Noise Robust Speaker Recognition with Convolutive Sparse Coding

Antti Hurmalainen[1], *Rahim Saeidi*[2], *Tuomas Virtanen*[1]

[1]Department of Signal Processing, Tampere University of Technology, Tampere, Finland
[2]Department of Signal Processing and Acoustics, Aalto University, Espoo, Finland
antti.hurmalainen@tut.fi, rahim.saeidi@aalto.fi, tuomas.virtanen@tut.fi

Abstract

Recognition and classification of speech content in everyday environments is challenging due to the large diversity of real-world noise sources, which may also include competing speech. At signal-to-noise ratios below 0 dB, a majority of features may become corrupted, severely degrading the performance of classifiers built upon clean observations of a target class. As the energy and complexity of competing sources increase, their explicit modelling becomes integral for successful detection and classification of target speech. We have previously demonstrated how non-negative compositional modelling in a spectrogram space is suitable for robust recognition of speech and speakers even at low SNRs. In this work, the sparse coding approach is extended to cover the whole separation and classification chain to recognise the speaker of short utterances in difficult noise environments. A convolutive matrix factorisation and coding system is evaluated on 2nd CHiME Track 1 data. Over 98% average speaker recognition accuracy is achieved for shorter than three second utterances at +9 ... -6 dB SNR, illustrating the system's performance in challenging conditions.
Index Terms: speaker recognition, noise robustness, compositional models, sparse coding, non-negative matrix factorization

1. Introduction

Speech processing systems intended for real-world environments must be able to cope with signals corrupted by a large variety of interferences, including reverberation and additive noise [1, 2]. Both speech and speaker recognition require assessing small details of vocalisations, which becomes increasingly difficult when a majority of spectro-temporal features is contaminated by competing sound sources. Loud audio events and non-stationary noise can mask large areas of spectro-temporal content in any feature representation. Especially, non-target speakers introduce additional phonetic content, which is inherently problematic for recognition and difficult to separate from actual target speech. Therefore practical systems should be made robust against many levels and types of signal degradation.

As one common case of robust speech processing, recognising speakers from short utterances contaminated by noise and in presence of room reverberation is a rather challenging task and has been of interest in several recent studies [3, 4, 5, 6]. In dealing with short duration utterances in test phase, employing duration sensitive transforms [7, 8] and uncertainty estimation and propagation [3, 9] are state-of-the-art techniques to compensate for adverse effects of limited data on recognition performance. There are several strategies to attain robustness against ambient noise and reverberation. Despite the fact that conventional and state-of-the-art speech enhancement techniques introduce artifacts, employing signal-level enhancement has shown to be effective in improving speaker recognition in noisy environments [5, 10, 11]. Extracting spectral features that represent reduced mismatch between clean and noisy speech proved to be essential in handling noise and reverberation [4, 12]. Feature-level enhancement using uncertainty-of-observation techniques [6, 13] or vector Taylor series [14] aids in robust speaker modelling and recognition. Training multiple parallel models for each noise condition [15] or employing informed score calibration [16] brings further robustness against interferences. However, at low SNRs, explicit modelling of individual sound sources and separation of target speech become increasingly important.

Sparse models have gained popularity in several branches of signal processing, including recognition of images [17], music genres [18], sound events [19], and generalised components of audio [20]. Sparsity-based speaker recognition algorithms have also been proposed [21, 22, 23, 24, 25, 26], although typically the main principle has been achieving robustness by low-rank modelling, not explicit factorisation to source components. Meanwhile, compositional models, especially *non-negative matrix factorisation* (NMF), have been used for speech enhancement and robust automatic speech recognition (ASR), motivated by simultaneous modelling of multiple additive sources such as speech and noise [27, 28, 29, 30]. Most often the algorithms are used to enhance features for external ASR back-ends. However, in recent years there has been increasing success with *sparse coding* or *classification* (SC), where speech signal content is derived directly from sparse component weights without returning to a spectral or time-domain feature representation [27, 29, 31].

