NMF-based Temporal Feature Integration for Acoustic Event Classification

Jimmy Ludeña-Choez[1,2], *Ascensión Gallardo-Antolín*[1]

[1]Dept. of Signal Theory and Communications, Universidad Carlos III de Madrid,
Avda. de la Universidad 30, 28911 - Leganés (Madrid), Spain
[2]Facultad de Ingenierías, Universidad Católica San Pablo, Arequipa, Perú

jimmy@tsc.uc3m.es, gallardo@tsc.uc3m.es

Abstract

In this paper, we propose a new front-end for Acoustic Event Classification tasks (AEC) based on the combination of the temporal feature integration technique called Filter Bank Coefficients (FC) and Non-Negative Matrix Factorization (NMF). FC aims to capture the dynamic structure in the short-term features by means of the summarization of the periodogram of each short-term feature dimension in several frequency bands using a predefined filter bank. As the commonly used filter bank has been devised for other tasks (such as music genre classification), it can be suboptimal for AEC. In order to overcome this drawback, we propose an unsupervised method based on NMF for learning the filters which collect the most relevant temporal information in the short-time features for AEC. The experiments show that the features obtained with this method achieve significant improvements in the classification performance of a Support Vector Machine (SVM) based AEC system in comparison with the baseline FC features.

Index Terms: acoustic event classification, temporal feature integration, non-negative matrix factorization

1. Introduction

In recent years, the problem of automatically detecting and classifying acoustic non-speech events has attracted the attention of numerous researchers. Although speech is the most informative acoustic event, other kind of sounds (such as laughs, coughs, keyboard typing, etc.) can give relevant cues about the human presence and activity in a certain scenario (for example, in an office room). This information could be used in different applications, mainly in those with perceptually aware interfaces such as smart-rooms [1]. Additionally, acoustic event detection and classification systems, can be used as a pre-processing stage for automatic speech recognition (ASR) in such a way that this kind of sounds can be removed prior to the recognition process increasing its robustness. In this paper, we focus on acoustic event classification (AEC).

A design of a suitable feature extraction process for AEC is an important issue. Several front-ends have been proposed in the literature, some of them based on Mel-Frequency Cepstral Coefficients (MFCC) [1], [2], [3], [4], log filter bank energies [3], Perceptual Linear Prediction (PLP) [5], log-energy, spectral flux, entropy and zero-crossing rate [1]. Most of these features are short-time characteristics in the sense that they are computed on a frame-by-frame basis (typically, the frame period used for speech/audio analysis is about 10-20 ms).

In other approaches (often denoted as *temporal feature integration* [6]), features at larger time scales are extracted by combining somehow the short-time characteristics information over a longer time-frame composed of several consecutive frames. The most common temporal integration technique consists of mapping the short-time features to their statistics (mean, standard deviation, skewness, etc.) computed over a certain temporal window [7], [8].

Recently, another temporal integration approach based on Filter Bank Coefficients (FC), which was initially proposed for general audio and music genre classification [6], [9], [10], has been experimented for AEC with promising results [7]. In contrast to the statistics-based features, FC allows to capture the dynamic structure in the short-time features. The idea behind FC is to summarize the periodogram of each short-time feature dimension by computing the power in several predefined frequency bands using a filter bank, which is usually the one proposed in [9]. However, as pointed in [10], this fixed filter bank is not general enough since the relevance of the dynamics in the short-time features for classification can be expected to be task-dependent. In this context, in [10] a supervised method for learning an optimal filter bank for music genre classification is presented. In this paper, we present an unsupervised method based on Non-Negative Matrix Factorization (NMF) for the design of a filter bank more suitable for AEC and show that the proposed method outperforms the baseline FC parameters in an AEC task. In addition, our method is very versatile, in the sense that it is not specific for AEC and therefore, it can be applied to other speech/audio classification tasks.

This paper is organized as follows: Section 2 describes the audio feature extraction process for AEC. Section 3 presents the mathematical background of NMF and its application for the unsupervised design of the filter bank for the FC-based front-end. Section 4 describes the experiments and results to end with some conclusions and ideas for future work in Section 5.

2. Audio Feature Extraction

Figure 1 represents the block diagram of the feature extraction process for AEC. It consists of two main stages: short-time feature extraction and temporal feature integration.

2.1. Short-time feature extraction

In this work, we have considered two different acoustic parameters as short-time features: the well-known MFCC and a modification of this baseline parameterization denoted as MFCC_HP.

