PERCEPTUALLY MOTIVATED TEMPORAL MODELING OF FOOTSTEPS IN A CROSS-ENVIRONMENTAL DETECTION TASK

M. Umair Bin Altaf, Taras Butko, Biing-Hwang (Fred) Juang

Center for Signal and Image Processing, School of Electrical and Computer Engineering,
Georgia Institute of Technology, Atlanta, Georgia.

ABSTRACT

Real world sounds are ubiquitous and form an important part of the edifice of our cognitive abilities. Their perception combines signatures from spectral and temporal domains, among others, yet traditionally their analysis is focused on the frame based spectral properties. We consider the problem of sound analysis from perceptual perspective and investigate the temporal properties of a "footsteps" sound, which is a particularly challenging from the time-frequency analysis viewpoint. We identify the irregular repetition of the self similarity and the sense of duration as significant to its perceptual quality and extract features using the Teager-Kaiser energy operator. We build an acoustic event detection system for "footsteps" which shows promising results for detection in cross-environmental conditions when compared with conventional approach.

Index Terms— Temporal modeling, HMM, Teager-Kaiser operator, Acoustic event detection

1. INTRODUCTION

Conventional modeling of audio processing is dominated by the short-time Fourier-spectral perspective in which one views the significant audio features as arising out of a short-time linear Fourier power spectrum analyzer. In audio signal processing this framework endures, in part, because the human ear is considered as a frequency analyzer [1]. The ubiquity is evident in the design and implementation of almost all coding, detection and recognition systems which work with general audio as input. In fact, the perceptual information in any audio signal is encoded in the temporal variation of the various attributes of the signal. Short-time Fourier power analysis (STFA) limits the domain of attributes to static Fourier power spectral features within a short-time interval of 20-40 msec, while the temporal variation of spectrum is limited to short-term time derivatives of the STFA, computed via orthonormal polynomial interpolation involving centi-second level raw measurements [2]. This methodology works for most components of speech, which is a slowly varying signal (i.e., pseudo-stationary) with a known production model and a hierarchical meta-structure imposed by language, but is found to be inadequate, at times, in characterizing general acoustic events, e.g., gunshots, steps and many environmental sounds [3, 4], where such assumptions are unjustified. Even within speech, unvoiced plosives are not pseudo-stationary. Attempts to alleviate these weaknesses – while acknowledged [5, 6] – have been sporadic.

We discuss the general problem of signal representation in the context of a prototypical acoustic event (AE), the "footsteps" which proved to be particularly challenging for classification and detection in the CLEAR campaigns [4, 7]. These were conducted on sets of datasets recorded in meeting-rooms [4] and, in addition to "footsteps" , contain 13 AEs such as "applause", "cup clinks", "door knock", among others. A sample "footstep" AE is shown in fig. 1. It consists of repeated impulses and each impulse is followed by a gradual decay, collectively called the peak. Each peak is produced by an interaction of a foot or shoe with the ground surface. Moreover, it is the series of such roughly similar peaks, irregularly separated and interspersed with background sounds that gives rise to the perceptual quality of a "footstep" .

Conventional STFA would fail to account for these perceptual cues in at least two ways. The onset of the peak cannot be represented accurately due to the time-frequency uncertainty principle. STFA may be adequate for differentiating between audio signals with such impulsive onsets and signals where such characteristics are absent but for differentiating between similar impulsive signals such as steps from two different people or between "footsteps" and another audio class such as "applause", the accurate representation of the relative onset time does become an important criterion. Secondly, the short time interval is not long enough to account for the long-term (> 100 msec) repeatability. Increasing the frame-size to account for this will lead to unreliable estimates due to non-stationarity of the signal at such time-scales.

Another technique to model such long term behaviour within the short-time framework is leave it to the states in the hidden Markov model (HMM) to model the temporal sequence of the events. Nevertheless, a well-known limitation of the HMM is that the underlying Markov assumption constrains the state occupancy duration to be exponentially distributed independent of the data distribution [8]. This problem is also accentuated with general audio signals as, when compared with speech, such signals do not have a hierarchical language model which could provide a high level description of the audio signal.

In general, the idea of including temporal information into audio processing system is not new. In [9], the authors suggest using amplitude modulation features extracted in 1 sec analysis windows for robust speech detection. This approach is motivated by previous studies that indicate that this type of information is explicitly coded in the auditory cortex. The physiologically inspired analysis approach for audio classification presented in [10] is based on an advanced model of the auditory system. The authors propose modeling of the neural response over analysis window of the same 1 sec duration. Inspired by auditory scene analysis, a number of auditory features based on temporal analysis of the waveform were derived from amplitude histograms, amplitude onset maps, spectral and harmonic profiles of the waveform in 1 sec window. These have been shown to help in sound detection [11].

