GAMMATONE WAVELET FEATURES FOR SOUND CLASSIFICATION IN SURVEILLANCE APPLICATIONS

Xavier Valero, Francesc Alías

Grup de Recerca en Tecnologies Mèdia
La Salle - Universitat Ramon Llull, Barcelona, Catalonia, Spain
email: xvalero@salle.url.edu, falias@salle.url.edu, web: www.salle.url.edu

ABSTRACT

Sound can deliver highly informative data about the environment, which can be of particular interest for home-teleassistance and surveillance purposes. In the sound event recognition process, the signal parameterisation is a crucial aspect. In this work, we propose Gammatone-Wavelet features (GTW) by merging Wavelet analysis, which is well-suited to represent the characteristics of surveillance-related sounds, and Gammatone functions, which model the human auditory system. An experimental evaluation that consists of classifying a set of surveillance-related sounds employing Support Vector Machines has been conducted at different SNR conditions. When compared to typical Wavelet analysis with *Daubechies* mother function (DWC), the GTW features show superior classification accuracy both in noiseless conditions and noisy conditions for almost any SNR level. Finally, it is observed that the combination of DWC and GTW yields the highest classification accuracies.

Index Terms— Gammatone function, Wavelet analysis, audio classification, feature extraction, audio-based surveillance, Ambient Assisted Living.

1. INTRODUCTION

Traditionally, surveillance systems have been based on video information. However, significant improvements may be achieved by adding audio as input information [1], providing several advantages since they *i)* work in absence of light, *ii)* need cheaper sensors, *iii)* fix the image limitations associated to the blind spots, *iv)* preserve the personal privacy (no image is stored from people), *v)* enable simple alarm triggering for the emergency or police services, etc. [2]. Audio-based surveillance systems may be used for security purposes in many contexts, such as offices [1], metro stations [3] or home environments [4], [5]. In addition, the technology might be applied to Ambient Assisted Living, so as to design smart homes that maintain safety, comfort and well-being of its inhabitants [6].

In the related literature, several works have focused on different aspects of audio signal parameterization and classification. However, there is still room for improvement given the novelty of the research field. This work especially focuses on the former, since it is found to be of paramount importance when it comes to recognize the occurring sound events [3].

In this paper, we propose linking two different concepts to come up with a signal parameterization that effectively describes surveillance-related sound events. The first one is Wavelet analysis, which is a technique commonly used in the field given the characteristics of the surveillance-related sound signals [4]-[6]. The second one is Gammatone filtering, which has been used to model the human auditory response [7]. In this work, we put forward a Wavelet analysis employing Gammatone mother functions, which are previously adapted to satisfy the Wavelet admissibility conditions. With regards to the machine learning technique asked to perform the audio classification, we choose Support Vector Machines, given its proved performance not only in general pattern recognition problems but also in audio classification tasks [4], [8].

Finally, it should be noticed that real-world applications require sound recognition systems robust to noise [4]. Therefore, in the experiments we paid especial attention to test the proposed system under different adverse noise (SNR) conditions.

The remainder of this paper is organized as follows. Section 2 reviews the related work. Section 3 presents the proposed Wavelet analysis employing auditory inspired Gammatone functions. Section 4 and Section 5 present the experimental setup and the obtained results. Finally, Section 6 draws up the conclusions and the future work.

2. RELATED WORK

Up to our knowledge, one of the seminal approaches to perform audio recognition for surveillance applications was conducted by Cowling and Sitte less than 10 years ago [2]. That work compiled several signal features and machine learning techniques previously employed in related research fields. The experiments, consisting in the classification of several surveillance-related sounds (such as *footsteps*, *wood snapping* or *glass breaking*), showed that both Mel Frequency Cepstral Coefficients and Continuous Wavelet Transform yielded the highest accuracy in combination with Dynamic Time Warping for classification.

Wavelet techniques have been commonly employed in the following works. In [6], Discrete Wavelet Transform was used to detect salient audio events in noisy environments for medical surveillance applications. The authors argued the good adaptation of the technique to signals with time-localized features such as *door slaps* or *footstep* sounds. The advantage of multi-resolution Wavelet techniques for sound classification in noisy-environments was also discussed in [5]. The signal representation introducing both time and frequency location improved the recognition of time varying sounds as *water* or *voices*. Unlike the aforementioned works, in [4] Discrete Wavelet Coefficients were merged with other temporal and frequency-based signal features into a single feature vector. This combined signal parameterization showed very good performance in classifying sounds for surveillance and security applications with Support Vector Machines.

