Computer Speech & Language (2015). 30(1), 32-42.
DOI: http://dx.doi.org/10.1016/j.csl.2014.04.001

Feature Extraction Based on the High-Pass Filtering of Audio Signals for Acoustic Event Classification

Jimmy Ludeña-Choez[a,b], Ascensión Gallardo-Antolín[a,*]

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

Abstract

In this paper, we propose a new front-end for Acoustic Event Classification tasks (AEC). First, we study the spectral characteristics of different acoustic events in comparison with the structure of speech spectra. Second, from the findings of this study, we propose a new parameterization for AEC, which is an extension of the conventional Mel Frequency Cepstrum Coefficients (MFCC) and is based on the high pass filtering of the acoustic event signal. The proposed front-end have been tested in clean and noisy conditions and compared to the conventional MFCC in an AEC task. Results support the fact that the high pass filtering of the audio signal is, in general terms, beneficial for the system, showing that the removal of frequencies below 100-275 Hz in the feature extraction process in clean conditions and below 400-500 Hz in noisy conditions, improves significantly the performance of the system with respect to the baseline.

Keywords: Acoustic Event Classification, High-Pass Filtering, Auditory Filterbank

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

*Corresponding author.
Email address: gallardo@tsc.uc3m.es (Ascensión Gallardo-Antolín)

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 (Temko and Nadeu, 2006), automotive applications (Muller et al., 2008), mobile robots working in diverse environments (Chu et al., 2006) or surveillance systems (Clavel et al., 2005). 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).

Several front-ends have been proposed in the literature, some of them based on short-term features, such as Mel-Frequency Cepstral Coefficients (MFCC) (Temko and Nadeu, 2006; Zieger, 2008; Zhuang et al., 2010; Kwangyoun and Hanseok, 2011), log filterbank energies (Zhuang et al., 2010), Perceptual Linear Prediction (PLP) (Portelo et al., 2009), log-energy, spectral flux, fundamental entropy and zero-crossing rate (Temko and Nadeu, 2006). Other approaches are based on the application of different temporal integration techniques over these short-term features (Meng et al., 2007; Mejía-Navarrete et al., 2011; Zhang and Schuller, 2012). Finally, other relevant works in the literature have shown that the activation coefficients produced by the application of Non-Negative Matrix Factorizarization (NMF) on audio spectrograms can be used as acoustic features for AEC and other related tasks (Weninger et al., 2011; Cotton and Ellis, 2011). In order to distinguish between the different acoustic classes, some classification tools are then applied over these acoustic features, as for example, Gaussian Mixture Models (GMM) (Temko and Nadeu, 2006), Hidden Markov Models (HMM) (Cotton and Ellis, 2011), Support Vector Machines (SVM) (Temko and Nadeu, 2006; Mejía-Navarrete et al., 2011), Radial Basis Function Neural Networks (RBFNN) (Dhanalakshmi et al., 2008) and Deep Neural Networks (DNN) (Kons and Toledo, 2013). The high correlation between the performance of different classifiers suggests that the main problem is not the classification technique, but a design of a suitable feature extraction process for AEC (Kons and Toledo, 2013).

In fact, as pointed in (Zhuang et al., 2010), conventional acoustic features are not necessarily the more appropriate for AEC tasks because they have been design according to the spectral characteristics of speech which are quite different from the spectral structure of acoustic events. To deal with

this issue in (Zhuang et al., 2010) it is proposed a boosted feature selection method to construct a more suitable parameterization for AEC.

In this work, we follow a different approach. Based on the empirical study of the spectral characteristics of different acoustic events in comparison with the structure of speech spectra, we propose a new parameterization for AEC, which is an extension of the conventional MFCC and is based on the high pass filtering of the acoustic event signal. The proposed front-end has been tested in clean and noisy conditions achieving, in both scenarios, significant improvements with respect to the baseline system.

This paper is organized as follows: in Section 2 the main spectral characteristics of acoustic events are described. Section 3 is devoted to the explanation of the proposed parameterization. Section 4 describes the experiments and results to end with some conclusions and ideas for future work in Section 5.

2. Spectral Characteristics of Acoustic Events

As it is well known, the spectrograms of speech signals are characterized by the presence of a higher energy in the low-frequency regions of the spectrum. However, in general, non-speech sounds do not show this speech spectral structure. In fact, in many cases, their relevant spectral contents are located in other frequency bands, as it will be shown in the empirical study of the spectral characteristics of several AEs performed in this Section.

As an example, Figure 1 represents the spectrograms of two instances of the same acoustic event, *Phone ring*. Although it is possible to extract conclusions about the spectral nature of this AE by means of the visual inspection of these spectrograms, their high variability due in part to the intrinsic frequency characteristics of the acoustic event and in part to the presence of noise (microphone, environment noise, etc.), motivates us to use an automatic method such as Non-Negative Matrix Factorization (NMF) (Lee and Seung, 1999), which is capable of providing a more compact parts-based representation of the magnitude spectra of the AEs.