In [32] we presented a system, where NMF activation weights are used for speaker recognition using conventional vector classification schemes. In [33], a classifier taking place within the non-negative framework was proposed, producing speech state likelihoods for ASR. In this work, the NMF-based learning and classification algorithm is applied to speaker recognition, exploiting the factorisation model's temporal information and non-negativity for significantly higher robustness than previous vector classifiers. Results are evaluated on GRID-based Track 1 of the 2$^{\text{nd}}$ CHiME Challenge, where one speaker out of 34 candidates speaks a short, $<$2.7 s utterance in difficult, non-stationary noise conditions down to -6 dB SNR [34].

The convolutive sparse coding approach is introduced in Section 2. Our experimental setup is described in Section 3. Results, discussion, and further ideas are given in Section 4, whereafter we conclude in Section 5.

2. Convolutive modelling and mapping

2.1. Matrix convolution

The framework is based on *matrix convolution*, here denoted by operator $\circledast$. The central element is an $L \times W$ *activation matrix* $\mathbf{X}$, which reflects the non-negative weights of L *atoms* over W

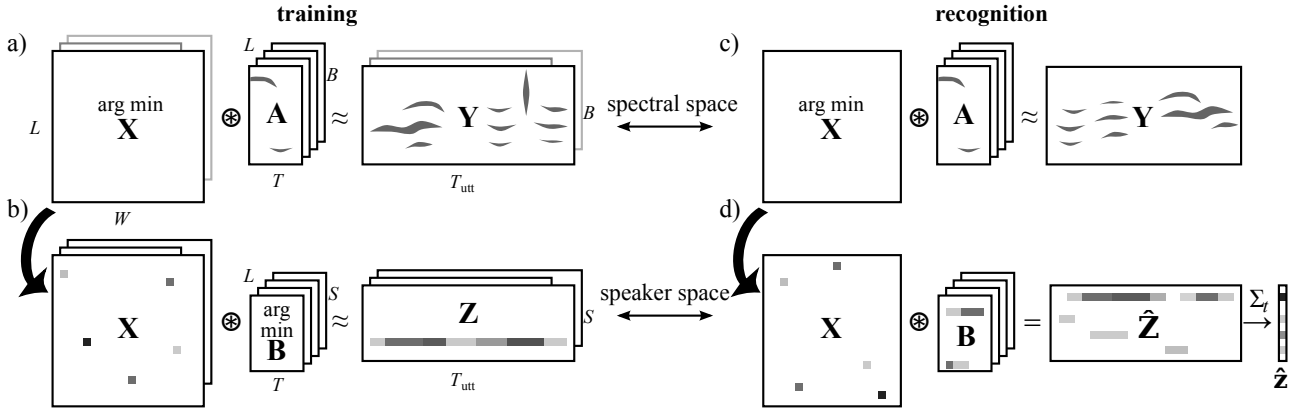

Figure 1: The proposed convolutive training and recognition procedure: a) Activations ($\mathbf{X}$) are solved for each training spectrogram, b) Mapping matrices ($\mathbf{B}$) are learnt jointly from training activations in the speaker space, c) The observation spectrogram is factored, d) Speaker magnitude estimate over time ($\hat{\mathbf{Z}}$) is computed with matrix convolution and summed to vector $\hat{\mathbf{z}}$ for recognition.

time indices. In a B-dimensional spectrogram space, a $B \times T_{\text{utt}}$ *observation matrix* $\mathbf{Y}$ covering the T_{utt} frames of an utterance is approximated by matrix $\mathbf{\Psi}$ as a convolution between activations $\mathbf{X}$ and a three-dimensional *basis array* $\mathbf{A}$ comprising L atoms, each a $B \times T$ spectrogram patch,

$$\begin{aligned} \mathbf{\Psi} &= \mathbf{X} \circledast \mathbf{A} \\ &= \sum_{t=1}^{T} \mathbf{A}_t \overset{\rightarrow(t-1)}{\mathbf{X}} . \end{aligned} \tag{1}$$

Each $B \times L$ matrix $\mathbf{A}_t$ contains the t^{th} columns of all atoms. Operator $\rightarrow$ shifts $\mathbf{X}$ right in an $L \times T_{\text{utt}}$ zero-padded matrix starting from its leftmost position in a convention, where activations are permitted up to time index $W = T_{\text{utt}} - T + 1$ so that all resulting spectral content fits in $\mathbf{\Psi}$ without cropping. All spectrogram data is also non-negative.