MFCC_HP has been motivated by the study performed in [11] in which the relevance of medium and high frequencies for distinguishing between different acoustic events is observed, suggesting that a high pass filtering of the short-term spectrum of the audio signal can be beneficial for improving the discrimination capabilities of the AEC system. In practice, this is ac-

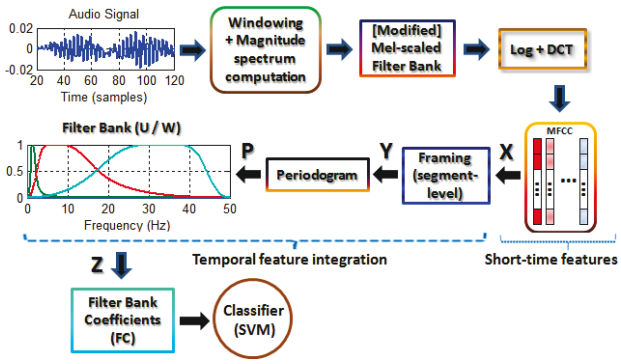

Figure 1: Block diagram of the feature extraction process.

complished by the explicit removal of several low-frequency filters from the original mel-scaled filter bank. In [11] it is shown that good results are obtained by eliminating the first low-frequency filters corresponding to frequencies below 75-275 Hz. Once the audio spectrum is filtered with this modified filter bank and the corresponding log filter bank energies are computed, a Discrete Cosine Transform (DCT) is applied over them as in the case of the conventional MFCC, yielding to a set of cepstral coefficients.

In both, MFCC and MFCC_HP, the log-energy of each frame and the first derivatives (where indicated) are computed and added to the cepstral coefficients.

2.2. Temporal feature integration

Once the cepstral coefficients are extracted, temporal feature integration is applied over audio segments of a given length in order to obtain a set of feature vectors at a larger time scale. In this work, we focus on the approach called Filter Bank Coefficients (FC) [6], [9], [10], which aims at capturing the temporal short-time features' behaviour.

First, the sequence of T short-time coefficients of dimension D_x, $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_T\}$ is divided into K segments, $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_K\}$ as follows,

$$\mathbf{y}_k = \{\mathbf{x}_{k \cdot H_s}, \mathbf{x}_{k \cdot H_s + 1}, ..., \mathbf{x}_{k \cdot H_s + L_s - 1}\} \quad (1)$$

where L_s is the segment size and H_s is the hop size, both defined in number of short-time frames.

Second, the periodogram of each dimension of the short-time features contained in the k-th segment $\mathbf{y}_k$ is estimated and, then, it is summarized by calculating the power in different frequency bands using a predefined filter bank,

$$\mathbf{z}_k = \mathbf{P}_k \mathbf{U} \quad (2)$$

where $\mathbf{P}_k$ comprises the periodograms of the sequence of the short-time coefficients belonging to the k-th segment, $\mathbf{U}$ is the frequency magnitude response of the filter bank and $\mathbf{z}_k$ is the final feature vector. The dimensions of $\mathbf{P}_k$, $\mathbf{U}$ and $\mathbf{z}_k$ are, respectively, D_x x D_p, D_p x n_f and D_x x n_f, where D_p is the dimensionality of each individual periodogram and n_f is the number of filters in the bank. The FC parameters $\mathbf{Z} = \{\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_K\}$ are the input to the AEC system, which, in this case, is based on Support Vector Machines (SVM).

Previous works [6], [9], in which the FC approach has been applied for general audio and music genre classification tasks, use a filter bank $\mathbf{U}$ composed of four filters corresponding to the following frequency bands:

- Filter 1: 0 Hz (DC filter)
- Filter 2: 1 - 2 Hz (modulation energy)
- Filter 3: 3 - 15 Hz (modulation energy)
- Filter 4: 20 - 43 Hz (perceptual roughness)

As the importance of the different dynamics in short-time features for classification may depend on the task, it can be argued that this fixed filter bank is not optimal for all audio classification problems. In other words, some modulation frequencies can be relevant for distinguishing between, for example, different acoustic events, and not between music genres. In next section, we present an unsupervised method for designing the FC filter bank and its application to AEC.