There are two characteristics that define the Wavelet analysis: the irregularity and asymmetry of the Wavelet mother functions and the variable length of windows to better adapt to the frequency components being analysed. These characteristics make the Wavelet analysis suitable for representing surveillance-related sound events, which frequently present a short duration and impulsive characteristics (e.g., *gunshots*, *footsteps*). Previous works employed different Wavelet mother functions: *Morlet* [2], *Coiflet* [5], or more typically, *Daubechies* [2], [4]-[6].

In this work, we propose to use Gammatone mother functions instead, which are well known for their application to human auditory modelling (specifically to model the cochlear frequency response). They are asymmetric and have a variable duration that depends on their central frequency. Thus, filtering a signal with a Gammatone filter bank is similar to a Wavelet transform in the sense that all basis functions are scaled versions of the mother function at the first central frequency [9]. The connection between both techniques has already been mentioned in the literature [9], [10], being used for acoustic source segregation in [11].

3. AUDITORY-BASED WAVELET FEATURES

The Gammatone filter takes its name from the impulse response *g(t,B)* (see Fig. 1), which is the product of a Gamma distribution function and a sinusoidal tone centred at the f_c frequency, being computed as [7]:

$$g(t,B) = K\, t^{(n-1)} e^{-2\pi B t} \cos(2\pi f_c t + \varphi) \qquad t > 0 \qquad (1)$$

where K is the amplitude factor; n is the filter order; f_c is the central frequency in Hertz; φ is the phase shift; and B represents the duration of the impulse response.

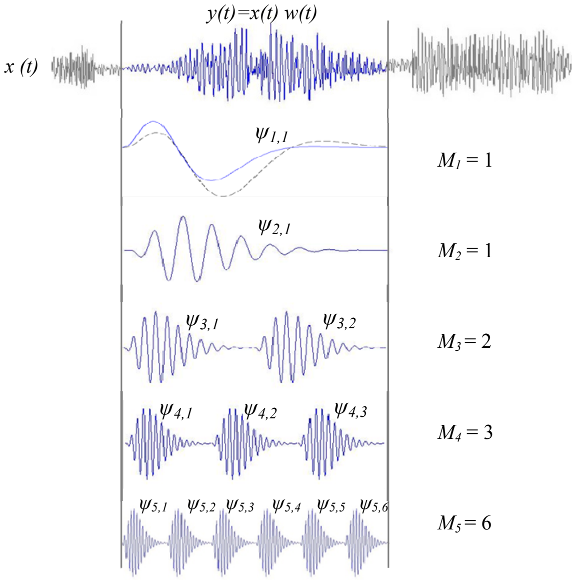

Figure 1. Shifted Gammatone mother functions $\psi_{i,m}(t,B)$ used to decompose the windowed signal y(t). M_i is the times $\psi_{i,}(t,B)$ is shifted to cover the whole signal frame. The dashed line represents $\psi_1(t,B)$ before substracting the D.C. component (see (4)).

A family of admissible Wavelets must satisfy the following two conditions so as to accomplish Parseval's equation and ensure the existence of its inverse [10]:

$$\int_{-\infty}^{+\infty} \left|\psi(t,B)\right|^2 dt \quad < \quad 0 \qquad (2)$$

$$\hat{\Psi}(w=0) = \int_{-\infty}^{+\infty} \psi(t,B)\, e^{-jwt}\; dt \quad = \quad 0 \qquad (3)$$

where *ψ(t,B)* states for the Wavelet mother function. Our Wavelet candidate function *g(t,B)* satisfies (2). Equation (3) implies that the function must have zero mean, which is satisfied by *g(t,B)* except for very low *B* (see Fig. 1). Thus, in order to satisfy (3), the mean value of the signal (D.C. component) is subtracted from *g(t,B)* by multiplying the function per an exponential function (similarly as in [12]), thus yielding an admissible family of Wavelet mother functions *ψ(t,B)*:

$$\begin{aligned} \psi(t,B) &= g(t,B)\; e^{j\xi t} \quad = \\ &= K\, t^{(n-1)} e^{-t(2\pi B - j\xi)} \cos(2\pi f_c t + \varphi) \end{aligned} \qquad (4)$$

where ξ a factor empirically set to cancel the D.C. component of *g(t,B)*.