In the spectral factorisation stage, $\mathbf{X}$ is solved with a matrix factorisation algorithm for fixed $\mathbf{A}$ and $\mathbf{Y}$ arrays as

$$\mathbf{X}_{\text{opt}} = \arg\min_{\mathbf{X}} \left[d(\mathbf{X} \circledast \mathbf{A},\ \mathbf{Y}) + c(\mathbf{X}) \right], \tag{2}$$

the target function comprising a *spectral estimate distance* d between the observation and its estimate, and an optional cost function c for the structure of $\mathbf{X}$. For audio spectrograms, generalised Kullback-Leibler (KL) divergence and weighted L_1 norm have been found suitable and commonly applied for these purposes, respectively [20]. Solving takes place with iterative updates, which can be found for different cost functions in literature [35, 36]. Solving components of these problems is referred to as *non-negative matrix deconvolution* (NMD).

2.2. Learning speaker mapping with deconvolution

To apply the model to speaker recognition, activations $\mathbf{X}$ are computed using a joint set of speech bases from all speaker candidates to be considered. In noisy conditions, a noise basis is also included. Because the multi-frame atoms model a lot of speaker-dependent spectro-temporal content, largest activation weights can typically be seen in atom indices corresponding to matching speakers. Indeed, simply observing the most active speaker-dependent sub-basis acts as a basic classifier with reasonable baseline results [32]. However, not all atoms are equally discriminative. Some of them may match other speakers or noise patterns, producing spurious activation weights.

To improve the accuracy, we have previously used various classifiers for activation vectors $\mathbf{x}$, produced by averaging $\mathbf{X}$ matrices of training and evaluation utterances over time [32]. Notable improvements were observed over baseline maximum activity scoring. Nevertheless, these classifiers still have their shortcomings. They discard all temporal information of activations, and do not exploit the model's inherent non-negativity. Therefore we propose a new learning and classification scheme, reflecting earlier work on phonetic state mapping in ASR [33].

Instead of using temporally averaged vectors, mapping from activations $\mathbf{X}$ to S speakers is learnt between $S \times T_{\text{utt}}$ *speaker matrices* $\mathbf{Z}$ and an $S \times T \times L$ *label array* $\mathbf{B}$. These correspond to the spectral model's $\mathbf{Y}$ and $\mathbf{A}$ arrays, only with the spectral space replaced with S-dimensional *speaker space* with each index representing one speaker. In $\mathbf{Z}$ matrices, only the true speaker's row contains non-zero values, thus the algorithm is expected to assign corresponding speaker content to $\mathbf{B}$ for atoms that activate during the target speaker's utterances.

After a set of $\mathbf{X}$ matrices is solved in the spectrogram domain for utterances from all speakers, the mappings are learnt as

$$\mathbf{B}_{\text{opt}} = \arg\min_{\mathbf{B}} d(\mathbf{X} \circledast \mathbf{B},\ \mathbf{Z}) \tag{3}$$

over the set of $\mathbf{X}$ and $\mathbf{Z}$ matrix pairs of training files. The procedure is illustrated in the left half of Figure 1. To emphasise actual speech regions of training signals, values of the active row of $\mathbf{Z}$ are assigned from frame-level speech signal magnitudes of the training utterances.

The following properties apply to learnt $\mathbf{B}$ data:

- Each atom-dependent $\mathbf{B}_l$ matrix ($S \times T$) represents the estimated match to speakers over the atom's frames.
- All label data is strictly non-negative.
- The model will assign variable amounts of label weight to atoms, depending on how consistently they get activated during utterances of specific speakers.
- In an ideal case, computing $\mathbf{X} \circledast \mathbf{B}$ would produce a $\mathbf{Z}$ matrix, where only the correct speaker's row is active. These coefficients reflect the magnitude of speech.

Therefore the algorithm is consistent with the non-negativity and temporal sensitivity of the spectrogram factorisation model, unlike earlier training schemes for vector classifiers.

2.3. Recognition

To recognise speakers with the trained mappings, activations are again computed for test utterances in the spectrogram domain. Then we compute $\hat{\mathbf{Z}} = \mathbf{X} \circledast \mathbf{B}$ from speech activations, which produces the $S \times T_{\text{utt}}$ estimate $\hat{\mathbf{Z}}$ reflecting each speaker's activity over utterance frames as seen in the right half of Figure 1.