3. NMF-based design of the FC filter bank

Our goal is to develop an unsupervised approach to find the optimal filter bank in such a way that the resulting FC parameters $\mathbf{z}$ carry the most significant information about the underlying temporal structure of the short-time features. This problem can be formulated as the decomposition of the periodograms $\mathbf{P}$ into their main components (i.e., into their more relevant frequency bands).

Non-Negative Matrix Factorization (NMF) provides a way to decompose a signal into a convex combination of non-negative building blocks (called Spectral Basis Vectors, SBV) by minimizing a given cost function. As both, the power spectrum of the MFCCs and the frequency response of the elements of the filter bank, are inherently positive, NMF can offer a suitable solution to our problem, as will be explained in next subsections. Along the rest of the paper, we denote the filter bank obtained by NMF as $\mathbf{W}$ in order to distinguish it from the fixed filter bank $\mathbf{U}$.

3.1. Non-Negative Matrix Factorization (NMF)

Given a matrix $\mathbf{V} \in \mathbb{R}_+^{A \times B}$, where each column is a data vector, NMF approximates it as a product of two matrices of non-negative low rank $\mathbf{W}$ and $\mathbf{H}$, such that

$$\mathbf{V} \approx \mathbf{WH} \quad (3)$$

where $\mathbf{W} \in \mathbb{R}_+^{A \times C}$ and $\mathbf{H} \in \mathbb{R}_+^{C \times B}$ and normally $C \leq min\,(A, B)$. This way, each column of $\mathbf{V}$ can be written as a linear combination of the C basis vectors (columns of $\mathbf{W}$), weighted with the coefficients of activation or gain located in the corresponding column of $\mathbf{H}$. NMF can be seen as a dimensionality reduction of data vectors from an $A-$dimensional space to the $C-$dimensional space. This is possible if the columns of $\mathbf{W}$ uncover the latent structure in the data [12]. The factorization is achieved by an iterative minimization of a given cost function as, for example, the Euclidean distance or the generalized Kullbak Leibler (KL) divergence,

$$D_{\mathrm{KL}}\,(\mathbf{V} \| \mathbf{WH}) = \sum_{ij} \left(\mathbf{V}_{ij} log \frac{\mathbf{V}_{ij}}{(\mathbf{WH})_{ij}} - (\mathbf{V} - \mathbf{WH})_{ij} \right) \quad (4)$$

In this work, we consider the KL divergence because it has been recently used with good results in speech processing tasks, such as speech enhancement and denoising for ASR tasks [13] [14] or feature extraction [15]. In order to find a local optimum value for the KL divergence between $\mathbf{V}$ and $(\mathbf{WH})$, an iterative

scheme with multiplicative update rules can be used as proposed in [12] and stated in (5),

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

where $\mathbf{1}$ is a matrix of size $\mathbf{V}$, whose elements are all ones and the multiplications $\otimes$ and divisions are component wise operations. NMF produces a sparse representation of the data, reducing the redundancy.

3.2. Constructing the FC filter bank with NMF

As mentioned before, the matrix to be descomposed is formed by the periodograms of the short-time features. As a unique filter is learnt for all the components, the matrix $\mathbf{P}$ consists of the row-wise concatenation of the D_x periodograms of the short-time parameters extracted from the training set of the different acoustic events considered. Therefore, the dimension of $\mathbf{P}$ is $(D_x \text{ x } n_s) \text{ x } D_p$, where n_s is the total number of segments in the training set.

Once this matrix is transposed ($\mathbf{P^T}$), its corresponding factored matrices $\mathbf{WH}$ are obtained using the learning rules in (5). The dimensions of $\mathbf{W}$ and $\mathbf{H}$ are, respectively, $D_p \text{ x } n_f$ and $n_f \text{ x } (D_x \text{ x } n_s)$. The resulting matrix $\mathbf{W}$ contains the SBVs which represent the basis of the power spectrum of the short-time features, as it is verified that $\mathbf{P^T} \approx \mathbf{WH}$, and, therefore, they could be interpreted as the filters of the required filter bank.

In order to compute the NMF-based FC parameters, equation (2) is applied substituting the fixed filter bank $\mathbf{U}$ by $\mathbf{W}$.

4. Experiments and results

4.1. Database and baseline system

The database used for the experiments consists of a total of 2,114 instances of target events belonging to 12 different acoustic classes: applause, cough, chair moving, door knock, door open/slam, keyboard typing, laugh, paper work, phone ring, steps, spoon/cup jingle and key jingle. The composition of the whole database was intended to be similar to the one used in [3] and it is shown in Table 1. Audio files were obtained from different sources: websites, the FBK-Irst database [16] and the UPC-TALP database [17]. The total number of audio segments of 2 s length in the database (see subsection 4.2) is 7,775.

