SYNTHETIC SOUND EVENT DETECTION BASED ON MFCC

J.M. Gutiérrez-Arriola, R. Fraile, A. Camacho, T. Durand, J.L. Jarrín, S.R. Mendoza

Escuela Técnica Superior de Ingeniería y Sistemas de Telecomunicación
Universidad Politécnica de Madrid

ABSTRACT

This paper presents a sound event detection system based on mel-frequency cepstral coefficients and a non-parametric classifier. System performance is tested using the training and development datasets corresponding to the second task of the DCASE 2016 challenge. Results indicate that the most relevant spectral information for event detection is below 8000 Hz and that the general shape of the spectral envelope is much more relevant than its fine details.

Index Terms— Sound event detection, spectral envelope, cepstral analysis

1. INTRODUCTION

Automatic sound event detection is a rather recent research issue and any advance related to it may impact a variety of application fields [1]. Probably, the most intuitive approach to sound description for event detection consists in parameterising its spectrum. Specifically, mel-frequency cepstral coefficients (MFCC) provide a low-dimensional procedure for coding the shape of the spectral envelope that has been successfully applied to speech processing tasks such as speaker verification [2] or laryngeal pathology detection [3]. In fact, this type of coefficients has also been applied to sound event detection [1, 4, 5]. Yet, it is known that sound perception not only works in spectral domain, but also in temporal domain [6]. Such temporal dimension may be included in sound event detection by different means such as calculating MFCC derivatives, training hidden Markov models for classification, or both [1, 4].

When it comes to detecting several sound events happening simultaneously, proposed approaches include decomposition of sound spectra in several components prior to classification [7], adding complexity to the classification stage to allow for multiple event detection [1], or combinations of both [8].

In our view, *a priori* decomposition of sound spectra in several components is problematic, since the addition of two signals in temporal domain does not necessarily result in the addition of their power spectra. For this reason, we approach the problem by directly coding the spectrum of the recorded signal using MFCC. The temporal dimension of the event detection problem is acknowledged by calculating the first derivatives of MFCCs and by splitting the sound signal into frames before processing. In this work, we concentrate on the design of the datasets and the signal analysis; consequently no assumption is made regarding the distribution of the calculated signal parameters. For this reason, a non-parametric classifier is chosen.

This work has been partially financed by the Spanish Government, through project grant number TEC2012-38630-C04-01.

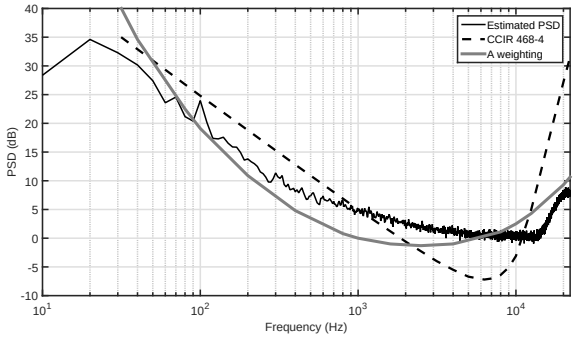

Figure 1: Power spectral density (PSD) of synthetic noise, estimated from a 6.5 second-length fragment using the Welch method [9]. For reference purposes, 'A Weighting' and 'CCIRR 468-4' curves [10] have also been plotted.

2. MATERIALS

Audio recordings were provided by IRCCYN, École Centrale de Nantes. They correspond to 11 sound event types (see Tab.1) recorded in a quiet environment, using a condenser microphone (AT8035, manufactured by Audio-Technica) connected to a portable recorder (H4n, manufactured by Zoom). Audio signals were sampled at 44.1 kHz and recorded with a single microphone (monophonic recordings). The microphone pass band ranges from 40 to 20,000 Hz.

20 events from each type were recorded, hence resulting 220 recordings each one containing a single sound event. For validation purposes, an additional dataset was built using the previous 220 recordings as a basis. This consists of 18 recordings with 2 minute durations. These were obtained by combining some of the single-event recordings into a single file and adding noise recorded in an independent session. Overlapping between events was allowed in 50% of the resulting files. Noise was approximately grey (Fig.1) and several levels of event-to-background ratio (EBR) were allowed: -6, 0 and 6 dB.

