INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Speaker-dependent Dictionary-based Speech Enhancement for Text-Dependent Speaker Verification

Nicolai Bæk Thomsen, Dennis Alexander Lehmann Thomsen, Zheng-Hua Tan, Børge Lindberg, Søren Holdt Jensen

Department of Electronic Systems, Aalborg University, Denmark
{nit,dalth,zt,bli,shj}@es.aau.dk

Abstract

The problem of text-dependent speaker verification under noisy conditions is becoming ever more relevant, due to increased usage for authentication in real-world applications. Classical methods for noise reduction such as spectral subtraction and Wiener filtering introduce distortion and do not perform well in this setting. In this work we compare the performance of different noise reduction methods under different noise conditions in terms of speaker verification when the text is known and the system is trained on clean data (mis-matched conditions). We furthermore propose a new approach based on dictionary-based noise reduction and compare it to the baseline methods.
Index Terms: Noise reduction, speaker verification, dictionary learning

1. Introduction

Speaker verification (SV) is becoming more and more relevant in commercial products as it allows for hands-free authentication. In this setting a speaker claims to have identity X and speaks an utterance and it is up to the system either to accept or reject the claim based on the recorded speech. The setup can be further strengthened by requiring a fixed passphrase from the speaker, which should be the same at enrollment and test time. One problem with using speech for authentication is the corruption of the recorded speech due to additive noise. Depending on the setting the noise can be different and have different characteristics, which will in general degrade the performance of SV.

One way of dealing with noise, is to apply pre-processing before submitting it to feature extraction. There are many speech enhancement and noise reduction algorithms for this purpose, such as Wiener filtering, Spectral Substraction and minimum mean-square estimation, see e.g.,[1, 2, 3] and the references in [4]. These are all derived from signal-level metrics, which do not guarantee good performance when used as front-ends for SV. Sadjadi et. al. [5] investigate the performance of speech enhancement techniques for speaker identification under noisy and mis-matched conditions and conclude that for low signal-to-noise ratios (0, 5 and 10 dB), there is generally a performance gain, but that enhancement can degrade performance at higher SNR and perform worse than using the noisy signal directly. The authors furthermore propose to use mean Hilbert envelope coefficients (MHEC) as features and show that they outperform conventional mel-frequency cepstrum coefficients (MFCC). In [6] different front-ends for speaker identification were compared and it was found that noise reduction based on minimum statistics noise estimation (MSNE) did not improve performance compared to no processing., but it helped when fused with other front-ends.

A new method for noise suppression or speech enhancement, denoted sparse coding or exemplar-based sparse representation, has attracted a lot of interest within the last decade [7]. The basic idea is to learn dictionaries of either speech, noise or both and then at test time model the noisy speech as a linear combination of exemplars from the speech and noise dictionary. The clean speech can thus be estimated by only retaining the contribution from the speech dictionary. It has been demonstrated to perform well for different types of noise [8, 9]. The advantage of this method for speaker verification is that the enhancement is based on previously seen speech from the speaker, thus it is not likely to introduce distortions as may be the case for standard approaches. The disadvantage of this method is the need for training data both for speech and noise. However given that most systems have a voice activity detector (VAD), it becomes possible to update the noise dictionary using the recordings of non-speech segments as done in [10].

In this paper we compare the performance of different well-known enhancement algorithms for the task of speaker verification when the back-end is trained on clean data, i.e., mismatched conditions. This is similar to [5], but here we consider text-dependent speaker verification instead of speaker recognition and the evaluation is done on the recent RSR2015 database. We furthermore propose to apply sparse coding to the task of front-end speech enhancement for text-dependent speaker verification. For this task a few enrollment recordings are available for each speaker, which can be used to train a speaker- and utterance dependent dictionary/codebook. Because the text to be spoken at training and test time are the same, the speech dictionary can model this reduced signal space very well with a limited number of entries/exemplars. At test time the dictionary corresponding to the claimed identity is then used to enhance the speech.

In Sec. 2 the problem is formally defined and in Sec. 3 we briefly introduce the concept of sparse coding for speech enhancement and how we apply it in this task. Finally, the experimental results are reported in Sec. 4 and a conclusion on the work along with directions for future work are given in Sec. 5.