Given a nonnegative matrix $V_{\mathrm{e}} \in \mathbb{R}_{+}^{F \times T}$, where each column is a data vector (in our case, V_e contains the short-term magnitude spectrum of a set of audio signals), NMF approximates it as a product of two nonnegative matrices W_{e} and H_{e}, such that

$$V_e \approx W_e H_e \tag{1}$$

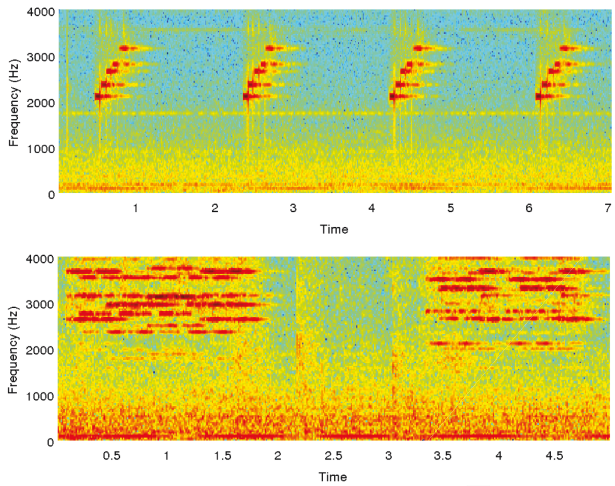

Figure 1: Spectrograms of two different instances of the acoustic event *Phone ring*.

where $W_{\mathrm{e}} \in \mathbb{R}_{+}^{F\times K}$ and $H_{\mathrm{e}} \in \mathbb{R}_{+}^{K\times T}$ and F, T and K represent frequency bins, frames and basis components, respectively. This way, each column of V_{e} can be written as a linear combination of the K building blocks (columns of W_{e}), weighted by the coefficients of activation located in the corresponding column of H_{e}. In this work, we are interested on retrieving the matrix W_{e} as it contains the building blocks or Spectral Basis Vectors (SBVs) which encapsule the frequency structure of the data in V_{e} (Smaragdis, 2004).

For each for the acoustic events considered, their SBVs were obtained by applying NMF to the corresponding matrix V_{e} composed by the short-term magnitude spectrum of a subset of the training audio files belonging to this particular class. The magnitude spectra were computed over 20 ms windows with a frameshift of 10 ms. In total, 364,214 magnitude spectral examples were used for performing NMF, which corresponds to approximately 60 minutes of audio. The NMF matrices were initialized using a multi-start initialization algorithm (Cichocki et al., 2009), in such a way that 10 pairs of uniform random matrices (W_{e} and H_{e}) were generated and the factorization producing the smallest euclidean distance between V_{e} and $(W_{\mathrm{e}}H_{\mathrm{e}})$ was chosen for initialization. Then, these initial matrices were refined by minimizing the KL divergence between the magnitude spectra V_{e} and their corresponding factored matrices $(W_{\mathrm{e}}H_{\mathrm{e}})$ using an iterative scheme and the learning rules proposed in (Lee and Seung, 1999) until the maximum number of iterations (in our case, 200) was reached.

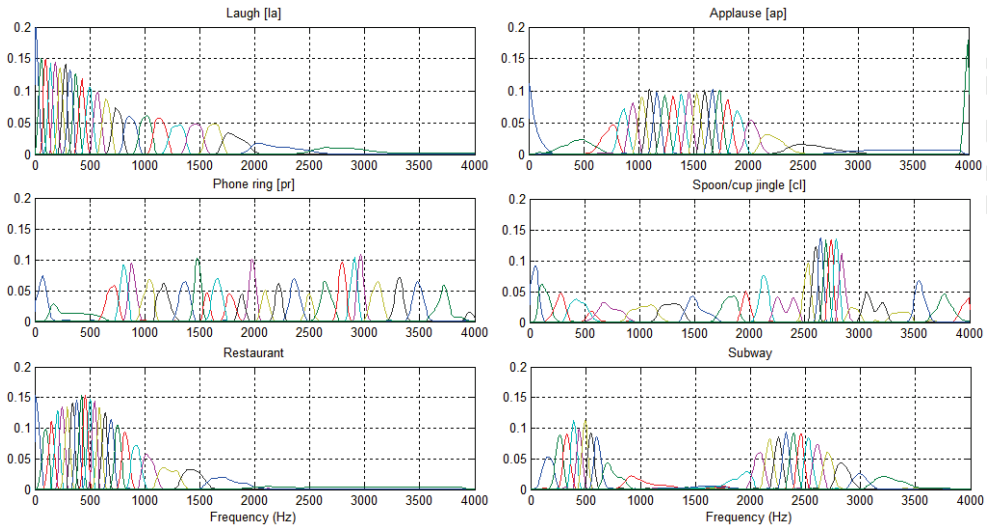

Figure 2: Spectral Basis Vectors (SBVs) for different acoustic events and types of noise.