The scaling of the proposed Gammatone Wavelet function is controlled by B, which is related to the Equivalent Rectangular Bandwidth (ERB), a psychoacoustic measure of the auditory filter width at each point along the cochlea [7]. Following Glasberg&Moore's model, the bandwidth B is calculated as [13]:

$$B = 1.019\, ERB = 1.019\left(24.7 \;+ \frac{f_c}{9.26}\right) \qquad (5)$$

According to (4), B fixes the length of $\psi(t,B)$ and the approximation detail: the narrower, the longer the window, and thus, the coarser approximation.

The i-th scaled versions of the Wavelet mother function $\psi_{i,}$ (t,B) are obtained by varying f_c from (5). The central frequency f_{ci} from each ψ_i (t,B) is computed as:

$$f_{ci} = (f_{high} + 228.72)e^{-\frac{i\,step}{9.26}} - 228.72 \qquad (6)$$

$$step = \frac{9.26}{N}\log\left(\frac{f_{high}+228.72}{f_{low}+228.72}\right) \qquad (7)$$

where f_{high} is the highest frequency considered; *step* is the logarithmic gap between consecutive filters; and N is the number of GT filters [13].

In order to incorporate the Wavelet translation, let us define the shifted Gammatone mother functions $\psi_{i,,m}$ (t,B) as:

$$\psi_{i,m}(t,B) = \psi_i(t - j\,L_i, B) \qquad 0 \le m < M_i \qquad (8)$$

where L_i is the length of ψ_i (t,B) and M_i is:

$$M_i = \left\lfloor \frac{L_{\max}}{L_i} \right\rfloor = \left\lfloor \frac{L_1}{L_i} \right\rfloor \qquad (9)$$

The scalar product of the windowed input signal (with a Hamming window) $y(t)=x(t)w(t)$ with the shifted Gammatone mother functions $\psi_{i,m}$ (t,B) yields the Wavelet time-frequency representation $\gamma_{i,m}$ (t,B):

$$\gamma_{i,m}(t,B) = < y(t), \psi_{i,m}(t,B) > \qquad (10)$$

Since large values of B result into short ψ_i (t,B), the Gammatone mother function needs to be shifted as much as necessary to cover the windowed signal $y(t)$. Thus, M_i grows with i and $\gamma_{i\,,m}(\tau,B)$ contains a larger number of components for high i, indirectly giving an extra weight to the high frequencies. Given that low frequencies are also important for the recognition of environmental sounds [14], we avoid this bias by computing the sum of the energy of the $\gamma_{i\,,m}(\tau,B)$ components for a certain i:

$$\hat{\gamma}_i(t,B) = \sum_{m=1}^{M_i} \left|\gamma_{i,m}(t,B)\right|^2 \qquad (11)$$

Finally, the Gammatone Wavelet (GTW) feature vector is obtained as:

$$\mathrm{GTW} = \{\ \hat{\gamma}_1(t,B), \hat{\gamma}_2(t,B), \ldots, \hat{\gamma}_N(t,B)\ \} \qquad (12)$$

5. EXPERIMENTAL EVALUATION

In order to empirically evaluate the performance of the proposed Wavelet analysis with Gammatone functions (hereafter GT-Wavelet features), a corpus of typical sounds from surveillance applications is used. Similarly as in [1]-[4], the specific sound classes are: *dog barks (90), screams (70)*, *voices (80)*, *gunshots (85)*, *footsteps (90)* and *thunders (75)*. The sound samples were taken from common sound libraries [16], [17] and present a variable duration (between 0.3 seconds and 4 seconds, depending on the file). All signals were normalized to 16 bits resolution and sampled at 22050 Hz.