3. SIGNAL ANALYSIS

3.1. Inspection of sound spectra

Fig.2 depicts the estimated spectra, averaged for each type of event. While some types have distinct spectral envelope shapes, such as key drops or phone ringing, there are others for which the spectral envelopes are similar. This is especially the case of cough, throat clearing, laughter and speech, since all these sounds are produced as outputs of the same acoustic filter: the human vocal tract. Such fact

Type#	Type name	Event
1	Clearthroat	Throat clearing
2	Cough	Cough
3	Doorslam	Door slam
4	Drawer	Drawer sliding
5	Keyboard	Typewritting
6	Keys	Keys dropping on a desk
7	Knock	Knocking on a door
8	Laughter	Laughter
9	Pageturn	Paper page turning
10	Phone	Phone ringing
11	Speech	French speech
12	Back	Background noise

Table 1: Event types. Recordings corresponding to the 12^{th} type (*back*) were obtained by cutting out event-free segments from the validation dataset.

suggests that parameterisation schemes based only on estimating the average spectral envelope are likely to have poor performances.

From another point of view, all spectra exhibit a decay at frequencies above 13 kHz. However, the power spectral density of background noise (*back* type in Fig.2) grows from 13 to 22 kHz, as also shown in Fig.1. As a consequence, the EBR above 13 kHz is a decreasing function of frequency.

3.2. Parameter computation

Considering aforementioned characteristics of the target sound event spectra, we propose a parameterisation scheme based on the calculation of mel-frequency cepstral coefficients (MFCCs) and their derivatives. The proposed signal processing scheme comprises the next stages:

1. *Windowing*: Each digital audio signal is first normalised to yield a unit power discrete-time signal $x\,[n]$, composed by N samples ($n = 0 \ldots N-1$). This signal is segmented in speech frames of length equal to L samples through multiplication by a framing window $w\,[n]$:

$$x_p\,[n] = x\,[n + p\,(L - l_0)] \cdot w\,[n] \tag{1}$$

where l_0 is the number of overlapping samples between consecutive frames and p is the frame index.

2. *Fourier transform*: From each speech frame, the short-term Discrete Fourier Transform (stDFT) is computed as:

$$X_p\,(k) = \sum_{n=0}^{L-1} x_p\,[n] \cdot e^{-j\frac{2\pi nk}{N_{\text{DFT}}}} \tag{2}$$

where N_{DFT} is the number of points of the stDFT, $N_{\text{DFT}} \geq L$ and $k = 0 \ldots N_{\text{DFT}} - 1$.

The absolute frequency value that corresponds to each stDFT coefficient is:

$$f_k = \begin{cases} f_{\text{s}} \cdot \frac{k}{N_{\text{DFT}}} & \text{if } k \leq \frac{N_{\text{DFT}}}{2} \\ f_{\text{s}} \cdot \frac{k - N_{\text{DFT}}}{N_{\text{DFT}}} & \text{if } k > \frac{N_{\text{DFT}}}{2} \end{cases} \tag{3}$$

being f_{s} the sampling frequency.

3. *Mel distortion*: After the computation of the stDFT, the next step is frequency distortion in spectral domain. This is made according to [11, chap. 2]:

$$f_k^{\text{mel}} = \text{sgn}\,[f_k] \cdot 2595 \cdot \log_{10}\left(1 + \frac{|f_k|}{700}\right) \tag{4}$$

4. *Mel spectrum smoothing*: This is done by integrating the energy present in the spectrum of the processed speech frame along a set of pre-defined mel-frequency bands. These are M equal-width bands linearly distributed between $f_{\text{MIN}}^{\text{mel}}$ and $f_{\text{MAX}}^{\text{mel}}$ with 50% overlap between consecutive bands. Each one is characterised by its centre mel frequency and its width. The i^{th} centre frequency is

$$f_{\text{c},i}^{\text{mel}} = f_{\text{MIN}}^{\text{mel}} + \left(f_{MAX}^{\text{mel}} - f_{\text{MIN}}^{\text{mel}}\right) \cdot \frac{i}{M+1} \tag{5}$$

where $i = 1 \ldots M$. Thus, each band covers the range $I_i^{\text{mel}} = \left[f_{\text{c},i-1}^{\text{mel}}, f_{\text{c},i+1}^{\text{mel}}\right]$, yielding bandwidth

$$\Delta f^{\text{mel}} = 2 \cdot \frac{f_{MAX}^{\text{mel}} - f_{\text{MIN}}^{\text{mel}}}{M+1} \tag{6}$$