The number of basis vectors K was set taking into account a trade-off between an accurate reconstruction of the magnitude spectra (i. e. the average approximation error between V_e and $(W_e H_e)$ computed over all AEs) and a good visualization of the SBVs. In particular, we used $K = 23$ which corresponds to the case in which the relative change in the average approximation error between two successive numbers of SBVs is less than 2%. It is also worth mentioning that when the number of basis vectors increases, NMF tends to place more and smaller bands in the areas of the spectrum with high energy (i. e. provides more resolution in these regions) and therefore reduces the overall reconstruction error (Bertrand et al., 2008). Nevertheless, for the purpose of this analysis, a larger value of K does not provide relevant information and produces a worse visualization of the SBVs.

Figure 2 represents the 23 SBVs of four different non-speech sounds (*Laugh*, *Applause*, *Phone ring* and *Spoon/cup jingle*) and two different kind of noises (*Restaurant* and *Subway*). From this figure, the following observations can be extracted:

- The spectral content of the AEs are very different each other, presenting, in general, relevant components in medium-high frequencies. As it is well-known that the spectral components of speech are concentrated in low frequencies, it is possible to infer that the parameterizations de-

signed for speech (as the conventional MFCC) are not suitable enough for representing non-speech sounds.

- In all cases, low frequency components are presented to a greater or lesser extent, so this part of the spectrum seems not to very discriminative when comparing different types of AEs.

- Comparing the SBVs of the non-speech sounds, it can be observed that large differences can be found in the medium-high part of the spectrum, suggesting that these frequency bands are more suitable (or at least, they can not be negligible) than the lower part of the spectrum for discriminating between different acoustic events.

- Different environment noises present very different spectral characteristics. For example, in the case of *Restaurant*, most of the frequency content is located in the band below 1 kHz, whereas the SBVs of the *Subway* noise are distributed in two different regions of the spectrum: a low frequency band below 750 Hz and a medium-high band of frequencies between 2 and 3 kHz. The analysis of other kind of noises (*Airport*, *Babble*, *Train* and *Exhibition Hall*) yields to similar observations. This way, the distortion produced over the AE signals due to the presence of additive noise will vary considerably depending of the nature of the noise. As a consequence of this fact, some noises will be presumably more harmful than others, producing more noticeable degradations in the performance of the AEC system.

3. Feature extraction for AEC derived from the high-pass filtering of the acoustic event signal

The observation of the SBVs of the different acoustic events shown in Section 2 motivated us to derive an extension of the conventional MFCC more suitable for AEC. As it is well known, MFCC is the most popular feature extraction procedure in speech and speaker recognition and also in audio classification tasks. The basic idea behind the new front-end is to take explicitly into account the special relevance of certain frequency bands of the spectrum into the feature extraction procedure through the modification of the characteristics of the conventional mel-scaled auditory filterbank.

One of the main conclusions drawn from the empirical study in Section 2 is that medium and high frequencies are specially useful for discriminating

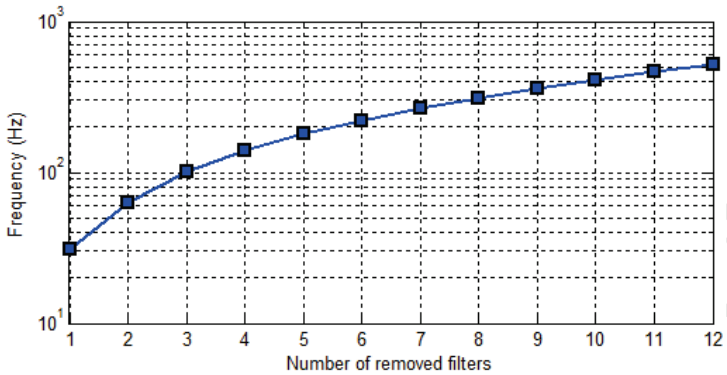

Figure 3: Upper frequency of the stopband vs. number of removed filters for the Mel scale.

between different acoustic events. For this reason, this band of frequencies should be emphasized in some way into the parameterization process. This can be accomplished by high-pass filtering the short-term frames of the signal (using the appropriate filter) prior to the application of the auditory filterbank and the cepstral parameters computation. However, in this work, we have adopted a straightforward method which consists of modifying the auditory filterbank by means of the explicit removal of a certain number of the filters placed on the low frequency region of the spectrum. In Figure 3 it can be observed the upper frequency of the complete stopband as a function of the number of removed filters in the auditory filterbank for the Mel scale.

In practice, this procedure consists of setting to a small value the energies corresponding to the outputs of the low-pass filters which are required to be removed. This threshold must be different to zero in order to avoid numerical problems with the logarithm, being, in our particular case, equal to 2^{-52} (the value of the roundoff level *eps* in the programming language Matlab).

Once the high-pass filtering is carried out following the procedure previously described and the remaining log filterbank 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[1]. Finally, it is applied a temporal feature integration technique which consists of dividing

[1]Another alternative to this method was considered in which the cepstral coefficients were obtained by applying the logarithm and the DCT exclusively on the outputs of the non-removed filters. The first method was finally adopted in this work because a preliminary experimentation showed that it outperformed this second approach.