2. Problem Formulation

In this work we consider a single-channel signal model with additive noise which is un-correlated with the speech of interest. The signal models in time- and STFT domain are thus respectively given by

$$x(n) = s(n) + w(n) \tag{1a}$$
$$X(k,f) = S(k,f) + W(k,f) \tag{1b}$$

http://dx.doi.org/10.21437/Interspeech.2016-763

where $x(n)$, $s(n)$ and $w(n)$ are the recorded signal, the speech signal and the noise in the time domain, respectively, and $X(k,f)$, $S(k,f)$ and $W(k,f)$ are the corresponding STFT domain signals with k and f denoting the frame index and the frequency index, respectively.

3. Methods

3.1. Non-negative Sparse Coding (NNSC)

Given a noisy recording of the speech we can compute the magnitude short-time Fourier-transform (STFT), $|\boldsymbol{X}|$, and represent it as a linear combination of columns from a speech dictionary in the following way

$$|\boldsymbol{X}| \approx \boldsymbol{D}\boldsymbol{H} \tag{2}$$

where $\boldsymbol{D}$ is a dictionary and $\boldsymbol{H}$ is the activation matrix. For the rest of the paper all STFT are magnitude unless stated otherwise. The decomposition can be achieved by applying non-negative matrix factorization (NMF), which solves the following optimization problem

$$\begin{aligned} &\underset{\boldsymbol{D},\boldsymbol{H}}{\text{minimize}}\ \frac{1}{2}||\boldsymbol{X}-\boldsymbol{D}\boldsymbol{H}||_F^2 + \alpha\cdot||\boldsymbol{H}||_1 \\ &\text{s.t. } \boldsymbol{D},\boldsymbol{H} \geq 0 \end{aligned} \tag{3}$$

where $||\cdot||_F$ is the Frobenius norm, $||\cdot||_1$ is the ℓ_1-norm, α is a small positive number and $\geq$ means that the elements of $\boldsymbol{D}$ and $\boldsymbol{H}$ must be non-negative. The optimization problem is solved using a multiplicative gradient method which alternates between updating $\boldsymbol{H}$ and $\boldsymbol{D}$, and is given by [11],

$$\boldsymbol{H} = \boldsymbol{H} \circ \frac{\boldsymbol{D}^T\boldsymbol{X}}{\boldsymbol{D}^T\boldsymbol{D}\boldsymbol{H}+\alpha} \tag{4a}$$

$$\boldsymbol{D} = \boldsymbol{D} \circ \frac{\boldsymbol{X}\boldsymbol{H}^T}{\boldsymbol{D}\boldsymbol{H}\boldsymbol{H}^T} \tag{4b}$$

where $\circ$ denotes the Hadamard product (element-wise). It is noted that other cost functions can be used as well [11] and that the original form does not contain the sparsity-inducing term on $\boldsymbol{H}$ in (3). However, it has been demonstrated to be beneficial for modeling speech [9]. Normalization was applied to the columns of $\boldsymbol{D}$ after every iteration to accomodate the scaling ambiguity, which is inherent to the decomposition.

3.2. NNSC for noise reduction

Under the assumption of additive noise, we can further decompose the dictionary and activation matrix into a speech part and a noise part,

$$\begin{aligned} \boldsymbol{X} &\approx \boldsymbol{D}_s\boldsymbol{H}_s + \boldsymbol{D}_n\boldsymbol{H}_n \\ &= \begin{bmatrix}\boldsymbol{D}_s & \boldsymbol{D}_n\end{bmatrix}\begin{bmatrix}\boldsymbol{H}_s \\ \boldsymbol{H}_n\end{bmatrix} \end{aligned} \tag{5}$$

The general approach for denoising under NNSC is to train $\boldsymbol{D}_s$ and $\boldsymbol{D}_n$ individually on suitable data, i.e., data that matches the condition at test time, prior to test time using the procedure in Sect. 3.1. To reduce noise at test time, we then decompose the magnitude STFT of the noisy speech according to(6a) using NMF and only update $\boldsymbol{H}_s$ and $\boldsymbol{H}_n$ via the following rules

$$\boldsymbol{H}_s = \boldsymbol{H}_s \circ \frac{\boldsymbol{D}_s^T\boldsymbol{X}}{\boldsymbol{D}_s^T\boldsymbol{D}\boldsymbol{H}+\alpha_s} \tag{6a}$$