The proposed GT-Wavelet feature vector is computed on frames of 45 ms length (fixed by the GT function with the largest time duration), with a 50% of overlap. As a baseline, we consider a typical Wavelet decomposition using the *Daubechies* mother function, with 4 vanishing moments and 6 decomposition levels, as in [4]. Hereafter, we will refer to this baseline parameterisation as Discrete Wavelet Coefficients (DWC). It should also be noted that two versions of the proposed GT-Wavelet features were computed: the first one using 7 GT functions (GTW-7), so as to make it compliant with the baseline (same amount of coefficients); and the second one employing 16 GT functions (GTW-16), thus covering the entire spectrum according to (7).

As a consequence of having sound samples with variable duration, the sound feature patterns result in variable dimension. However, a procedure is needed to fix the feature pattern dimensionality (since the classifiers generally need fixed length input data) while keeping the time evolution information of the sound signals. Similarly as in [4], each feature vector was divided into three portions of equal length. After computing one mean vector per portion, these three vectors are merged into the final feature vector used by the classifier

Support Vector Machines (SVM) perform the classification of the parameterised audio signals. The basis of the SVM is mapping the input samples into a high dimensional space and finding the hyperplane that optimally separates the two classes [15]. In this work, we employ a SVM with a Radial Basis Function kernel, given the good performance attained in [4] and [8]. To adapt the binary SVM to multi-class classification, we follow a *one vs. all* scheme, given the lower computational cost when compared to *one vs. one* approach [8].

Three experiments have been conducted. In the first one, the feature performance is evaluated in noiseless conditions, whereas in the second one, it is tested in noisy environments. The third experiment analyzes the performance of all the possible feature combinations. In the three experiments, a 10-fold cross validation scheme is employed to distribute the corpus data between training and test sets. The obtained accuracy is computed as the averaged percentage of correctly classified samples.

6. RESULTS

6.1. Classification in noiseless environment

The first experiment considers the different audio events in the ideal case without interfering noise. In these conditions, the accuracy achieved by the system is really high, with classification rates above 90% regardless of the audio feature employed (see Fig. 2). Specifically, the GTW-16 yields the highest averaged accuracy (96.2%), closely followed by the GTW-7 (95.3%) and in third position, the DWC (91.6%). Hence, both versions of the Gammatone Wavelet features outperform the DWC in noiseless conditions by 4.6% and 3.7% in average, respectively.

The confusion matrices obtained when using the baseline DWC and the proposed GTW-16 features are shown in Table 1 and Table 2, respectively. The most relevant improvements yielded by GTW-16 with respect to DWC are observed in the reduction of misclassifications between *footsteps* and *thunder* (10.8%), *dog barks* and *scream* (8%) and *gunshot* and *thunder* (6.2%).

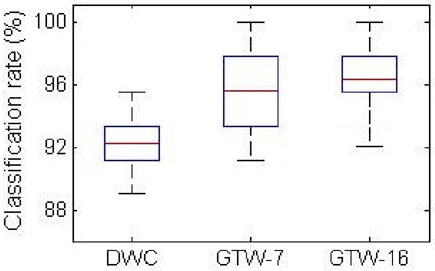

Figure 2. Boxplot of the classification rate (%) using DWC, GTW-7 and GTW-16 features.

	Dog	Scream	Voices	Gunshot	Footsteps	Thunder
Dog bark	98.9	14		1.2		
Scream	1.1	84				1.5
Voices			98.8	2.4		1.5
Gunshot		2	1.3	83.5		7.7
Footsteps				2.4	97.8	15.4
Thunder				4.7	2.2	66.2

Table 1. Confusion matrix using DWC features. Rows depict the system outputs and columns the targets.

	Dog	Scream	Voices	Gunshot	Footsteps	Thunder
Dog bark	95.6	6		1.2		1.5
Scream	2.2	92				
Voices	1.1		100			3.1
Gunshot	1.1	2		91.8		1.5
Footsteps					100	4.6
Thunder				1.2		81.5

Table 2. Confusion matrix using GTW-16 features. Rows depict the system outputs and columns the targets.

6.2. Classification in noisy environment

To recreate noisy conditions, the audio samples were contaminated with city background noise recordings at different SNR levels, ranging from 10dB to -20dB.