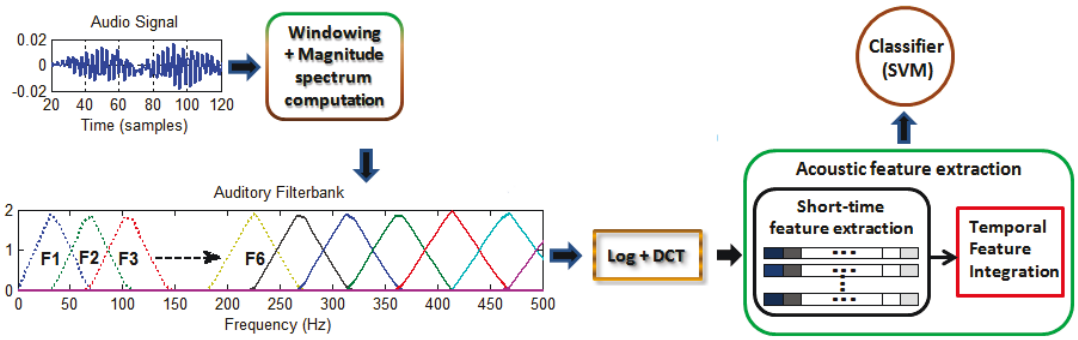

Figure 4: Block diagram of the proposed front-end.

the sequence of cepstral coefficients into sliding windows of several seconds length and computing the statistics of these parameters (in this case, mean, standard deviation and skewness) over each window. These segment-based parameters are the input to the acoustic event classifier, which is based on Support Vector Machines (SVM). The whole process is summarized in Figure 4.

4. Experiments

4.1. Database and Experimental Protocol

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 (Zhuang et al., 2010) and it is shown in Table 1. Audio files were obtained from different sources: websites[2], the FBK-Irst database[3] (FBK-Irst, 2009) and the UPC-TALP database[4] (UPC-TALP, 2008) and were converted to the same format and sampling frequency (8 kHz). The total number of segments of 2 seconds length (the window size used for the segment-based features computation as indicated in Subsection 4.2) in the whole database is 7,775.

[2]http://www.freesound.org/
[3]http://catalog.elra.info/product_info.php?products_id=1093
[4]http://catalog.elra.info/product_info.php?products_id=1053

Figure 5 shows the histogram of the number of segments per target event in the database, being the average about 3.75 segments.

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

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.

For the experiments in noisy conditions, the original audio recordings were contaminated with six different types of noise (*Airport*, *Babble*, *Restaurant*, *Train*, *Exhibition Hall* and *Subway*) obtained from the AURORA framework (Pearce and Hirsch, 2000) at SNRs from 0 dB to 20 dB with 5 dB step. In order to calculate the amount of noise to be added to the clean recordings, the audio and noise powers were calculated following the procedure indicated in (Steeneken, 1991), which takes into account the non-stationary characteristics of the signals.

The AEC system is based on a one-against-one SVM with RBF kernel on normalized features (Mejía-Navarrete et al., 2011). The system was developed using the LIBSVM software (Chang and Lin, 2011). Concerning SVM training, for each one of the subexperiments, a 5-fold cross validation was used for computing the optimal values of the RBF kernel parameters using

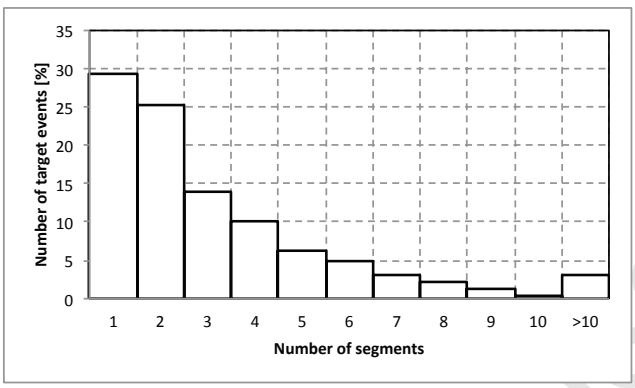

Figure 5: Histogram of the number of segments per target event for the database used in the experimentation.

clean data (i.e., these parameters were not optimized for noisy conditions). In the testing stage, as the SVM classifier is fed with segmental-based features computed over sliding windows, the classification decisions are made at segment level. In order to obtain a decision for the whole instance (target event level), the classifier outputs of the corresponding windows are integrated using a majority voting scheme, in such a way that the most frequent label is finally assigned to the whole recording (Geiger et al., 2013).

4.2. Experiments in clean conditions

This set of experiments were carried out in order to study the performance of the proposed front-end in clean conditions (i.e. when no noise is added to the original audio files). For the baseline experiments, 12 cepstral coefficients (C_1 to C_{12}) were extracted every 10 ms using a Hamming analysis window of 20 ms long and a mel-scaled auditory filterbank composed of 40 spectral bands. Also, the log-energy of each frame (instead of the zero-order cepstrum coefficient) and the first derivatives (where indicated) were computed and added to the cepstral coefficients. The final feature vectors consisted of the statistics of these short-term parameters (mean, standard deviation and skewness) computed over segments of 2 s length with overlap of 1 s.