Table 1: *Database used in the experiments.*

Class	Event type	No. of occurrences
1	Applause [ap]	155
2	Cough [co]	199
3	Chair moving [cm]	115
4	Door knock [kn]	174
5	Door open/slam [ds]	251
6	Keyboard typing [kt]	158
7	Laugh [la]	224
8	Paper work [pw]	264
9	Phone ring [pr]	182
10	Steps [st]	153
11	Spoon/cup jingle [cl]	108
12	Key jingle [kj]	131
Total		2,114

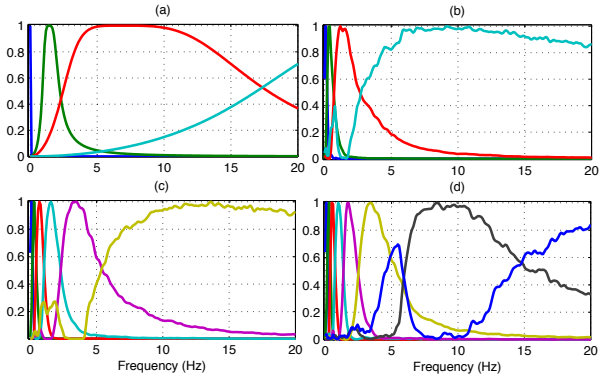

Figure 2: Frequency responses of the filter banks used in the temporal feature integration process. (a) Fixed filter bank ($\mathbf{U}$), 4 filters; Filter banks determined by NMF ($\mathbf{W}$): (b) 4 filters; (c) 6 filters; (d) 8 filters.

Since this database is too small to achieve reliable classification results, we have used a 6-fold cross validation to artificially extend it, averaging the results afterwards. Specifically, we have split the database into six disjoint balanced groups. One different group is kept for testing in each fold, while the remainder are used for training.

The AEC system is based on a one-against-one SVM with RBF kernel and a majority voting scheme for the final decision [7]. For each one of these experiments, a 5-fold cross validation was used for computing the optimal values of RBF kernel parameters.

4.2. Feature extraction

Two different types of short-time features are considered: MFCC and MFCC_HP. The difference between them is that for MFCC_HP the two first filters in the mel-scaled filter bank are removed, so frequencies below approximately 75 Hz are not considered in the cepstral coefficients computation. In both parameterizations, 12 cepstral parameters are extracted every 10 ms using a Hamming analysis window of 20 ms long and a mel-scaled filter bank composed of 40 and 38 triangular bands for MFCC and MFCC_HP, respectively. Also, the log-energy of each frame and the first derivatives (where indicated) are computed and added to the cepstral coefficients, yielding to a $D_x = 13$ (or 26 when the first derivatives are used) dimension short-time feature vector.

For the temporal feature integration, audio segments of 2 s length with overlap of 1 s are used. The periodograms of each short-time feature dimension are computed over these segments and filtered using the filter bank $\mathbf{U}$ defined in subsection 2.2, for the baseline FC parameters and the filter bank $\mathbf{W}$ obtained with the NMF method, for the NMF-based FC ones.

The filters of $\mathbf{U}$ are 2$^{\text{nd}}$ order Butterworth filters. On the contrary, in the NMF-based method, for each fold, the filter bank $\mathbf{W}$ is obtained by applying the method described in section 3 over the corresponding training set. In all folds, NMF is initialized by generating 10 random matrices ($\mathbf{W}$ and $\mathbf{H}$), in such a way that the factorization with the smallest euclidean distance between $\mathbf{P^T}$ and ($\mathbf{W}$ $\mathbf{H}$) is chosen for initialization. Then, these initial matrices are refined by minimizing the KL divergence using the multiplicative update rules given in equation (5) and a maximum of 200 iterations. Finally, the resulting $\mathbf{W}$ contains the filters of the required bank.