$$\boldsymbol{H}_n = \boldsymbol{H}_n \circ \frac{\boldsymbol{D}_n^T\boldsymbol{X}}{\boldsymbol{D}_n^T\boldsymbol{D}\boldsymbol{H}+\alpha_n} \tag{6b}$$

where α_s and α_n are the sparsity paramters for speech and noise, respectively. An estimate of the magnitude STFT of clean speech is then found by multiplying the Wiener gain function onto the noisy STFT

$$\hat{\boldsymbol{S}} = \boldsymbol{X} \circ \frac{\frac{\boldsymbol{D}_s\boldsymbol{H}_s}{\boldsymbol{D}_n\boldsymbol{H}_n}}{1+\frac{\boldsymbol{D}_s\boldsymbol{H}_s}{\boldsymbol{D}_n\boldsymbol{H}_n}} \tag{7}$$

The time-domain estimate is obtained using the phase from the noisy signal and applying the overlap-add method. Finally the time-domain signal is submitted to a feature extraction module.

4. Experiments

4.1. Evaluation Database and Noise Database

In this study we used a part of the RSR2015 database [12] for evaluation, which is designed for text-dependent speaker verification. Part 1 of RSR2015 consists of nine sessions for each speaker with short and fixed pass-phrases. For development we chose a subset of 47 male speakers and used prompt ID 1 with sessions 1, 4 and 7 for enrollment data and the remaining for testing, resulting in a total of 298 utterances. The trial protocol consisted of $47\cdot 298 = 14006$ trials. For the evaluation we used prompt ID 2.

We used five different noise types; white Gaussian noise (WGN), babble, cantine, PC-fan and car noise. WGN was generated in MATLAB, babble noise was generated by adding speech from Librispeech corpus [13], cantine and PC-fan was recorded and car noise was taken from [14]. All noise material was split into a training set used for training the dictionaries and a test set for evaluation. The noise was added by taking out a random segment matching the length of the relevant speech signal and then scaling it according to the desired SNR and finally adding it to the speech.

4.2. Speaker Verification Back-end

We used normal 13-dimensional MFCC features with log-energy and including delta and delta-delta coefficients and a conventional i-Vector system [15] was setup as back-end for speaker verification. The UBM consisted of 64 components and was trained on the TIMIT training data from the male speakers only. The same data was used for training the total-variability matrix which was set to have a rank of 400. An energy-based VAD was used to discard noise-only feature frames in both training and testing procedures. We used the cosine-similarity measure for scoring. It should be noted, that this back-end did not utilize temporal information as would normally be the case for text-dependent speaker verification. However, this is not an issue when only comparing the speech enhancement methods. Since three utterances, session 1, 4 and 7, were available for enrollment for each speaker, an i-Vector was computed for each utterance and the mean of these three was used as the final i-Vector.

4.3. Baseline Frontends

A number of different well-known speech enhancement algorithms were chosen for comparison. We chose spectral subtraction [16] (Ssub, implementation from [4]), minimum mean-squared error (MMSE, implementation from [17]) and Wiener filtering [18] (Wiener, implementaion from [4]) as they are theoretically well-founded algorithms designed according to signal level metrics and have been successfully applied in a number of applications.

4.4. Settings for NNSC noise reduction

A dictionary for each speaker was learned by computing the STFT of utterances from session 1, 4 and 7 (same data as enrollment for SV system) and concatenating speech-only frames and then applying (4) until convergence. Becasue each of the three utterances was recorded in different sessions, we found that it was important to normalize the time-domain signal before computing the STFT. All STFT were computed using a 32ms window with 16ms overlap and an FFT length of 512. For determining the best settings for the dictionaries in terms of sparsity, α, and number of columns, initial experiments were carried out using exhibition noise along with the development set described earlier. Based on this the size of the dictionaries (speech and noise) were set to 64 and the sparsity for speech, α_s, was set to 0.001 and the sparsity for noise, α_n, was set to 0.1 at training time but 0 at test time. At test time the algorithm was set to terminate when the relative change was less than 10^{-4} or after 500 iterations.