As shown in Fig. 3, the performance of the features is notably affected by the presence of noise, finding three differentiated regions. Firstly, for high and intermediate SNR levels (0 dB or higher), the best performing feature is GTW-16. The averaged classification rates are 3.2% and 5.3% higher than those yielded by DWC and GTW-7, respectively. Secondly, for low SNR levels (-5 dB to -10 dB), the differences between the three features are narrower, showing DWC and GTW-16 quite similar performances. Finally, for extremely low SNR levels (below -10 dB), the previously observed behaviour dramatically changes and GTW-7 yields the highest accuracies, i.e., 8.8% and 6.6% superior to DWC and GTW-16, respectively. Nevertheless, GTW-16 yields on average the best performance across the SNR sweep among the three tested features, with a 62.3%, followed by GTW-7 (61.5%) and DWC (60.5%).

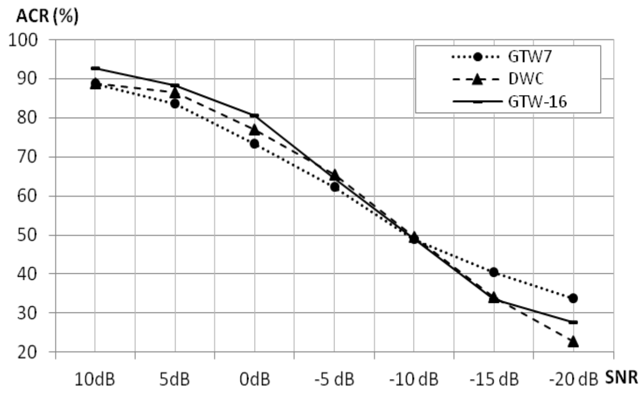

Figure 3. Averaged classification rate (%) using DWC, GTW-7 and GTW-16 features at different SNR levels.

6.3. Feature combination

The last experiment builds feature vectors from different combinations of the three signal features. Audio classification tasks were repeated both in noiseless and noisy conditions, following the same scheme of the previous experiments. Considering the averaged classification rate at all SNR conditions (see Table 3), the highest percentage is yielded by the feature vector merging all three features (68.65%), followed by the combination of DWC+GTW16 (68.36%). The results suggest that the more features the greater the information about the signal and hence, the higher performance obtained. Thus, the feature combination provides an advantage respect to using any single feature when conducting sound classification for surveillance applications. It is also noticeable that combining DWC with any of the GTW versions yields a better performance than combining both GTW versions, which suggests that there is a positive complementation between the baseline (DWC) and the proposed analysis technique (GTW).

If we rather take a look on the results at each SNR level (see Fig.4), for SNR higher than -10dB (leftmost part of Fig. 4), the feature vector combination of the 3 features yields the best performance. However, this feature vector is outperformed at extremely low SNR levels by the winner single feature vector at those categories i.e., GTW-7 (rightmost part of Fig. 4). These results suggest that the simplest version of the Gammatone Wavelet analysis is less sensitive to extremely noisy conditions, although it is worth noting the poor classification results obtained in this context. This effect could be due to the coarser representation provided by GTW-7 when compared to that from GTW-16, being less affected by the undesired background noise.

Feature	Averaged accuracy
GTW16 (best single feature)	66.52%
GTW16 + GTW17	67.27%
DWC + GTW7	67.47%
DWC + GTW16	68.36%
DWC + GTW16 + GTW7	68.64%

Table 3. Averaged accuracy yielded by the different feature vectors. The accuracy value is obtained by averaging the classification rates at every SNR level.

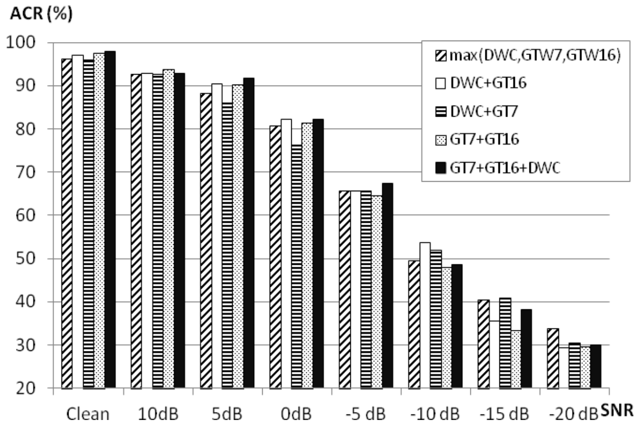

Fig 4. Averaged classification rates (%) obtained by feature combinations at different SNR levels.