Table 2 and Table 3 show, respectively, 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) by varying the number of eliminated low frequency bands in the auditory filterbank. Results for the baseline systems (when no frequency

Table 2: Average classification rate [%] (segment) in clean conditions.

Param.	Number of Eliminated Filters												
	Base.	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
CC	75.10	77.47	77.66	77.58	77.63	78.16	76.95	78.11	76.87	76.12	77.23	77.23	76.10
CC+ΔCC	77.57	79.43	79.45	79.22	79.36	79.07	79.20	79.55	79.41	78.47	77.81	78.77	78.55

Table 3: Average classification rate [%] (target event) in clean conditions.

Param.	Number of Eliminated Filters												
	Base.	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
CC	81.07	82.28	82.04	82.42	82.42	81.89	81.31	83.20	81.27	80.78	80.69	81.75	79.72
CC+ ΔCC	81.41	82.62	83.39	83.58	83.49	83.15	82.38	82.71	82.81	80.06	81.12	81.55	81.22

bands are eliminated) are also included. Both tables contain the classification rates for two different set of acoustic parameters, CC (cepstral coefficients + log-energy) and CC+ΔCC (cepstral coefficients + log-energy + their first derivatives).

As can be observed for the CC parameterization, the high pass filtering of the acoustic event signal outperforms the baseline, being the improvements more noticeable when the number of eliminated filters varies from 3 to 7. From Figure 3, it can be seen that these ranges of eliminated filters roughly correspond to a stopband from 0 Hz to 100-275 Hz. In particular, the best performance is obtained when the seven first low frequency filters are not considered in the cepstral coefficients computation. In this case, the difference in performance at segment level with respect to the baseline is statistically significant at 95% confidence level. The relative error reduction with respect to the corresponding baseline is around 12.1% at segment level and around 11.2% at target event level.

Similar observations can be drawn for the CC+ΔCC parameterization: best results are obtained when low frequencies (below 100-275 Hz) are not considered in the feature extraction process. When comparing to CC for the case in which the first 7 filters are eliminated, it can be observed that CC+ΔCC achieves an improvement about 1.4% absolute and a decrement around 0.5% absolute at, respectively, segment and target event level over CC. However, these differences are not statistically significant.

Other frequency scales (in particular, ERB and Bark) have been experimented observing, as is expected, a similar behaviour than the Mel scale with

respect to the elimination of low frequency bands. Nevertheless, the Mel scale produces slightly better results than ERB and Bark. More details about these experiments can be found in (Ludeña-Choez and Gallardo-Antolín, 2013).

In order to perform a more detailed analysis of the AEC system performance, we have analysed the confusion matrices produced by the baseline and the proposed front-end. As an example, Figures 6 (a) and (b) show, respectively, the confusion matrices at segment level for the baseline CC+ΔCC parameters and the modified version of this parameterization in which the first 7 low frequency filters are removed. In both tables, the columns correspond to the correct class, the rows to the hypothesized one and the values in them are computed as the average over the six folds. As can be observed, in the baseline system the less confusable classes (with a classification rate greater than 80%) are *Applause*, *Keyboard typing*, *Laugh*, *Paper work* and *Phone ring*, whereas the highest confusable ones are *Cough*, *Chair moving*, *Door knock* and *Spoon/cup jingle*. In particular, 23% of the *Cough* instances are classified as *Laugh* and 12% of the *Chair moving* and *Door knock* instances are assigned to the class *Steps*. It is worth mentioning the large amount of confusions between the human vocal-tract non-speech sounds (i. e. *Cough* and *Laugh*) which has been previously reported in the literature (Temko and Nadeu, 2006). In the proposed front-end, the recognition rates of all the acoustic classes increase with the exception of *Cough* and *Key jingle*. The acoustic events which are better classified are the same than in the baseline, whereas there are only two AEs with a classification rate less than 70% (*Cough* and *Chair moving*). This is because classes *Door knock* and *Spoon/cup jingle* reduce significantly their amount of confusions in comparison to the baseline.

4.3. Experiments in noisy conditions

In order to study the impact of noisy environments on the performance of the AEC system, several experiments were carried out with six different types of noise (*Airport*, *Babble*, *Restaurant*, *Train*, *Exhibition Hall* and *Subway*) at SNRs from 0 dB to 20 dB with 5 dB step. For the sake of brevity, we only report in this subsection the results for the baseline and for the proposed front-end in the case of CC+ΔCC parameters.