Integration along bands is commonly done using triangular windows [12, chap. 6]. Thus, the result for each band is:

$$\tilde{X}_p\,(i) = \frac{1}{A_i} \cdot \sum_{f_k^{\text{mel}} \in I_i^{\text{mel}}} \left| \frac{f_k^{\text{mel}} - f_{\text{c},i-1}^{\text{mel}}}{\frac{\Delta f^{\text{mel}}}{2}} - 1 \right| |X_p\,(k)| \tag{7}$$

where the normalising term A_i ensures that for each band the mean energy is computed without any bias:

$$A_i = \sum_{f_k^{\text{mel}} \in I_i^{\text{mel}}} \left| \frac{f_k^{\text{mel}} - f_{\text{c},i-1}^{\text{mel}}}{\frac{\Delta f^{\text{mel}}}{2}} - 1 \right| \tag{8}$$

5. *Transformation into cepstral domain*: The last step in MFCC computation is transformation of the afore-mentioned smoothed mel spectrum into cepstral domain. Such transformation can be realised by calculating the inverse DFT of the logarithm of the power spectrum [13]. Given that the speech signal is real-valued, it may be assumed that its spectrum is symmetric. Furthermore, if $\tilde{X}_p\,(0)$ is defined to be equal to 1, which simply means adding a constant value to the signal in temporal domain, then the power cepstrum of the mel-wrapped and spectrally smoothed signal can be written as:

$$\begin{aligned} \mathcal{X}_p\,[q] &= \tfrac{1}{2M+1} \textstyle\sum_{i=-M}^{M} \log\left(\tilde{X}_p\,(i)\right) e^{j\frac{2\pi i}{2M+1}q} \\ &= \tfrac{1}{M+\frac{1}{2}} \textstyle\sum_{i=1}^{M} \log\left(\tilde{X}_p\,(i)\right) \cos\left(\frac{\pi iq}{M+\frac{1}{2}}\right) \end{aligned} \tag{9}$$

The coefficients $\mathcal{X}_p\,[q]$ are called MFCC and they may be computed using an expression that resembles the discrete cosine transform (DCT) of the logarithm of the smoothed mel-wrapped spectrum of the speech frame $x_p\,[n]$. In fact, the original MFCC formulation [14] directly uses the second form of the DCT (DCT-2) [15, chap. 8]. Herein, (9) is preferred because it has a simpler relation to the DFT.

6. *Derivation*: Derivation of MFCC to obtain ΔMFCC is performed using a eighth-order discrete differentiating filter:

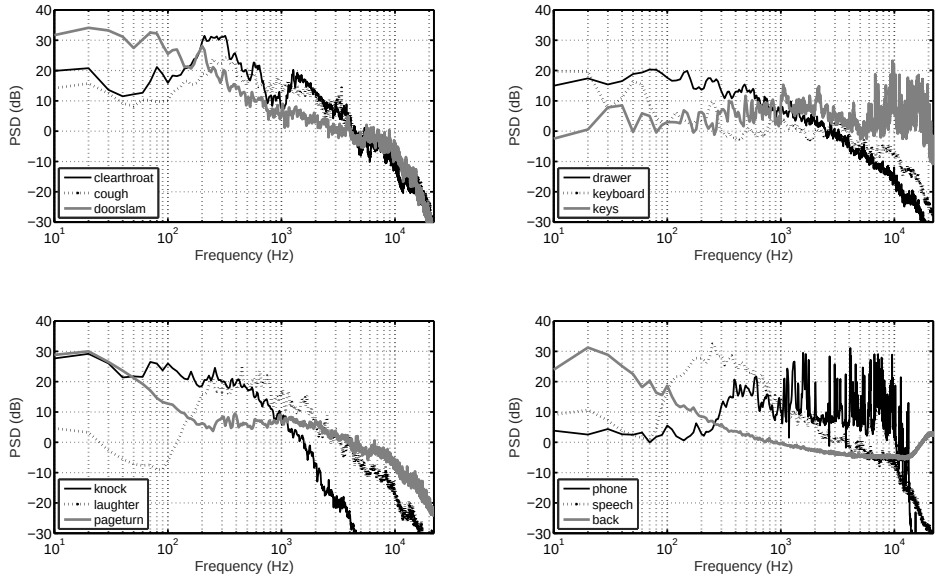

Figure 2: Average power spectral density for each type of event. Spectra have been averaged for all 20 recordings belonging to each type. Estimation has been carried out using the Welch method [9].