Assuming that only one target speaker is active, the straightforward recognition method is to sum $\hat{\mathbf{Z}}$ over time to vector $\hat{\mathbf{z}}$, and then pick the index of its largest value. However, competing voices and other speech-like interferences may produce large outlier entries to $\hat{\mathbf{Z}}$, calling for filtering, compression, or thresholding steps to favour candidates whose activity is more consistent over time. In our experiments, square root compression of $\hat{\mathbf{Z}}$ entries was found to improve the overall robustness. In multi-speaker scenarios, overlapping and changing speaker hypotheses should be evaluated over appropriate time intervals.

3. Experimental setup

3.1. Evaluation database

The method was evaluated on 2nd CHiME Challenge Track 1 data, consisting of GRID command utterances over non-stationary room noise. The original challenge task was robust keyword recognition within its 51-word vocabulary with full knowledge of the active speaker's identity [34, 37]. However, in this work we use the corpus for speaker recognition, treating the active speaker's identity among a closed set of 34 speakers as the unknown parameter to solve. The corpus is very closely related to its predecessor used in our earlier experiments [32, 38]. The notable difference is inclusion of a noisy training set in the 2nd CHiME, permitting multi-condition training of models.

The corpus has 500 training utterances for each speaker, both as noiseless files and with varying levels of noise from +9 to -6 dB in six 3 dB steps. Development and test sets consist of 600 utterances from mixed speakers, here considered unknown, repeated over the same six SNRs and always with different noise content. The development set is also available without additive noise. Utterance length ranges approximately from 1.2 to 2.7 s with a mean length of about 1.9 s. All audio is sampled at 16 kHz stereo and has room reverberation [34].

3.2. Factorisation framework

The system setup was effectively identical to our earlier work [32, 38], only updated for the 2nd CHiME corpus. For each speaker, we used 300 clean training utterances to generate 250 speech atoms, each a 40×25 template in a 40-dimensional mel-spectral space with 25 consecutive frames, averaged to mono. Frame length was 25 ms and frame shift 10 ms, standing for approximately 250 ms of temporal context in the atoms. The other 200 training utterances per speaker and development/test sets were represented as spectrogram matrices $\mathbf{Y}$.

Utterances were factored with NMD, using the 34 concatenated speaker-dependent bases for a total of 8500 speech atoms and also 250 noise atoms sampled from each utterance's noise context when additive noise was present in the set [29, 32]. The number of NMD iterations was 300. Weighted atom level L_1 sparsity constraint was used for $\mathbf{X}$ as in earlier work [29, 32].

To learn the $\mathbf{B}$ array, a $\mathbf{Z}$ matrix was generated for each training utterance using signal magnitude as the target value on the row of the speaker's index. As in ASR experiments, four NMD iterations with Euclidean d cost were used to solve (3), only updating the label array $\mathbf{B}$ over the set of $34 \cdot 200$ clean or noisy training utterances [33]. Recognition was performed by finding the index of maximum value in temporally summed $\hat{\mathbf{z}}$ vectors ('SC / sum'). As an alternative to suppress large outliers from loud non-target events, the $\hat{\mathbf{Z}}$ matrix bins were first square root compressed before temporal summing ('SC / compr').

Table 1: Speaker recognition accuracy over SNR for the 2nd CHiME challenge Track 1 test set. Baseline results (Section 3.3) are given in the first block. The second and third block list results for clean and noisy trained NMD classifiers, respectively. The proposed sparse coding is denoted by 'SC'.

SNR	9 dB	6 dB	3 dB	0 dB	-3 dB	-6 dB	avg
Reference methods							
i-vector	62.5	57.0	50.7	46.7	40.3	34.5	48.6
GMM-ML	96.7	91.2	88.7	76.2	68.2	60.8	80.3
max. act	99.0	98.2	97.5	93.2	88.5	69.0	90.9
Clean-trained NMD activation classifiers							
SLDA	99.7	99.2	98.8	97.3	95.3	85.8	96.0
SC / sum	99.8	99.5	99.2	97.7	96.7	86.0	96.5
SC / compr	99.8	99.7	99.5	97.3	97.2	88.3	97.0
Noisy-trained NMD activation classifiers							
SLDA	98.3	98.0	97.7	96.5	94.8	88.3	95.6
SC / sum	100.0	99.8	99.3	98.2	98.3	93.5	98.2
SC / compr	100.0	99.8	99.7	98.8	98.0	95.0	98.6