However, despite the recent success in neurophysiological and magneto-encephalographic studies, much of the structure, mechanism, and interactions of the stimuli in the auditory cortex remain unknown.

In this paper we apply a perceptually motivated approach which

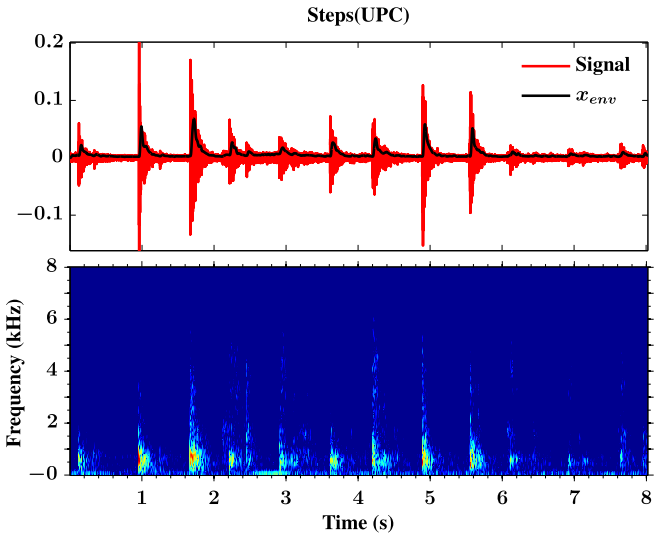

Fig. 1 – *The "footsteps" signal along with its smoothed Hilbert envelope and its Fourier spectra.*

consists of modeling analysis concepts that are important for humans to recognize sounds. We introduce two such concepts that arise from temporal analysis of AEs: the irregular repetition of the self similarity of the audio signal and the sense of duration [12]. This can provide the high level perceptual description of AEs derived from long-term temporal analysis of audio signal. We should note that, in general, audio signals may not exhibit the irregularly repeating self similarity but the dimension of perceptual duration is always present. We focus on organizing a perceptually motivated representation model for the "footsteps" AE for the purpose of detecting this sound against other AEs with the objective to generalize the approach to other AEs.

2. ACOUSTIC EVENT DETECTION PROBLEM

In AE detection (AED) task we aim at processing the acoustic signals collected by a set of distant microphones and convert them into symbolic descriptions corresponding to a listener's perception of the different sound events that are present in the signals and their sources.

For meeting-room environments, the task of AED has already been adopted as a semantically relevant technology in the European CHIL project (2004-2007) and two international evaluation campaigns. In the last evaluation CLEAR 2007 [4], five out of six submitted systems showed accuracies below 25%, and the best system had a 33.6% accuracy. In most submitted systems the standard combination of cepstral coefficients and HMM classifiers, widely used in speech recognition, was exploited. One of the major problems of such a low recognition rate is the presence of signal overlaps [13]. Another source of mistakes comes from mismatch conditions in training and testing. In fact, the database used in CLEAR 2007 consists of interactive seminars that were recorded in several rooms by different research groups (AIT, ITC, IBM, UKA, and UPC). The cross-environmental effect in these recordings manifests itself in the form of:

1. Different room impulse responses, different objects producing sounds and the way of their production.
2. Background noise.

Class Labels	*UPC vs UPC*	*FBK vs FBK*	*UPC vs FBK*	*FBK vs UPC*
applause	0.0	0.0	11.1	13.3
cup clink	0.0	0.0	69.4	4.7
chair moving	2.6	3.0	68.6	42.1
cough	1.5	0.0	13.9	23.1
door close	1.6	0.0	61.5	13.1
door open	0.0	0.0	63.9	96.7
key jingle	1.6	2.8	41.7	21.5
door knock	0.0	0.0	28.6	16.3
keyboard typing	1.7	0.0	60.0	15.2
laugh	4.7	0.0	19.4	48.4
phone ringing	0.8	0.0	19.7	25.0
paper wrapping	2.2	2.8	41.7	28.6
footsteps	1.2	0.0	76.3	80.8
Average	**1.37**	**0.66**	**44.3**	**33.0**

Table 1 – *Experimental results in terms of EER measure*

To demonstrate challenge of AED in mismatched conditions, we show the recognition results obtained using two databases of isolated AEs recorded at UPC [14] and FBK [15] meeting-rooms. Note that the recorded sounds of the events do not overlap. Both databases include 13 sound classes (excluding "unknown" class) and approximately 50-60 sounds per each class. People who participated in recordings took different places in the room during each recording session.