$$\begin{aligned}\Delta\mathcal{X}_p[q] &= \frac{1}{4}\mathcal{X}_{p-4}[q] - \frac{1}{3}\mathcal{X}_{p-3}[q] + \frac{1}{2}\mathcal{X}_{p-2}[q] \\ &- \mathcal{X}_{p-1}[q] + \mathcal{X}_{p+1}[q] - \frac{1}{2}\mathcal{X}_{p+2}[q] \\ &+ \frac{1}{3}\mathcal{X}_{p-3}[q] - \frac{1}{4}\mathcal{X}_{p+4}[q] \end{aligned} \quad (10)$$

4. CLASSIFICATION

The feature vectors describing sound frames that result from the previous signal analysis scheme have probability distributions with shapes that significantly differ between distinct sound events. For instance, the distributions for the *speech* and *keys* classes illustrated in Fig.3 present different shapes. From another point of view, it is known that for classification problems, the choice of classifier is much less relevant than the availability of as many data as possible [16]. For these reasons, a non-parametric discriminant approach based on the k-nearest-neighbours (kNN) rule [17] was selected.

5. POST-PROCESSING

Let $N_t(p)$ be the number of neighbours belonging to event type t assigned to the p^{th} sound frame by the kNN rule, therefore:

$$\sum_{t=1}^{12} N_t(p) = k \quad (11)$$

Then, a straightforward application of this classification rule would lead to assigning event type $\mathcal{T}(p)$ to the p^{th} sound frame such that:

$$\mathcal{T}(p) = \arg\max_t N_t(p) \quad (12)$$

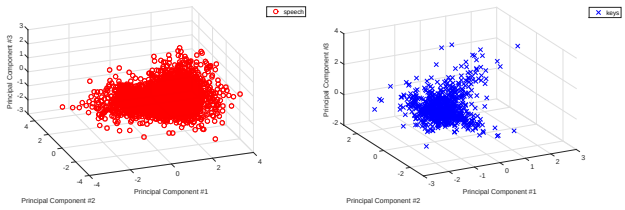

Figure 3: Distribution of frames belonging to *speech* (left) and *keys* (right) classes in the feature space defined by the three first principal components of the feature vectors including 15 MFCC + 15 ΔMFCC parameters.

However, the following procedure was used in order to smooth the effect of outlier frames:

1. Low-pass filtering of the number of neighbours by computing the local average using a sliding Hamming window w_{h}:

$$\tilde{N}_t(p) = \frac{\sum_{\Delta p=-P_1}^{P_1} N_t(p+\Delta p)\cdot w_{\text{h}}[\Delta p]}{\sum_{\Delta p=-P_1}^{P_1} w_{\text{h}}[\Delta p]} \quad (13)$$

2. Discarding events for which the filtered number of neighbours is below a certain threshold:

$$\hat{N}_t(p) = \begin{cases} \tilde{N}_t(p) & \text{if } \tilde{N}_t(p) \geq N_{\text{thres}} \\ 0 & \text{otherwise} \end{cases}, \; t = 1\ldots 11 \quad (14)$$

3. Assigning an event type to each frame, in case the smoothed number of neighbours corresponding to some class is above the threshold; otherwise, the frame is considered to belong to the *back* class:

Param.	Value	Explanation
L	1324	30 ms frames with $f_s = 44.1$ kHz
l_0	331	25% overlap between adjacent frames
$\omega[n]$		Hamming window
N_{DFT}	1324	Same as frame length
$f_{\mathrm{MIN}}^{\mathrm{mel}}$	62.63 mel	Corresponding to $f = 40$ Hz
$f_{\mathrm{MAX}}^{\mathrm{mel}}$	3582 mel	Corresponding to $f = 13$ kHz
M	40	
k	25	
P_1	5	200 ms filter length
N_{thres}	6.75	
P_2	1	Corresponding to 25 ms
$\Delta\tau_{\mathrm{MIN}}$	2 s	
T_{MIN}	300 ms	

Table 2: Parameter values for the reference system.