4.5. Speaker Verification Results

In this section we state the results in terms of equal-error-rate (EER) for the different noise types, SNR levels and algorithms (Tabs. 2-6), and we furthermore state the average improvement relative to merely using the un-processed noisy signal as input directly, in Tab. 1. It is noted that the performance on the proposed method changes under clean condition depending on noise type due to using different noise dictionaries, whereas the baseline methods do not, which is to be expected. The general trend across all noise types is that all the enhancement algorithms does improve performance for SNR$\leq$10 dB, which is in line with the results obtained in [5]. The proposed method outperforms the other algorithms with a substantial margin in this range of SNR levels except for the case of babble noise. For SNR levels of 15 and 20 dB, both spectral subtraction and Wiener filtering perform better than MMSE and the proposed method. We believe that the limited number of entries in the dictionary (64) is not able to span the full signal subspace at these SNR-levels, thus the added distortion becomes greater than the noise removed. However, this should be investigated further. Finally, only Wiener filtering does not reduce performance when clean speech is used as input. It is also seen that MMSE seems to give the overall worst performance.
To explain the performance of the proposed method on different noise types, we plotted the average values of entries in the noise dictionaries and speaker dictionaries for each frequency bin, i.e., $D_{AVG}(f) = 1/K \cdot \sum D(f,k)$. It is seen in Fig. 1 that noise types where the proposed method perform well (PC-fan, car and WGN) are not very similar to the characteristics of the speaker dictionaries. As the opposite case, babble and cantine noise are very similar to the speaker dictionaries. We can furthermore explain the lower performance of babble with the fact that this is non-stationary and different speakers were used to generate train and test data, thus unseen speakers have occured during testing.

To further analyse the performance of the different methods we have included spectrograms of an utterance for the clean, noisy and processed signals in Fig. 2. When comparing the spectrograms it is seen, that spectral subtraction and MMSE are very aggressive and distort the signal, whereas the proposed method does not seem to distort the speech as much, which explains why it performs so well in this scenario.

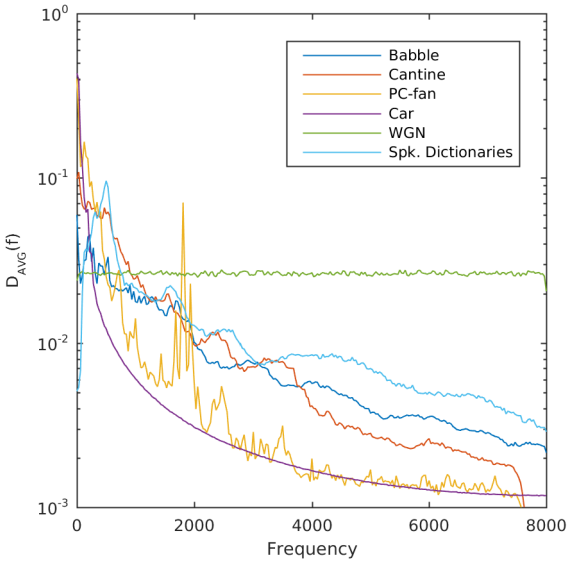

Figure 1: Plot of $D_{AVG}(f)$ for the trained noise dictionaries and speaker-dependent dictionaries.

Table 1: Average improvement (%) in EER relative to using the unprocessed noisy signal. Note that bigger is better in this case.

SNR	Ssub	MMSE	Wiener	Prop.
0 dB	9.17	11.11	12.03	**35.03**
5 dB	19.41	9.73	22.38	**33.90**
10 dB	21.78	7.00	25.47	**26.93**
15 dB	**21.44**	-5.89	21.31	17.40
20 dB	12.75	-22.11	**13.37**	1.25
Clean	-76.13	-38.13	**4.80**	-15.36

Table 2: EER (%) for white gaussian noise.

SNR	Raw	Ssub	MMSE	Wiener	Prop.
0 dB	38.70	33.21	30.36	33.21	**23.57**
5 dB	28.21	22.58	22.61	20.35	**17.40**
10 dB	18.57	15.36	18.21	13.57	**12.78**
15 dB	11.79	11.81	15.00	**10.40**	10.71
20 dB	9.29	9.29	13.21	**8.57**	10.36
Clean	7.50	13.21	10.36	7.14	**7.09**

Table 3: EER (%) for cantine noise.