3.3. Reference systems

The following baseline systems were used for comparison:

- i-vector classification [10, 39]
- Maximum likelihood estimation with Gaussian mixture models (GMM-ML) [40]
- Maximum total activation weight of speaker-dependent bases in the convolutive factorisation (max. act) [32]
- Sparse linear discriminant analysis (SLDA) from temporally averaged NMD activation vectors [32]

The first baseline is an i-vector based system used for a NIST SRE'12 submission [10]. The system is trained using previous NIST SRE corpora. In training the probabilistic linear discriminant analysis (PLDA) transform, short and noisy segments are included to help PLDA in capturing variabilities caused by noises as well as utterances with variable duration. In training speaker templates, the extracted features from training utterances are pooled together. For the second baseline, Gaussian mixture models with 64 components were trained for each speaker with a maximum likelihood criterion (GMM-ML) [40].

Two further baselines were derived from the same activation matrices as the proposed method. Maximum activity is computed directly from test factorisations by using the largest total $\mathbf{X}$ weight of speaker-dependent bases for scoring with no further training. SLDA models were trained from temporally averaged activations using 500 non-zero components as in [32].

4. Results and Discussion

4.1. Results

Results for baseline and proposed methods are listed in Table 1. Correct recognition rate is reported for each SNR of the 2nd CHiME Track 1 test set and as averages over all noise levels.

Despite minor deviations in individual SNR scores, the main trends of performance are easily observed. The i-vector

based system represents a rather poor performance, which can be mainly attributed to extremely short test utterances (effective speech duration of $\sim$1 s after voice activity detection). The GMM-ML system is mostly accurate in clean conditions, but loses quality over increasing noise levels. All methods derived from the NMD system surpass more conventional baselines by a clear margin. A major factor is explicit modelling of speech and noise components, whose speech-only activations provide significantly more reliable input for classification.

The new sparse coding approach outperforms direct maximum activation scoring and SLDA of averaged vectors from the same NMD system. Noisy-trained SLDA does not reach perfect accuracy at clean conditions due to larger variability in training, but it turns more robust in noisy conditions. Conversely, all proposed SC variants were 100% accurate for clean development data, thus for this method noisy training produced uniformly better results than clean training. Some improvements are achieved by applying additional compression to $\hat{\mathbf{Z}}$ matrices before summing. Overall, the highest average recognition rate is 98.6% with even the -6 dB case yielding 95.0% accuracy.

While directly comparable results were not found in literature at the moment of writing, reported average results for the effectively identical 1st CHiME corpus include up to 72.3% in [41] using NMF enhancement for a GMM-UBM system, and up to 94.0% in [42] with uncertainty modelling and self-mixed multi-condition training for a GMM recogniser. Further results for the GRID corpus are listed in [43], where a wavelet and multimodal neural network system achieves 97.5% accuracy for clean data with emphasis on fast (50 ms) identification.

4.2. Discussion

According to the results, proposed convolutive sparse coding appears very accurate for robust recognition of speakers from short utterances in difficult noise conditions. The foremost contributing factors are explicit modelling of speech and noise, and relatively long temporal context used in the representation of spectrogram patterns. The system can thus separate and classify components from additive mixtures based on their spectro-temporal behaviour over extended periods. Conversely, the total amount of audio content required for recognition is low, because a few characteristic patterns suffice for reliable recognition.

Although this particular task is simplified by its small vocabulary, which is easy to model with long templates, we have already described algorithms for segmentation and recognition with variable-length units of speech, suggesting that the general approach is also applicable to large vocabulary tasks [44].

Concerning complexity, all steps of the proposed system use large-scale linear algebra operations, which are well suited for parallel computing. The Matlab NMD solver used for these experiments achieves real-time performance on a GTX 750 Ti desktop GPU. However, the complexity scales linearly to basis size, atom length, observation length, and iteration count, which may differ greatly for other tasks and configurations. On the other hand, the setup was originally designed for ASR via sparse coding, which requires discovery of a full phonetic sequence rather than just recognising the speaker. Therefore the setup could probably be simplified for the latter task.

