAD based denoising.

Further works would include fine tuning of the method, such as determining the optimal number of iterations to obtain the best trade off between enhancement and computing cost, as well as the use of spectro temporal patches as elements of dictionary.

7 Acknowledgement

This work is part of the ALIAS project, funded by the Ambient Assisted Living (AAL) Joint Programme, by the German BMBF, the French ANR and the Austrian BMVIT (see `http://www.aal-europe.eu` for details).

References

G. Aragon-Camarasa, H. Fattah, and J. Paul Siebert. 2010. Towards a unified visual framework in a binocular active robot vision system. *Robotics and Autonomous Systems*, 58(3):276–286.

S. Boll, W. Heuten, E.M. Meyer, and M. Meis. 2010. Development of a multimodal reminder system for older persons in their residential home. *Informatics for Health and Social Care*, 35(3-4):104–124.

S. Chew, W. Tay, D. Smit, and C. Bartneck. 2010. Do social robots walk or roll? *Social Robotics*, pages 355–361.

A. Cichocki, R. Zdunek, and S. Amari. 2004. New algorithms for non-negative matrix factorization in applications to blind source separation. In *Proc. of Acoustics, Speech and Signal Processing, 2006. ICASSP 2006.*, volume 5, pages V–V. IEEE.

European Commission. 2007. Europe's demographic future: facts and figures. *Brussels: European Commission.*

C.V. Cotton and D.P.W. Ellis. 2011. Spectral vs. spectro-temporal features for acoustic event detection. In *Proc. of Applications of Signal Processing to Audio and Acoustics (WASPAA), 2011 IEEE Workshop on*, pages 69 –72, oct.

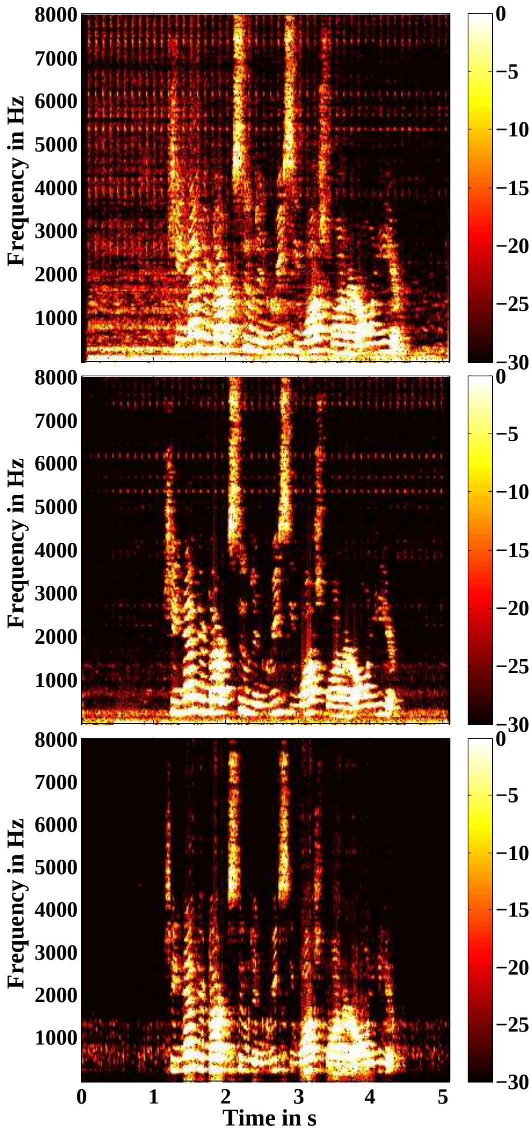

Figure 8: Spectrogram of a denoised signal using the three different methods, MS (top), MMSE (middle) and NMF.

Y. Ephraim and D. Malah. 1985. Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 33(2):443–445.

T. Gerkmann and R.C. Hendriks. 2011. Noise power estimation based on the probability of speech presence. In *Proc. of Applications of Signal Processing to Audio and Acoustics (WASPAA), 2011 IEEE Workshop on*, pages 145–148. IEEE.

S. Goetze, N. Moritz, J.E. Appell, M. Meis, C. Bartsch, and J. Bitzer. 2010. Acoustic user interfaces for ambient-assisted living technologies. *Informatics for Health and Social Care*, 35(3):125.

D.D. Lee and H.S. Seung. 2001. Algorithms for non-negative matrix factorization. *Advances in neural information processing systems*, 13.

C. Lisetti, F. Nasoz, C. LeRouge, O. Ozyer, and K. Alvarez. 2003. Developing multimodal intelligent affective interfaces for tele-home health care. *International Journal of Human-Computer Studies*, 59(1):245–255.

P. Loizou. 2007. Speech enhancement: Theory and practice (crc, boca raton, fl:).

R. Martin. 2001. Noise power spectral density estimation based on optimal smoothing and minimum statistics. *Speech and Audio Processing, IEEE Transactions on*, 9(5):504–512.

P.D. O'Grady and B.A. Pearlmutter. 2006. Convolutive non-negative matrix factorisation with a sparseness constraint. In *Proc. of the 2006 16th IEEE Signal Processing Society Workshop on Machine Learning for Signal Processing*, pages 427–432.

B. Pfister and T. Kaufmann. 2008. *Speech Processing: Basics and Methods for Speech Synthesis and Speech Recognition (In German: Sprachverarbeitung: Grundlagen und Methoden der Sprachsynthese und Spracherkennung)*. Springer Publishing Company, Incorporated.

L. Rabiner and B.H. Juang. 1993. *Fundamentals of speech recognition*. Prentice hall.

M.N. Schmidt, J. Larsen, and F.T. Hsiao. 2007. Wind noise reduction using non-negative sparse coding. In *Proc. of 2007 IEEE Workshop on Machine Learning for Signal Processing,*, pages 431–436.

G. Van Den Broek, F. Cavallo, and C. Wehrmann. 2010. *AALIANCE ambient assisted living roadmap*. Ios Press.

T. Virtanen. 2007. Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(3):1066–1074.

M. Wölfel, J.W. McDonough, and Inc Ebrary. 2009. *Distant speech recognition*. Wiley Online Library.

N. Yamakawa, T. Takahashi, T. Kitahara, T. Ogata, and H. Okuno. 2011. Environmental sound recognition for robot audition using matching-pursuit. *Modern Approaches in Applied Intelligence*, pages 1–10.

K. Youssef, S. Argentieri, and J.L. Zarader. 2010. Binaural speaker recognition for humanoid robots. In *Proc. of 2010 11th International Conference on Control Automation Robotics & Vision (ICARCV)*, pages 2295–2300.

V. Zue, S. Seneff, and J. Glass. 1990. Speech database development at mit: Timit and beyond. *Speech Communication*, 9(4):351–356.