The features consist of 12 Mel-Frequency Cepstral Coefficients (MFCCs) including energy coefficient, extracted every 10 msec with a Hamming window of 25 msec The resulting parameters together with their first and second order time derivatives are arranged into a single observation vector of 39 components. Cepstral mean normalization is applied. Each sound is modeled by a 2-state full-connected HMM and each state is represented by a GMM of 64 mixtures with diagonal covariance matrix. The training was accomplished using the standard Baum-Welch training procedure. For evaluation the AEs were cut from the continuous audio according to the ground-truth labels. Then the isolated audio segments were fed to each HMM corresponding to a set of acoustic classes to perform Viterbi decoding. We evaluate the recognition rate of each sound class individually. Applying Bayes' rule and discarding the constant prior probabilities for class and out-of-class AE, the likelihood ratio in the log domain becomes:

$$\Lambda(x) = \log P(x|\lambda_C) - \log P(x|\lambda_{\bar{C}}) \quad (1)$$

The term $P(x|\lambda_C)$ is the likelihood of the utterance with observation vector x given that it is from the class model and $P(x|\lambda_{\bar{C}})$ is the likelihood of the utterance given it is from the corresponding out-of-class or non-class model, where $\lambda_{\bar{C}} = \arg\max_{\lambda \neq \lambda_C} P(x|\lambda_C)$. The likelihood ratio is compared to a threshold Θ and the class model is accepted if $\Lambda(x) > \Theta$ and rejected if $\Lambda(x) < \Theta$. The likelihood ratio essentially measures the degree to which the class model resembles the test utterance compared to some non-class model. In our experiments the decision threshold is set to obtain Equal Error Rate (EER) performance between rejecting the true class and accepting the non-class utterances. In table 1 we present the recognition results in terms of EER corresponding to four different experimental scenarios. The first two columns correspond to the case when training and testing is performed using different chunks of the same database, either UPC or FBK. The next two columns correspond to mismatched conditions where training is performed using sounds from one database and testing from another.

In matched conditions, the EER is relatively low (around 1%).

However, in mismatched conditions the EER increases drastically, indicating the fallout from the cross-environmental effect that is present in the rooms in form of noise and reverberation. Also, the intra-class variation of sounds plays a crucial role for sound recognition. In fact, "footsteps" is one of the classes that showed the highest error-rate, which is in line with the results from CLEAR 2007 [4].

3. SIGNAL SHAPE AND SELF SIMILARITY

In a broad sense we define shape as the evolution, the dynamics of signal parameters along time. If a signal does not change along with time, we could say that it has a constant shape. The basic question is: the evolution of *which* parameters need to be taken into account and over which time-scales to define signal shape. We address this from a perceptual point of view.

The perceptually identifying features of a "footstep" include a distinctive asymmetric peak shape due to the sudden rise, because of the contact of the foot with the ground, and a gradual fall, mainly because of the room impulse response. The peak duration is constrained to be in the vicinity of 100-200 ms and if we change this duration —for example by randomly changing the Fourier phase while keeping the Fourier power spectra constant for frame durations > 70 msec or < 10 msec —the changed sounds are no more recognized as "footsteps" .

Similarly, the minimum duration of the background sound between two successive peaks is perceptually important to the "footsteps" sound. For almost all sounds in the database, this interval of repetition is between 0.4-1 sec If we change the position of the peaks so that the repetition rate lies outside this range the perception changes —the sound is perceived as "door knock", if rate is less than 0.3 sec, and as disjointed strikes if the rate is more than 1 sec.

Estimation of the above temporal parameters requires energy estimates that give local measurements while preserving long term trends. In fig. 2, we show a self-similarity plot of MFCC, ΔMFCC and $\Delta\Delta$MFCC which is obtained by calculating the reciprocal of the Euclidean vector distance between frames. The grayscale intensity gives the similarity between frames centered at time location on the x-axis and y-axis. We notice that there is only a single prominent line at the main diagonal indicating self-similarity at zero lag τ, which does not provide any information. Thus MFCCs do not represent the long term self-similarity of the signal or the onset of the peaks.

With the failure of the frame based spectral approach, we turn to instantaneous energy measurements to characterize its shape. The Hilbert envelope provides one such measurement of the signal energy through the Hilbert transform [16].