4.3. Extensions and future work

As stated in Section 1, the proposed training and classification algorithm has its direct counterpart in ASR, if phonetic states are used as the target matrix instead of speakers. Indeed, exactly the same convolutive model can be used for ASR, even at the same time by concatenating the speaker and speech state matrices in learning and recognition. The system will then perform speaker-independent ASR and speaker recognition simultaneously. This dual functionality appears promising, because it opens several options for direct classification and model selection in multi-speaker scenarios. Furthermore, the model's additivity means that it should be applicable to *diarisation* of multiple, potentially overlapping speakers as studied in [45].

For now, it is an open question whether further gains could be achieved with *group sparsity*. One such model was proposed in [38] to favour solutions, where only a few speaker bases are active per utterance. However, shrinking the solution in the factorisation stage causes early selection of speakers, which may turn out incorrect especially when non-target speakers are present in the input. In an unfavourable case, group sparsity may attenuate true speakers to such extent that the later recognition step cannot compensate it. As an alternate approach, it could be viable to compute the $\hat{\mathbf{Z}}$ matrix continuously during factorisation, and to use the continuity of its speaker estimates as an additional criterion for optimisation of cost (2).

Overall, reliable speaker recognition in adverse conditions has its direct applications, but it is also useful in ASR for selecting correct speech models for enhancement and back-end recognition. We expect to study these options further with the proposed method.

5. Conclusions

A framework based on convolutive non-negative matrix modelling was presented to perform noise robust speaker recognition using a sparse coding approach. Spectrogram factorisation with concatenated long-context speech bases and an explicit noise model is used to separate and classify speakers using their characteristic spectro-temporal patterns. The system was evaluated on 2nd CHiME Track 1 data, achieving over 98% average accuracy for short utterances in non-stationary noise at +9 ... -6 dB SNR, also involving competing non-target voices.

The proposed system is fully contained within a non-negative modelling framework with no additional components like GMM evaluation. The system has a direct parallel in speech recognition, permitting joint recognition of speech content and speakers in complex multi-source scenarios. Potential future work includes speed optimisations, multi-speaker tasks like diarisation, and application to more diverse recognition scenarios.

6. Acknowledgements

This work has been partially funded by the Academy of Finland (projects 256961, 258708 and 284671). We acknowledge the computational resources provided by the Aalto Science-IT project.

7. References

[1] T. Virtanen, R. Singh, and B. Raj, Eds., *Techniques for Noise Robustness in Automatic Speech Recognition*. New York, NY, USA: Wiley, 2012.

[2] J. Li, L. Deng, Y. Gong, and R. Haeb-Umbach, "An Overview of Noise-Robust Automatic Speech Recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 4, pp. 745–777, 2014.

[3] S. Cumani, O. Plchot, and P. Laface, "On the use of i-vector posterior distributions in Probabilistic Linear Discriminant Analysis," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 4, pp. 846–857, 2014.

[4] S. Ganapathy, S. H. Mallidi, and H. Hermansky, "Robust Feature Extraction Using Modulation Filtering of Autoregressive Mod-

els," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 8, pp. 1285–1295, 2014.

[5] X. Zhao, Y. Wang, and D. Wang, "Robust Speaker Identification in Noisy and Reverberant Conditions," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 4, pp. 836–845, 2014.

[6] C. Yu, G. Liu, S. Hahm, and J. H. L. Hansen, "Uncertainty propagation in front end factor analysis for noise robust speaker recognition," in *Proc. of the 39th ICASSP*, Florence, Italy, 2014, pp. 4017–4021.

[7] T. Hasan, R. Saeidi, J. H. L. Hansen, and D. A. van Leeuwen, "Duration mismatch compensation for i-vector based speaker recognition systems," in *Proc. of the 38th ICASSP*, Vancouver, BC, Canada, 2013, pp. 7663–7667.

[8] A. Nautsch, C. Rathgeb, C. Busch, H. Reininger, and K. Kasper, "Towards duration invariance of i-Vector based adaptive score normalization," in *Odyssey speaker and language recognition workshop*, Joensuu, Finland, 2014.

[9] P. Kenny, T. Stafylakis, P. Ouellet, J. Alam, and P. Dumouchel, "PLDA for speaker verification with utterances of arbitrary duration," in *Proc. of the 38th ICASSP*, Vancouver, BC, Canada, 2013, pp. 7649–7653.