$$\mathcal{T}(p) = \begin{cases} \arg\max_t \hat{N}_t(p) & \text{if } \max_t \hat{N}_t(p) > 0 \\ 12 & \text{otherwise} \end{cases} \quad t = 1 \ldots 12 \tag{15}$$

4. Discarding events that are not detected in a minimum number of consecutive frames:

$$\tilde{\mathcal{T}}(p) = \begin{cases} \mathcal{T}(p) & \text{if } \sum_{\Delta p=-P_2}^{P_2} \mathcal{T}(p+\Delta p) = 2P_2+1 \\ 12 & \text{otherwise} \end{cases} \tag{16}$$

After classification of every sound frame, decision on the on/off times of sound events is made based on the next rules:

A. An event is considered to be formed by a set of consecutive frames corresponding to the same value of $\tilde{\mathcal{T}}(p)$. In such a case, the event type is defined by $\tilde{\mathcal{T}}(p)$ and its starting and ending times are defined by the central time instants of the first and last frames of the set, respectively.

B. Two events of the same type are merged into a single one if the time difference between the starting time of the second one and the ending time of the first one is less than a certain threshold $\Delta\tau_{\mathrm{MIN}}$. The resulting event duration is from the starting time of the first original event to the ending time of the second one.

C. A minimum event duration T_{MIN} is defined. If the duration of a given event is shorter, then its starting point is advanced and its ending point delayed so that its duration equals T_{MIN}.

6. EXPERIMENTS & RESULTS

The previously described system, with the parameter values summarised in Tab.2, was used as a reference and applied to the detection of sound events in the additional dataset described in section 2. Results for 20 MFCC + 20 ΔMFCC are summarised on the left column of Tab.3.

System performance can be significantly improved by building a training dataset with features as similar as possible to those of the validation dataset. In this case, if noise sequences extracted from the additional dataset are added to the 220 training recordings with the same levels of SNR as in the validation dataset, namely -6, 0 and 6 dB, and the resulting 660 sound signals are used as the new training dataset then the system performance can be significantly improved, as shown in the middle column of Tab.3.

Results in the right column of Tab.3 indicate that the most relevant information is concentrated below 8000 Hz ($f_{\mathrm{MAX}}^{\mathrm{mel}} = 2840$) and that it can be described using only 15 MFCCs plus their derivatives without any big loss of performance. This being a simpler configuration, a more robust performance is to be expected.

Last, it should be noted that the post-processing rule B implicitly allows event overlapping. In fact, detection performance for the recordings in the validation set with overlapped events ($F = 65.84\%$, $ER = 0.5030$ for the segment-based evaluation; $F = 59.18\%$, $ER = 0.6869$ for the event-based evaluation) is similar to the overall performance (Tab.3).

	Perform. Measure	Refer. System	Training with noise	8000 Hz 15 MFCC
Segment based	F	9.61%	70.06%	67.65%
	ER	0.9569	0.4706	0.4973
Event based	F	6.07 %	62.99%	60.22 %
	ER	1.059	0.6616	0.6902

Table 3: Event detection results in terms of F-score (F) and Error Rate (ER).

	Average	Clearthroat	Cough	Knock
F	34.2%	54.0%	25.0%	42.4%
ER	2.2537	0.7510	0.9053	1.8566
	Doorslam	**Drawer**	**Keyboard**	**Keys**
F	4.5%	33.1%	67.5%	12.5%
ER	3.0628	0.9033	0.5426	1.1156
	Laughter	**Pageturn**	**Phone**	**Speech**
F	47.6%	5.6%	71.3%	12.8%
ER	0.7647	0.9744	0.4793	13.4350

Table 4: Class average evaluation results (segment-based).

7. CONCLUSIONS

The reported results in sound event detection, obtained using a system based on MFCC parameters and a non-parametric classifier lead to two main conclusions. In the first place, system performance is critically affected by a proper selection of the sound recordings used for training the system. In this particular case, using recordings with noise levels similar to those in the testing set has allowed a significant improvement in performance. Secondly, the key spectral information for sound event detection seems to be concentrated below 8000 Hz. Additionally, the fact that 15 MFCCs provide almost the same performance as 20 MFCCs reveals that the essential information is in the overall shape of the spectral envelope and not in its fine details, be them either narrow peaks or narrow valleys.