Figure 7 represents, for each noise, the average of the relative error reduction with respect to the baseline (noisy conditions without high-pass filtering of the audio signal) computed across the SNRs considered (0 dB to 20 dB

	1	2	3	4	5	6	7	8	9	10	11	12
1	90,40	0,51	0,00	0,25	0,53	0,12	0,44	0,54	0,00	0,15	0,00	1,07
2	0,00	67,95	3,09	3,56	1,60	2,04	7,06	0,54	0,41	1,04	6,67	0,00
3	0,62	1,79	65,73	2,54	2,67	0,54	1,43	0,76	0,68	5,06	3,23	0,71
4	0,00	0,51	2,25	68,19	3,73	1,26	0,11	0,11	0,27	5,95	0,22	0,71
5	0,00	0,51	0,56	5,85	75,20	0,72	0,33	0,76	1,08	1,49	1,51	0,89
6	0,00	3,08	4,49	1,27	4,27	80,54	2,76	7,34	2,71	4,02	6,24	2,84
7	8,05	23,33	4,78	4,07	4,80	0,66	84,23	2,37	4,74	1,64	3,44	2,31
8	0,62	1,03	2,25	0,51	1,33	11,71	1,54	80,58	3,11	4,32	1,29	12,08
9	0,31	0,26	1,40	0,00	1,33	0,42	1,32	1,83	84,57	0,45	3,66	1,95
10	0,00	0,00	12,36	12,21	3,20	0,96	0,22	1,73	0,14	74,26	5,81	2,13
11	0,00	1,03	0,56	0,00	0,80	0,42	0,11	0,76	1,22	1,04	64,95	2,13
12	0,00	0,00	2,53	1,53	0,53	0,60	0,44	2,70	1,08	0,60	3,01	73,18

(a)

	1	2	3	4	5	6	7	8	9	10	11	12
1	93,50	0,26	0,00	1,02	0,80	0,00	0,33	0,54	0,54	0,15	0,43	0,89
2	0,31	65,64	2,25	3,56	1,60	0,78	5,73	0,76	0,27	0,60	3,66	0,53
3	0,62	1,54	68,26	1,27	2,13	0,18	1,65	0,86	0,14	2,83	1,94	0,71
4	0,00	2,05	3,93	74,55	4,80	0,42	0,66	0,11	0,00	5,95	0,22	0,00
5	0,00	2,82	1,97	5,85	76,00	0,78	0,44	0,86	0,95	1,64	1,72	1,24
6	0,00	2,56	4,21	0,51	2,13	83,00	2,21	6,69	3,11	2,98	5,38	4,97
7	4,64	22,05	3,93	4,07	3,20	0,90	85,01	0,76	2,30	1,64	5,81	1,60
8	0,31	1,28	1,97	0,25	1,87	9,61	1,76	81,55	2,30	4,61	2,15	9,59
9	0,31	1,03	1,40	0,25	2,93	0,60	0,66	2,16	86,87	0,89	4,52	1,60
10	0,31	0,26	8,99	7,38	2,93	1,68	0,44	1,73	0,14	77,38	1,08	4,97
11	0,00	0,00	0,28	0,76	0,80	1,08	0,66	0,32	2,57	0,60	70,75	1,78
12	0,00	0,51	2,81	0,51	0,80	0,96	0,44	3,67	0,81	0,74	2,37	72,11

(b)

Figure 6: Confusion matrices [%] at segment level for the CC+ΔCC parameterization: (a) Baseline; (b) Front-end with the first 7 low frequency filters removed.

with 5 dB step) as a function of the number of removed low frequency filters at both, segment and target event level. The mean of the relative error reduction over all noises and SNRs is also indicated. In order to observe in more detail the behaviour of the AEC system with respect to different SNRs and noises, we also show Table 4 which contains the classification rates at segment level for the baseline and for the proposed front-end at several selected SNRs (20, 10 and 0 dB) for the six noises evaluated and the range of number of eliminated filters from 7 to 12.

Although all the evaluated noises produce a dramatic decrease in the classification rates, results in Table 4 suggest that each type of noise has a different effect over the system performance, being some noises (*Airport*, *Babble*, *Restaurant* and *Train*) less harmful than others (*Exhibition Hall* and *Subway*). This fact can be explained by analysing the spectral characteristics

of each noise. In Figure 2, the SBVs of the *Restaurant* and *Subway* noises are represented. In the first case, most of the spectral content is concentrated in low frequencies and, for this reason, this kind of noise affects to lesser extent the most relevant frequencies of the AEs. However, in the second case, part of the SBVs spreads over medium-high frequencies, and therefore, this noise can considerably mask the underlying AEs. Note that, on the one hand, the SBVs of *Airport*, *Babble*, *Restaurant* and *Train* noises are concentrated in the same range of frequencies than those of the *Restaurant* noise and, on the other hand, the SBVs of *Exhibition Hall* and *Subway* have also similar characteristics.