SNR	Raw	Ssub	MMSE	Wiener	Prop.
0 dB	30.00	28.57	29.64	27.86	**23.93**
5 dB	23.95	18.93	21.11	19.69	**17.56**
10 dB	17.42	**12.68**	15.51	13.27	13.93
15 dB	12.50	9.51	13.57	**9.12**	10.71
20 dB	9.45	**7.50**	11.71	8.21	9.29
Clean	7.50	13.21	10.36	**7.14**	10.10

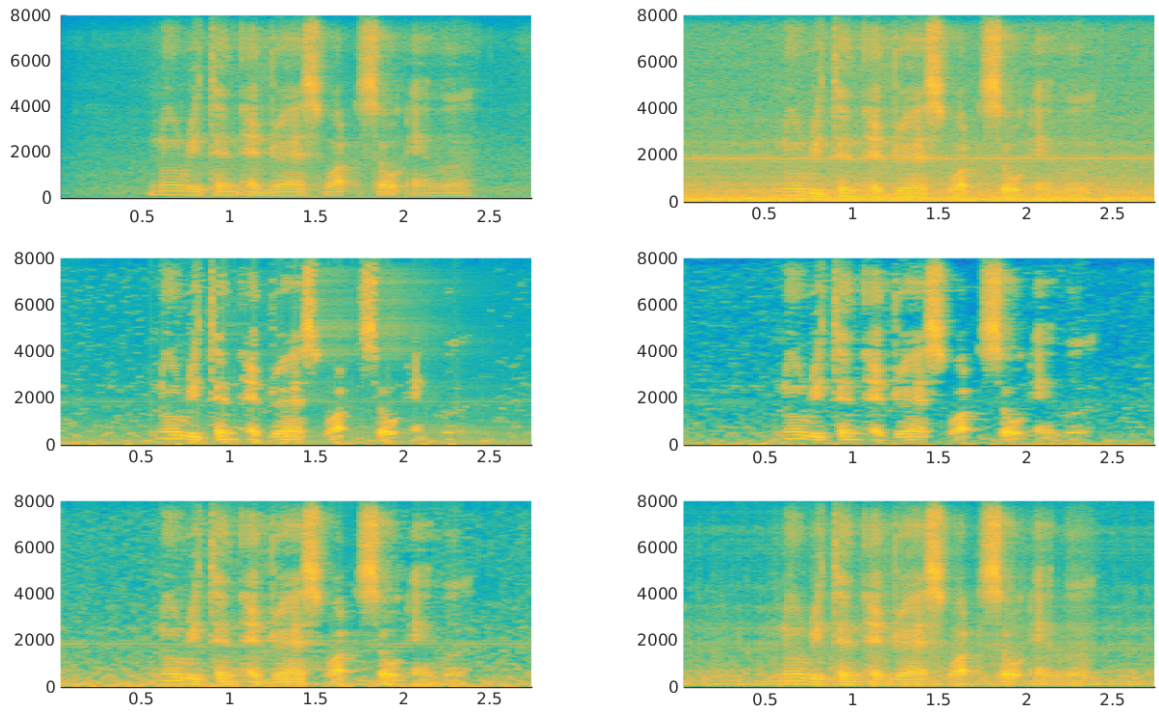

Figure 2: Spectrograms of an utterance with PC-fan noise at SNR-level of 0 dB. Top-left: clean speech, top-right: noisy speech, middle-left: Spectral subtraction, middle-right: MMSE, bottom-left: Wiener and bottom-right: proposed method.

Table 4: EER (%) for babble noise.

SNR	Raw	Ssub	MMSE	Wiener	Prop.
0 dB	36.79	36.07	38.37	**34.79**	34.82
5 dB	26.79	24.29	29.64	**23.28**	28.75
10 dB	18.38	15.36	19.64	**15.00**	21.41
15 dB	13.79	**9.75**	13.57	11.79	16.60
20 dB	10.00	**8.21**	12.50	8.55	13.99
Clean	7.50	13.21	10.36	**7.14**	12.14

Table 5: EER (%) for car noise.

SNR	Raw	Ssub	MMSE	Wiener	Prop.
0 dB	22.97	20.49	21.07	19.64	**7.86**
5 dB	20.57	15.00	18.22	14.67	**7.63**
10 dB	16.17	12.86	13.93	11.28	**7.61**
15 dB	12.18	10.00	12.14	8.93	**7.14**
20 dB	9.29	9.64	10.20	7.86	**7.13**
Clean	7.50	13.21	10.36	**7.14**	**7.14**

Table 6: EER (%) for PC-fan noise.