[10] R. Saeidi, K. A. Lee, T. Kinnunen *et al.*, "I4U submission to NIST SRE 2012: A large-scale collaborative effort for noise-robust speaker verification," in *Proc. of the 14th INTERSPEECH*, Lyon, France, 2013, pp. 1986–1990.

[11] S. O. Sadjadi and J. H. L. Hansen, "Assessment of single-channel speech enhancement techniques for speaker identification under mismatched conditions," in *Proc. of the 11th INTERSPEECH*, Makuhari, Japan, 2010, pp. 2138–2141.

[12] R. Saeidi, J. Pohjalainen, T. Kinnunen, and P. Alku, "Temporally Weighted Linear Prediction Features for Tackling Additive Noise in Speaker Verification," *IEEE Signal Processing Letters*, vol. 17, no. 6, pp. 599–602, 2010.

[13] J. Ming, T. J. Hazen, J. R. Glass, and D. A. Reynolds, "Robust Speaker Recognition in Noisy Conditions," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 5, pp. 1711–1723, 2007.

[14] Y. Lei, L. Burget, and N. Scheffer, "A noise robust i-vector extractor using vector Taylor series for speaker recognition," in *Proc. of the 38th ICASSP*, Vancouver, BC, Canada, 2013, pp. 6788–6791.

[15] D. Garcia-Romero, X. Zhou, and C. Y. Espy-Wilson, "Multicondition Training of Gaussian PLDA Models in I-Vector Space for Noise and Reverberation Robust Speaker Recognition," in *Proc. of the 37th ICASSP*, Kyoto, Japan, 2012, pp. 4257–4260.

[16] L. Ferrer, L. Burget, O. Plchot, and N. Scheffer, "A unified approach for audio characterization and its application to speaker recognition," in *Odyssey speaker and language recognition workshop*, Singapore, 2012.

[17] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust Face Recognition via Sparse Representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 210–227, 2009.

[18] Y. Panagakis, C. Kotropoulos, and G. R. Arce, "Music Genre Classification via Sparse Representations of Auditory Temporal Modulations," in *Proc. of the 17th EUSIPCO*, Glasgow, Scotland, UK, 2009, pp. 1–5.

[19] A. Mesaros, O. Dikmen, T. Heittola, and T. Virtanen, "Sound Event Detection in Real Life Recordings Using Coupled Matrix Factorization of Spectral Representations and Class Activity Annotations," in *Proc. of the 40th ICASSP*, Brisbane, Australia, 2015, pp. 151–155.

[20] T. Virtanen, J. F. Gemmeke, B. Raj, and P. Smaragdis, "Compositional models for audio processing," *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 125–144, 2015.

[21] Y.-H. Long, L.-R. Dai, E.-Y. Wang, B. Ma, and W. Guo, "Non-negative matrix factorization based discriminative features for speaker verification," in *Proc. of the 7th ISCSLP*, Tainan, Taiwan, 2010, pp. 291–295.

[22] I. Naseem, R. Togneri, and M. Bennamoun, "Sparse Representation for Speaker Identification," in *Proc. of the 20th ICPR*, Istanbul, Turkey, 2010, pp. 4460–4463.

[23] Q. Wu, L.-Q. Zhang, and G.-C. Shi, "Robust Feature Extraction for Speaker Recognition Based on Constrained Nonnegative Tensor Factorization," *Journal of Computer Science and Technology*, vol. 25, no. 4, pp. 745–754, 2010.

[24] J. M. K. Kua, E. Ambikairajah, J. Epps, and R. Togneri, "Speaker Verification Using Sparse Representation Classification," in *Proc. of the 36th ICASSP*, Prague, Czech Republic, 2011, pp. 4548–4551.

[25] C. Joder and B. Schuller, "Exploring Nonnegative Matrix Factorization for Audio Classification: Application to Speaker Recognition," in *Proc. of ITG Conference on Speech Communication*, Braunschweig, Germany, 2012.

[26] C. Tzagkarakis and A. Mouchtaris, "Sparsity Based Noise Robust Speaker Identification Using a Discriminative Dictionary Learning Approach," in *Proc. of the 21st EUSIPCO*, Marrakech, Morocco, 2013.