Table 2: *Classification rate [%] for different feature sets.*

Short-time features	Temporal integration	Classif. rate (segment) [%]	Classif. rate (event) [%]
MFCC	FC	65.68 ± 1.06	70.59 ± 1.94
MFCC_HP	FC	70.91 ± 1.01	76.39 ± 1.81
MFCC_HP	FC_NMF	74.39 ± 0.97	79.29 ± 1.73
MFCC + Δ	FC	67.92 ± 1.04	71.75 ± 1.92
MFCC_HP + Δ	FC	72.36 ± 0.99	76.39 ± 1.81
MFCC_HP + Δ	FC_NMF	76.15 ± 0.95	80.15 ± 1.70

Figures 2 (b), (c) and (d) represent the NMF-based filter bank $\mathbf{W}$ obtained on a single fold using the previous procedure for n_f = 4, 6 and 8 filters, respectively. For comparison purposes, the baseline filter bank $\mathbf{U}$ is also represented in Figure 2 (a). Note that, although the maximum modulation frequency is 50 Hz (the short-term features are extracted each 10 ms), for improving the readability of the figures, only frequencies up to 20 Hz are represented. From the comparison of Figures 2 (a) and (b), it can be seen that filters 1 and 2 of $\mathbf{U}$ roughly appears in $\mathbf{W}$. The highest frequency filter in $\mathbf{W}$ presents a high bandwidth and covers the modulation frequencies of the baseline filters 3 and 4. Finally, the filter 4 of $\mathbf{U}$ is substituted by a low-frequency filter in $\mathbf{W}$, suggesting that, for describing the temporal structure of the MFCCs, low modulation frequencies are more relevant than high ones. The same conclusion can be drawn from Figures 2 (c) and (d), in which it can be observed that, when the number of filters increases, NMF tends to place more filters in low and medium modulation frequencies than in high frequencies. Also, it is worth mentioning that the resulting filters do not differ very much between folds.

4.3. Experiments with NMF-based FC parameters

Table 2 shows the results achieved in terms of the average classification rate at segment level (percentage of segments correctly classified) and at target event level (percentage of target events correctly classified) as well as the corresponding 95 % confidence intervals for the different parameterizations considered. "FC" and "FC_NMF" indicates, respectively, the use of the fixed filter bank and the NMF-based one, both composed of 4 filters, in the temporal feature integration process. The suffix "$+\Delta$" indicates that the short-time feature set includes the first derivatives of the cepstral coefficients.

First of all, it can be observed that, in general, the use of Δ parameters improves the classification results with respect to the case in which Δs are not considered, although these differences are not statistically significant. Anyway, both cases follow the same trends. In fact, for either case (without and with Δs), when comparing MFCC with MFCC_HP, both using the baseline filter bank $\mathbf{U}$ (FC), it can be observed that MFCC_HP achieves better results, being the difference in performance with respect to MFCC statistically significant at 95% confidence. This result suggests that the high pass filtering of the acoustic event spectrum prior to the computation of the cepstral coefficients is useful for obtaining more discriminative features, and therefore, for improving the final results.

With respect to the use of the filter bank extracted by the NMF procedure in combination with the short-time features MFCC_HP (MFCC_HP + NMF_FC), it can be seen that this parameterization outperforms the fixed filter bank (MFCC_HP + FC). In this case, the relative error reduction with respect to MFCC_HP + FC is around 12.0% at segment level and 12.3% at target event level when Δ parameters are not considered and around 13.7% at segment level and 15.9% at target event level when Δs are included. In addition, in this latter case, the differences in performance are statistically significant. This result shows that the filters learned by NMF are capable to capture the dynamical structure of the cepstral coefficients, producing a filter bank more suitable for AEC than the fixed one.

Table 3: *Classification rate [%] for different number of filters in the filter bank extracted by NMF.*

Short-time features	Number of filters	Classif. rate (segment) [%]	Classif. rate (event) [%]
MFCC_HP	4	74.39 ± 0.97	79.29 ± 1.73
	6	74.02 ± 0.97	79.19 ± 1.73
	8	73.69 ± 0.98	78.99 ± 1.74
	10	73.65 ± 0.98	79.09 ± 1.73
MFCC_HP + Δ	4	76.15 ± 0.95	80.15 ± 1.70
	6	75.51 ± 0.96	78.37 ± 1.76
	8	74.70 ± 0.97	76.20 ± 1.82
	10	73.99 ± 0.98	74.36 ± 1.86

4.4. Experiments with different number of filters in the NMF-based filter bank

For either type of feature set, MFCC_HP or MFCC_HP + Δ, experiments were performed considering 4, 6, 8 and 10 bands in the NMF-based filter bank. Table 3 contains the corresponding classification rates as well as the corresponding 95 % confidence intervals.