We extract the Hilbert envelope, $|x_a(t)|$, of the signal $x(t)$ [16] and then low pass filter it at 20Hz to obtain the smoothed temporal envelope, $x_{env}(t)$, of the signal. This is shown in fig. 1 for the "footsteps" signal. The repeatability of the "footsteps" sound can be demonstrated with autocorrelation function, $r[.,.]$, of Hilbert envelope $x_{env}(t)$ shown in fig. 4, where the energy in each frame is normalized before calculating the autocorrelation as follows:

$$r_{M,L}^{x_{env}}[n_1, n_2] = \frac{\sum_{l=-L/2}^{L/2} x_{env}[Mn_1 - l]x_{env}[Mn_2 - l]}{\sum_{l=-L/2}^{L/2} x_{env}[Mn_1 - l] \sum_{l=-L/2}^{L/2} x_{env}[Mn_2 - l]} \quad (2)$$

Where $x[n]$ is the sampled $x(t)$, L is the frame duration and M is the frame shift. In fig. 4, the "footsteps" demonstrate a self-similarity at 0.6 sec seen clearly with a diagonal starting at this point. The contrast improvement in fig. 4 over that of fig. 2 is also obvious. We also observe that the diagonal lines are blurred at certain points.

The Teager-Kaiser energy operator (TKEO) [17] provides an unconventional perspective on the instantaneous energy of a signal. It relates energy to square of the signal amplitude *and* the square of its frequency. The discrete instantaneous energy, $x_{TKEO}[n]$ given by TKEO is:

$$x_{TKEO}[n] = x^2[n] - x[n+1]x[n-1] \quad (3)$$

Fig. 5 shows the autocorrelation of the smoothed TKEO energy of the signal using the definition in eq. (2), which is a marked improvement over the Hilbert envelope in fig. 4. Hilbert envelope brings out the apparent self-similarity, but its focus is only on the signal amplitude, which leads to a less sharp representation. TKEO provides the most crisp representation of this self-similarity which is robust against intra-class variations in "footsteps" because it includes instantaneous frequency, in addition to signal amplitude, in the energy estimate —combining the local and long-term measurement in a single estimate.

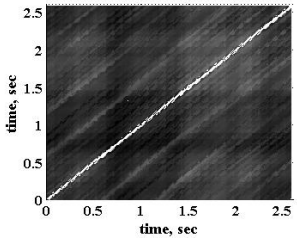

Fig. 2 – *Autocorrelation function of a "footsteps" sound using MFCC, ΔMFCC and $\Delta\Delta$MFCC*

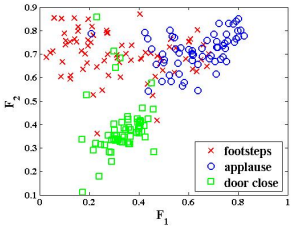

Fig. 3 – *Distribution of different AEs with similarity features F_1 and F_2 for UPC database.*

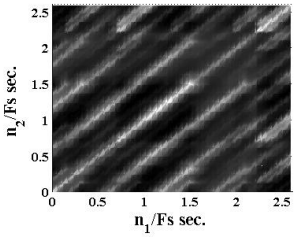

Fig. 4 – $r_{M,L}^{x_{env}}[n_1, n_2]$ *of a "footsteps" sound with $M/F_s = 10$ msec, $L/F_s = 800$ msec and F_s=16KHz.*

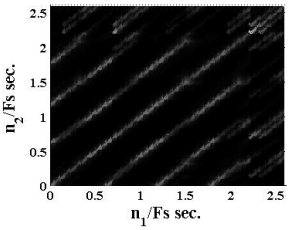

Fig. 5 – $r_{M,L}^{x_{TKEO}}[n_1, n_2]$ *of "footsteps" with $M/F_s = 10$ msec, $L/F_s = 800$ msec and F_s=16KHz.*

4. ACOUSTIC MODELING

4.1. Waveform self-similarity features

According to our perceptual observations, the discriminative characteristic of the "footsteps" AE in the time domain is the irregular repetition of certain self-similarities. One can notice that other AEs also exhibit the repetitive behavior: "applause" consists of the series of hands clapping; "door knock" consists of the successively hitting the door with human hand, etc. The basic concept behind repetition estimation is the similarity measurement.