[27] J. F. Gemmeke, T. Virtanen, and A. Hurmalainen, "Exemplar-based Sparse Representations for Noise Robust Automatic Speech Recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2067–2080, 2011.

[28] F. Weninger, J. Feliu, and B. Schuller, "Supervised and Semi-supervised Suppression of Background Music in Monaural Speech Recordings," in *Proc. of the 37th ICASSP*, Kyoto, Japan, 2012, pp. 61–64.

[29] A. Hurmalainen, J. F. Gemmeke, and T. Virtanen, "Modelling non-stationary noise with spectral factorisation in automatic speech recognition," *Computer Speech & Language*, vol. 27, no. 3, pp. 763–779, 2013.

[30] N. Mohammadiha, P. Smaragdis, and A. Leijon, "Supervised and Unsupervised Speech Enhancement Using Nonnegative Matrix Factorization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 10, pp. 2140–2151, 2013.

[31] E. Yılmaz, J. F. Gemmeke, and H. Van hamme, "Noise Robust Exemplar Matching Using Sparse Representations of Speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 8, pp. 1306–1319, 2014.

[32] R. Saeidi, A. Hurmalainen, T. Virtanen, and D. A. van Leeuwen, "Exemplar-based Sparse Representation and Sparse Discrimination for Noise Robust Speaker Identification," in *Odyssey speaker and language recognition workshop*, Singapore, 2012.

[33] A. Hurmalainen and T. Virtanen, "Learning State Labels for Sparse Classification of Speech with Matrix Deconvolution," in *Proc. of ASRU*, Olomouc, Czech Republic, 2013, pp. 168–173.

[34] E. Vincent, J. Barker, S. Watanabe, J. Le Roux, F. Nesta, and M. Matassoni, "The Second CHiME Speech Separation and Recognition Challenge: an Overview of Challenge Systems and Outcomes," in *Proc. of ASRU*, Olomouc, Czech Republic, 2013, pp. 162–167.

[35] P. Smaragdis, "Convolutive Speech Bases and their Application to Supervised Speech Separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 1, pp. 1–14, 2007.

[36] A. Cichocki, R. Zdunek, A. H. Phan, and S. Amari, *Nonnegative Matrix and Tensor Factorizations*. New York, NY, USA: Wiley, 2009.

[37] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An Audio-visual Corpus for Speech Perception and Automatic Speech Recognition," *Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 2421–2424, 2006.

[38] A. Hurmalainen, R. Saeidi, and T. Virtanen, "Group Sparsity for Speaker Identity Discrimination in Factorisation-based Speech Recognition," in *Proc. of the 13th INTERSPEECH*, Portland, OR, USA, 2012, pp. 2138–2141.

[39] N. Dehak, P. J. 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, 2011.

[40] D. A. Reynolds, "Speaker identification and verification using Gaussian mixture speaker models," *Speech Communication*, vol. 17, no. 1–2, pp. 91–108, 1995.

[41] N. Lyubimov, M. Nastasenko, M. Kotov, and D. Doroshin, "Exploiting Non-negative Matrix Factorization with Linear Constraints in Noise-Robust Speaker Identification," in *Proc. of the 16th SPECOM*, Novi Sad, Serbia, 2014, pp. 200–208.

[42] A. Ozerov, M. Lagrange, and E. Vincent, "Uncertainty-based learning of acoustic models from noisy data," *Computer Speech and Language*, vol. 27, no. 3, pp. 874–894, 2013.

[43] N. Almaadeed, A. Aggoun, and A. Amira, "Speaker identification using multimodal neural networks and wavelet analysis," *IET Biometrics*, vol. 4, no. 1, pp. 18–28, 2015.

[44] A. Hurmalainen, J. F. Gemmeke, and T. Virtanen, "Compact Long Context Spectral Factorisation Models for Noise Robust Recognition of Medium Vocabulary Speech," in *Proc. of the 2nd CHiME workshop*, Vancouver, Canada, 2013, pp. 13–18.

[45] J. T. Geiger, R. Vipperla, S. Bozonnet, N. Evans, B. Schuller, and G. Rigoll, "Convolutive Non-Negative Sparse Coding and New Features for Speech Overlap Handling in Speaker Diarization," in *Proc. of the 13th INTERSPEECH*, Portland, OR, USA, 2012, pp. 2054–2157.