For MFCC_HP, results vary with the number of filters, although the differences are rather small and not statistically significant. However, for MFCC_HP + Δ, the classification rate decreases along the number of frequency bands, suggesting that 4 filters are enough for representing the temporal behaviour of the short-time features (especially for the Δ parameters).

5. Conclusions

In this paper, we have presented a new front-end for AEC based on the combination of FC features and NMF. In particular, NMF is used for the unsupervised learning of the filter bank which captures the most relevant temporal behaviour in the short-time features. From the resulting NMF-based filter bank, we have observed that low modulation frequencies are more important than the high ones for distinguishing between different acoustic events. The experiments have shown that the features obtained with this method achieve significant improvements in the classification performance of a SVM-based AEC system in comparison with the baseline FC parameters.

For future work, we plan to extend our method in two directions: the design of a different filter bank for each dimension of the short-time features and the development of a semisupervised version for AEC.

6. Acknowledgements

This work has been partially supported by the Spanish Government grants TSI-020110-2009-103, IPT-120000-2010-24 and TEC2011-26807. Financial support from the Fundación Carolina and Universidad Católica San Pablo, Arequipa (Jimmy Ludeña-Choez) is thankfully acknowledged.

7. References

[1] A. Temko and C. Nadeu, "Classification of acoustic events using SVM-based clustering schemes," *Pattern Recognition*, vol. 39, pp. 684–694, 2006.

[2] C. Zieger, "An HMM based system for acoustic event detection," *Lecture Notes in Computer Science (LNCS), Springer*, pp. 338–344, 2008.

[3] X. Zhuang, X. Zhou, M. A. Hasegawa-Johnson, and T. S. Huang, "Real-world acoustic event detection," *Pattern Recognition Letters*, vol. 31, pp. 1543–1551, 2010.

[4] K. Kwangyoun and K. Hanseok, "Hierarchical approach for abnormal acoustic event classification in an elevator," *IEEE Int. Conf. AVSS*, pp. 89–94, 2011.

[5] J. Portelo, M. Bugalho, I. Trancoso, J. Neto, A. Abad, and A. Serralheiro, "Non speech audio event detection," *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 1973–1976, 2009.

[6] A. Meng, P. Ahrendt, and J. Larsen, "Temporal feature integration for music genre classification," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 15, pp. 1654–1664, 2007.

[7] D. Mejía-Navarrete, A. Gallardo-Antolín, C. Peláez, and F. Valverde, "Feature extraction assesment for an acoustic-event classification task using the entropy triangle," *INTERSPEECH*, pp. 309–312, 2011.

[8] Z. Zhang and B. Schuller, "Semi-supervised learning helps in sound event classification," *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 333–336, 2012.

[9] M. McKinney and J. Breebaart, "Features for audio and music classification," *International Society for Music Information Retrieval Conference (ISMIR)*, pp. 151–158, 2003.

[10] J. Arenas-García, J. Larsen, L. K. Hansen, and A. Meng, "Optimal filtering of dynamics in short-time features for music organization," *International Society for Music Information Retrieval Conference (ISMIR)*, pp. 290–295, 2006.

[11] J. Ludeña and A. Gallardo-Antolín, "NMF-based spectral analysis for acoustic event classification tasks," *Advances in Nonlinear Speech Processing (NOLISP 2013), Lecture Notes in Computer Science (LNCS)*, vol. 7911, pp. 9–16, 2013.

[12] D. Lee and H. Seung, "Algorithms for non-negative matrix factorization," *Nature*, vol. 401, pp. 788–791, 1999.

[13] K. Wilson, B. Raj, P. Smaragdis, and A. Divakaran, "Speech denoising using nonnegative matrix factorization with priors," *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 4029–4032, 2008.

[14] J. Ludeña and A. Gallardo-Antolín, "Speech denoising using nonnegative matrix factorization with kullback-leibler divergence and sparseness constraints," *Comm. in Computer and Information Science (CCIS)*, vol. 328, pp. 207–216, 2012.

[15] B. Schuller, F. Weninger, and M. Wollmer, "Non-negative matrix factorization as noise-robust feature extractor for speech recognition," *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 4562–4565, 2010.

[16] "FBK-Irst database of isolated meeting-room acoustic events," *ELRA Catalog no. S0296*.

[17] "UPC-TALP database of isolated meeting-room acoustic events," *ELRA Catalog no. S0268*.