7. CONCLUSIONS

This paper has proposed a parameterization technique to effectively representing the characteristics of audio signals for audio surveillance applications. The proposal is based on the fusion of Wavelet analysis with Gammatone filters. The results show that the proposed parameterization technique outperforms the typical Discrete Wavelet analysis with *Daubechies* mother functions, both at noiseless and noisy conditions, besides showing equivalent performance in the range of -5dB to -10dB SNR. The combination of both parameterisation techniques seems to be a better solution for the corpus at hand, yielding the highest classification rates at reasonable SNR levels. Future work lines will be addressed to test the proposed system in real environments, extending the range of sound events and considering non-stationary interfering background noise.

8. REFERENCES

[1] P.K. Atrey, C. Maddage, M.S. Kanjanhalli, "Audio based event detection for multimedia surveillance", *Proc. IEEE International Conference on Multimedia and Expo*, 2006.

[2] M. Cowling, R. Sitte, "Comparison of techniques for environmental sound recognition", *Pattern Recognition Letters*, vol. 24, pp. 2895-2907, 2003.

[3] S. Ntalampiras, I. Potamitis, N. Fakotakis, "On acoustic surveillance of hazardous situations", *Proc. ICASSP*, 2009.

[4] A. Rabaoui, M. Davy, S. Rossignol, N. Ellouze, "Using One-class SVMs and Wavelets for audio surveillance", *IEEE. Transactions Information Forensics and Security*, vol. 3, no. 4, December 2008.

[5] N. Mclachlan, D. K. Kumar, J. Becker, "Wavelet classification of indoor environmental sound sources", *International Journal of Wavelets, Multiresolution and Information Processing*, vol. 4, no. 1, pp. 81-96, 2006.

[6] D.Istrate, E.Castelli, M.Vacher, L.Besacier, J.F.Serignat, "Information extraction from sound for medical telemonitoring", *IEEE Trans. Information Technology in Biomedicine*, vol. 10, no. 2, April 2006.

[7] R. D. Patterson, J. Holdsworth, *A functional model of neural activity patterns and auditory images*, W. A. Ainsworth (Ed.), Advances in Speech, Hearing and Language Processing, vol. 3 part B, pp. 554-562, London: JAI Press, 1996.

[8] C-C. Ling, S-H. Chen, T-K. Truong, Y. Chang, "Audio classification and categorization based on Wavelets and Support Vector Machine", in *IEEE. Transactions on Speech and Audio Signal Processing,* vol. 13, no. 5, September 2005.

[9] A. Park, "Using Gammachirp filter for auditory analysis of speech", *18.327: Wavelets and Filter banks*, May 2003.

[10] R. Wöhrmann, L. Solbach, "Preprocessing for the automated transcription of polyphonic music: Linking Wavelet theory and auditory filtering", *Proc. ICMC*, 1995.

[11] A. Unoki, M. Akagi, "A method of signal extraction from noisy signal based on auditory scene analysis", *Speech Communication,* vol. 27, no. 3-4, pp. 261-279, 1999.

[12] T. Sing Lee, "Image Representation Using 2D Gabor Wavelets", *IEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 18, no. 10, October 1996.

[13] M. Slaney, "An Efficient Implementation of the Patterson-Holdsworth Auditory Filter Bank," *Apple Technical Report #35*, Apple Computer Library, Cupertino, CA 95014, 1993.

[14] B. Gygi, "Factors in the identification of environmental sounds", PhD. thesis, Indiana University, July 2001.

[15] N. Cristianini, J. Shawe-Taylor, *An introduction to Support Vector Machines and other kernel-based learning methods*, Cambridge University Press, Cambridge, 2000.

[16] "The BBC Sound Effects Library – Original Series," [Online]. Available: http://www.sound-ideas.com/bbc.html

[17] "The Freesound Project," [Online]. Available: http://www.freesound.org/.