From results in Figure 7 it can be observed that with respect to the performance of the proposed front-end, for the *Airport*, *Babble*, *Restaurant* and *Train* noises, the classification rates at segment level improve considerably when the number of eliminated filters increases, specially for medium-low SNRs (see the corresponding rows labeled as "0 dB" and "10 dB" in Table 4). In this case, optimal values are obtained when frequencies below 400-500 Hz are not considered in the feature computation, which corresponds to the elimination of the 10-11 first low-frequency filters. Similar observations can be drawn by analysing the results at target event level. For the *Exhibition Hall* and *Subway* noises, results at segment level suffer a slight variation with respect to the number of removed filters, achieving smaller improvements with respect to the baseline than in the case of the other noises. At target level, the variations are greater, yielding to a decrease in the classification rate in most of the cases for these two noises.

Nevertheless, in average the proposed front-end, when 11 filters are removed, obtains relative error reductions with respect to the baseline (see Figure 7) about 7.81% at segment level and 7.78% at target event level.

Further experiments were carried out for other scales (Bark and ERB) and the CC parameterization. In all the cases, the results follow similar trends in comparison to the Mel scale and the CC+ΔCC parameters.

5. Conclusion

In this paper, we have presented a new parameterization method for acoustic event classification tasks, motivated by the study of the spectral characteristics of non-speech sounds. First, we have performed an empirical study of the spectral contents of different acoustic events, concluding that medium and high frequencies are specially important for the discrimination

Table 4: Average classification rate [%] (segment) for the CC+ΔCC parameterization and different noise types and SNRs.

Noise	SNR (dB)	Number of Eliminated Filters						
		Baseline	**7**	**8**	**9**	**10**	**11**	**12**
AIRPORT	**20 dB**	66.51	69.47	67.82	68.33	68.82	68.52	68.17
	10 dB	49.92	53.45	52.94	54.29	55.63	56.08	55.28
	0 dB	29.01	33.60	34.59	35.57	37.48	38.20	36.97
BABBLE	**20 dB**	67.09	68.77	68.45	68.89	68.94	67.94	67.33
	10 dB	52.27	56.45	56.69	56.99	57.44	56.85	56.08
	0 dB	27.59	36.92	35.74	37.16	39.12	37.69	35.28
RESTAURANT	**20 dB**	67.43	69.40	68.89	68.89	69.22	68.8	68.62
	10 dB	53.09	57.14	56.97	57.34	57.26	57.32	56.69
	0 dB	25.65	37.35	37.91	38.22	38.22	38.65	36.72
TRAIN	**20 dB**	71.18	72.92	72.82	72.8	72.74	72.27	72.68
	10 dB	58.69	61.72	61.67	62.91	62.9	63.44	63.27
	0 dB	40.46	45.81	45.88	46.40	46.7	47.32	46.83
EXHIBITION HALL	**20 dB**	58.00	57.68	57.13	58.01	58.35	57.76	56.49
	10 dB	42.66	42.98	42.41	43.46	44.02	43.65	42.50
	0 dB	22.00	23.45	24.02	23.83	24.66	24.99	23.61
SUBWAY	**20 dB**	56.90	56.38	55.97	56.23	56.68	56.10	55.32
	10 dB	39.88	41.51	40.94	40.82	40.53	41.30	40.40
	0 dB	19.34	23.06	23.81	23.94	22.74	24.75	24.41

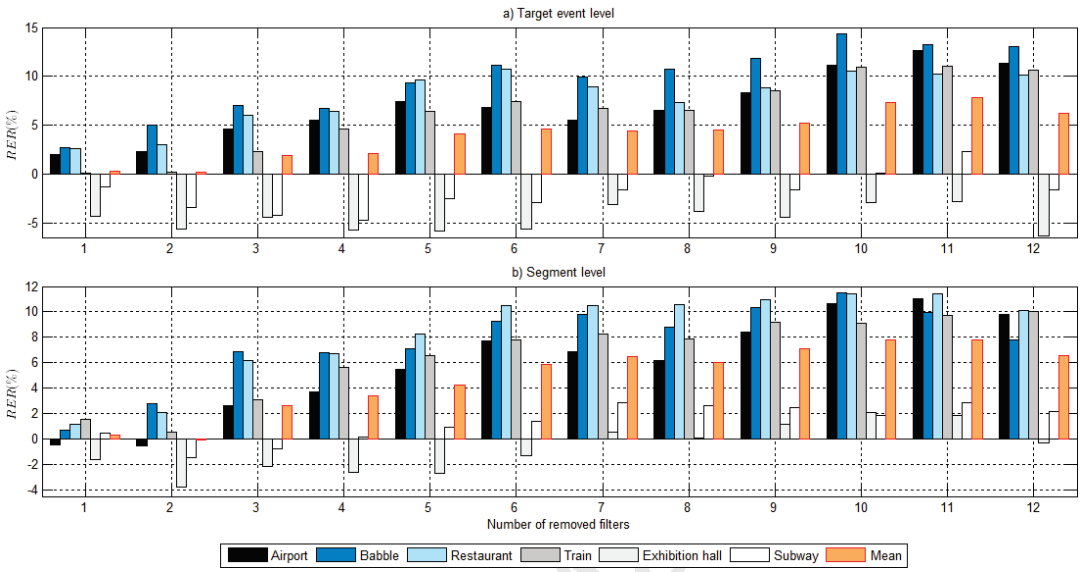

Figure 7: Relative error reduction [%] with respect to the baseline for the CC+ΔCC parameterization and the Mel scale: (a) at target event level; (b) at segment level.

between non-speech sounds. Second, from the findings of this study, we have proposed a new front-end for AEC, which is an extension of the MFCC parameterization and is based on the high pass filtering of the acoustic event signal. In practice, the proposed front-end is accomplished by the modification of the conventional mel-scaled auditory filterbank through the explicit elimination of a certain number of its low frequency filters.