APPENDIX: EVALUATION RESULTS

Performance of the proposed system (15 MFCCs; 40-8000 Hz) for the DCASE 2016 evaluation dataset is reported in [18]. The overall indicators for the segment-based evaluation were $ER = 2.0870$ and $F = 25.0\%$; for the event-based evaluations, they were $ER = 1.3064$ and $F = 25.7\%$. Per-class results are summarised in Tab.4.

8. REFERENCES

[1] T. Heittola, A. Mesaros, A. Eronen, and T. Virtanen, "Context-dependent sound event detection," *EURASIP J. Audio Speech Music Process.*, vol. 2013, no. 1, pp. 1–13, 2013.

[2] F. Bimbot, J. F. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-García, D. Petrovska-Delacrétaz, and D. A. Reynolds, "A tutorial on text-independent speaker verification," *EURASIP J. Adv. Signal Process.*, vol. 2004, no. 4, pp. 1–22, 2004.

[3] R. Fraile, N. Sáenz-Lechon, J. I. Godino-Llorente, V. Osma-Ruiz, and C. Fredouille, "Automatic detection of laryngeal pathologies in records of sustained vowels by means of mel-frequency cepstral coefficient parameters and differentiation of patients by sex," *Folia Phoniatr. Logop.*, vol. 61, no. 3, pp. 146–152, 2009.

[4] A. Diment, T. Heittola, and T. Virtanen, "Sound event detection for office live and office synthetic AASP challenge," in *Proc. IEEE AASP Challenge Detection Classif. Acoust. Scenes Events (WASPAA)*, 2013.

[5] E. Cakir, T. Heittola, H. Huttunen, and T. Virtanen, "Polyphonic sound event detection using multi label deep neural networks," in *Internat. Joint Conf. Neural Networks (IJCNN)*, 2015, pp. 1–7.

[6] P. Gómez-Vilda, J. M. Ferrández-Vicente, V. Rodellar-Biarge, A. Álvarez-Marquina, L. M. Mazaira-Fernández, R. Martínez-Olalla, and C. Muñoz-Mulas, "Detection of speech dynamics by neuromorphic units," in *Internat. Work-Conf. Interplay between Natural and Artificial Comput.* Springer, 2009, pp. 67–78.

[7] O. Dikmen and A. Mesaros, "Sound event detection using non-negative dictionaries learned from annotated overlapping events," in *IEEE Workshop Appl. Signal Process. Audio Acoust.* IEEE, 2013, pp. 1–4.

[8] T. Heittola, A. Mesaros, T. Virtanen, and A. Eronen, "Sound event detection in multisource environments using source separation," in *Workshop on Machine Listening in Multisource Environm.*, 2011, pp. 36–40.

[9] J. G. Proakis and D. G. Manolakis, *Digital Signal Processing: Principles Algorithms and Applications*. Macmillan Publishing Company, 1988.

[10] P. Skirrow, "Audio measurements and test equipment," in *Audio Engineers Reference Book*, M. Talbot-Smith, Ed. Focal Press, Oxford, 1999, ch. 3.6.

[11] X. D. Huang, A. Acero, and H. W. Hon, *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*. Prentice-Hall, 2001.

[12] J. R. Deller, J. G. Proakis, and J. H. L. Hansen, *Discrete-Time Processing of Speech Signals*. Macmillan Publishing Company, 1993.

[13] D. G. Childers, D. P. Skinner, and R. C. Kemerait, "The cepstrum: A guide to processing," *Proc. IEEE*, vol. 65, no. 10, pp. 1428–1443, 1977.

[14] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-28, no. 4, pp. 357–366, 1980.

[15] A. V. Oppenheim, R. W. Schafer, and J. R. Buck, *Discrete-Time Signal Processing*. Prentice-Hall, 1989, vol. 2.

[16] M. Banko and E. Brill, "Scaling to very very large corpora for natural language disambiguation," in *Proc. 39^{th} Annual Meeting Assoc. Computational Linguistics*, 2001, pp. 26–33.

[17] S. Theodoridis and K. Koutroumbas, *Pattern Recognition*. Academic Press - Elsevier, 2003.

[18] "Sound event detection in synthetic audio. Task results," Tampere University of Technology," DCASE, 2016. [Online]. Available: http://www.cs.tut.fi/sgn/arg/dcase2016/task-results-sound-event-detection-in-synthetic-audio[Visited: 04/08/2016]