SNR	Raw	Ssub	MMSE	Wiener	Prop.
0 dB	29.86	25.64	21.25	24.21	**16.52**
5 dB	24.47	19.64	20.52	18.48	**12.50**
10 dB	20.19	14.64	17.14	14.54	**10.70**
15 dB	15.69	10.00	15.00	11.53	**9.02**
20 dB	11.07	7.86	12.14	9.29	**7.45**
Clean	7.50	13.21	10.36	7.14	**6.79**

5. Conclusion

In this paper we have proposed non-negative sparse coding for speech enhancement for text-dependent speaker verification, and the performance was compared with spectral subtraction, MMSE and Wiener filtering for five different noise types at six different SNR levels. The method was superior for low SNR-levels (0-10dB) increasing performance by $\approx 27-35\%$ relative to the unprocessed signal. Wiener filtering and spectral subtraction performed better at higher SNR-levels while all algorithms decreased performance relative to using the un-processed signal, when no noise was added to the signal.

An interesting path to follow is to create multiple dictionaries with different size and sparsity for different SNR-levels, as the proposed method currently does not perform well for higher SNR-levels. Another interesting approach to investigate is to concatenate the frequency vector (columns of STFT) to capture temporal information which is highly desired in the case of text-dependent speaker verification. Finally it would be interesting to compare the methods under matched conditions, i.e., when the back-end is trained on processed data.

6. Acknowledgements

The paper reflects some results from the OCTAVE Project (#647850), funded by the Research European Agency (REA) of the European Commission, in its framework programme Horizon 2020. The views expressed in this paper are those of the authors and do not engage any official position of the European Commission.

7. References

[1] J. Chen, J. Benesty, Y. Huang, and S. Doclo, "New insights into the noise reduction wiener filter," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1218–1234, July 2006.

[2] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 27, no. 2, pp. 113–120, Apr 1979.

[3] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 33, no. 2, pp. 443–445, Apr 1985.

[4] P. C. Loizou, *Speech Enhancement: Theory and Practice*, 2nd ed. Boca Raton, FL, USA: CRC Press, Inc., 2013.

[5] S. O. Sadjadi, T. Hasan, and J. H. Hansen, "Mean hilbert envelope coefficients (MHEC) for robust speaker recognition," in *Proc. Interspeech*, 2012.

[6] O. Plchot, S. Matsoukas, P. Matjka, N. Dehak, J. Ma, S. Cumani, O. Glembek, H. Hermansky, S. H. Mallidi, N. Mesgarani, R. Schwartz, M. Soufifar, Z. H. Tan, S. Thomas, B. Zhang, and X. Zhou, "Developing a speaker identification system for the darpa rats project," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, May 2013, pp. 6768–6772.

[7] 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, Sept 2011.

[8] K. W. Wilson, B. Raj, P. Smaragdis, and A. Divakaran, "Speech denoising using nonnegative matrix factorization with priors," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, March 2008, pp. 4029–4032.

[9] M. N. Schmidt, J. Larsen, and F. T. Hsiao, "Wind Noise Reduction using Non-negative Sparse Coding," in *Machine Learning for Signal Processing, 2007 IEEE Workshop on*, Aug 2007, pp. 431–436.

[10] S. Mirsamadi and J. Hansen, "Multichannel feature enhancement in distributed microphone arrays for robust distant speech recognition in smart rooms," in *2014 IEEE Spoken Language Technology Workshop (SLT)*, Dec. 2014, pp. 507–512.

[11] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *In NIPS*. MIT Press, 2000, pp. 556–562.

[12] A. Larcher, K. A. Lee, B. Ma, and H. Li, "Text-dependent speaker verification: Classifiers, databases and RSR2015," *Speech Communication*, vol. 60, pp. 56 – 77, 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167639314000156

[13] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2015, pp. 5206–5210.

[14] J. Thiemann, N. Ito, and E. Vincent, "The diverse environments multi-channel acoustic noise database: A database of multichannel environmental noise recordings," *The Journal of the Acoustical Society of America*, vol. 133, no. 5, pp. 3591–3591, 2013.

[15] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-End Factor Analysis for Speaker Verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, May 2011.

[16] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '79.*, vol. 4, Apr 1979, pp. 208–211.

[17] R. C. Hendriks, R. Heusdens, and J. Jensen, "MMSE based noise PSD tracking with low complexity," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, March 2010, pp. 4266–4269.

[18] P. Scalart and J. V. Filho, "Speech enhancement based on a priori signal to noise estimation," in *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on*, vol. 2, May 1996, pp. 629–632 vol. 2.