The proposed front-end have been tested in clean and noisy conditions and compared to the conventional MFCC in an AEC task. Results support the fact that the high pass filtering of the audio signal is, in general terms, beneficial for the system, showing that the removal of frequencies below 100-275 Hz in the parameterization process in clean conditions and below 400-500 Hz in noisy conditions, improves significantly the performance of the system with respect to the baseline.

For future work, we plan to use feature selection techniques for automatically determining the most discriminative frequency bands for AEC. Other future lines include the unsupervised learning of auditory filterbanks by means of NMF and the use of the NMF activation coefficients as acoustic features for AEC.

Acknowledgments

This work has been partially supported by the Spanish Government grants 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.

References

Bertrand, A., Demuynck, K., Stouten, V., 2008. Unsupervised learning of auditory filter banks using non-negative matrix factorization. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP) , 4713–4716.

Chang, C.C., Lin, C.J., 2011. LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology 2, 27:1–27:27. Software available at `http://www.csie.ntu.edu.tw/ cjlin/libsvm`.

Chu, S., Narayanan, S., Kuo, C.C.J., Mataric, M.J., 2006. Where am I? Scene recognition for mobile robots using audio features. IEEE International Conference on Multimedia and Expo (ICME) .

Cichocki, A., Phan, A.H., Zdunek, R., 2009. Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-way Data Analysis and Blind Source Separation. Wiley, Chichester.

Clavel, C., Ehrette, T., Richard, G., 2005. Events detection for an audio-based surveillance system. IEEE International Conference on Multimedia and Expo (ICME) .

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

Dhanalakshmi, P., Palanivel, S., Ramalingam, V., 2008. Classification of audio signals using SVM and RBFNN. Expert Systems with Applications 36, 6069–6075.

FBK-Irst, 2009. FBK-Irst database of isolated meeting-room acoustic events. ELRA Catalog no. S0296 .

Geiger, J.T., Schuller, B., Rigoll, G., 2013. Large-scale audio feature extraction and SVM for acoustic scene classification. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics .

Kons, Z., Toledo, O., 2013. Audio event classification using deep neural networks. Conference of the International Speech Communication Association (INTERSPEECH) , 1482–1486.

Kwangyoun, K., Hanseok, K., 2011. Hierarchical approach for abnormal acoustic event classification in an elevator. IEEE Int. Conf. AVSS , 89–94.

Lee, D., Seung, H., 1999. Algorithms for non-negative matrix factorization. Nature 401, 788–791.

Ludeña-Choez, J., Gallardo-Antolín, A., 2013. NMF-based spectral analysis for acoustic event classification tasks. Advances in Nonlinear Speech Processing (NOLISP 2013), Lecture Notes in Computer Science 7911, 9–16.

Mejía-Navarrete, D., Gallardo-Antolín, A., Peláez, C., Valverde, F., 2011. Feature extraction assessment for an acoustic-event classification task using the entropy triangle. Conference of the International Speech Communication Association (INTERSPEECH) , 309–312.

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

Muller, C., Biel, J., Kim, E., Rosario, D., 2008. Speech-overlapped acoustic event detection for automotive applications. Conference of the International Speech Communication Association (INTERSPEECH) .

Pearce, D., Hirsch, H.G., 2000. The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions. Proc. Sixth International Conference on Spoken Language Processing (ICSLP/INTERSPEECH) , 29–32.

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

Smaragdis, P., 2004. Discovering auditory objects through nonnegativity constraints. Proc. Statistical and Perceptual Audio Processing .

Steeneken, H.J.M., 1991. Speech level and noise level measuring method. Technical Report. Document SAM-TN0-042. Esprit-SAM .

Temko, A., Nadeu, C., 2006. Classification of acoustic events using SVM-based clustering schemes. Pattern Recognition 39, 684–694.

UPC-TALP, 2008. UPC-TALP database of isolated meeting-room acoustic events. ELRA Catalog no. S0268 .

Weninger, F., Schuller, B., Wollmer, M., Rigoll, G., 2011. Localization of non-linguistic events in spontaneous speech by Non-Negative Matrix Factorization and Long Short-Term Memory. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP) , 5840–5843.

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

Zhuang, X., Zhou, X., Hasegawa-Johnson, M.A., Huang, T.S., 2010. Real-world acoustic event detection. Pattern Recognition Letters 31, 15431551.

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