We define similarity in terms of autocorrelation $r_{M,L}^{x_{TKEO}}$ of energy measurement from TKEO, where $M/F_s = 10$ msec, $L/F_s = 800$ msec and F_s=16KHz is the sampling rate. We compute two

features: $F_1[n]$ and $F_2[n]$ that represent the degree of the waveform self-similarity in short and long intervals, respectively. The feature $F_1[n]$ represents the maximum of the autocorrelation function for a frame centered at n within a time lag of 0.2-0.4 sec, i.e, $F_1[n] = \max_{\tau_1} r_{M,L}^{x_{TKEO}}[n, n + F_s\tau_1]$ where $\tau_1 \in [0.2, 0.4]$ sec and $F_2[n]$ is for the same frame within a time lag of 0.4-0.9 sec, i.e, $F_2[n] = \max_{\tau_2} r_{M,L}^{x_{TKEO}}[n, n + F_s\tau_2]$ where $\tau_2 \in [0.4, 0.9]$ sec.

In fig. 3 we show the distribution of AEs in the UPC database using features F_1 and F_2. As one can expect, the "footsteps" exhibit low short-term degree of self-similarity and high degree of long-term self-similarity. Events such as "applause" show high degree of self-similarity in short and long terms. On the other hand, features F_1 and F_2 show low degree of self-similarity for the AEs that are not repetitive like "door close".

4.2. Explicit duration HMM

For AE modeling we use HMM as a ready model for temporal characterization. HMMs incorporate the inherent temporal structure of audio and have shown to be particularly powerful in modeling sounds in which temporal structure is important, such as speech. Ergodic (full-connected) or left-to-right topologies can be chosen for general AEs. In either case, a well-known limitation of the HMM is that the underlying Markov assumption constrains the state occupancy duration to be exponentially distributed according to $P(d) = (1 - a_{ii})a_{ii}^{d-1}$, where d is the duration, and a_{ii} is the self-transition probability.

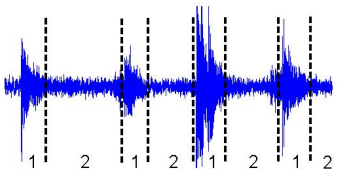

Fig. 6 – *State alignments of "footsteps" AE.*

In fig. 6 we show the waveforms of "footstep" AE together with its Viterbi state alignment that was obtained using ergodic 2-state "footstep" HMM applied to this AE (**1** and **2** are the two states of the HMM model). The first state corresponds to the impact sound constituting the "footstep" and another state corresponds to the background sound between two successive peaks. We note that the duration occupancy in each of these two states has certain temporal constraints. Up to 0.2 sec, the "footsteps" sound occupies the first state and then for the interval 0.4 - 0.9 sec, it remains in the second state. We model these constraints for "footsteps" AE using Ferguson's explicit duration HMM (EDHMM) [8]. No state in EDHMM in fig. 7 has self transitions, hence direct modeling of per-state duration distributions using state transition probabilities parameters (α_i, $\beta_i, \gamma_i, \delta_i$) becomes possible. In fig. 7 all states marked with the same number (**1** or **2**) have the same observation probability distributions but the transition probabilities between states monotonically change from left to right: α's and γ's decrease and β's and δ's increase.

4.3. Experimental Results

The detection results for the "footsteps" AE are presented in fig. 8. We used two approaches to incorporate the irregular repetition and the sense of duration analysis concepts: feature level fusion and

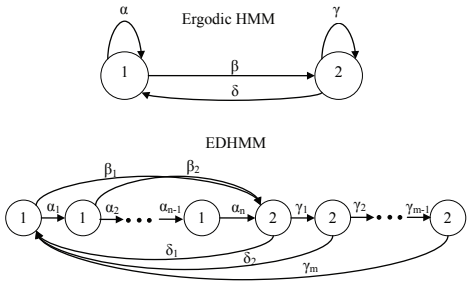

Fig. 7 – *Ergodic and explicit duration HMM*

EDHMM modeling. These approaches are compared with the baseline results for "footsteps" in table 1.

In feature level fusion, the features F_1 and F_2 are appended to the initial 39 MFCCs to form the composite 41-dimensional feature vector. In EDHMM the "footsteps" acoustic model is built as described in section 4.2. The estimation and inference of EDHMM parameters is performed using the forward-backward algorithm within the *allowable* duration intervals —chosen from previous perceptual observations to be 0.01-0.2 sec for peak and 0.4-0.9 sec for background sound between peaks. Note that both EDHMM and the baseline ergodic 2-state HMM have the same observation distributions in the corresponding states; the only difference between models lies in the transition probabilities between states. We achieved 14% of EER reduction in the case of feature level fusion approach and 27% of EER reduction in the case of EDHMM in cross-environmental scenario. Owing to our construction of EDHMM and the design of features, the error rate reduces significantly mainly because the temporal information, expressed in the form of long-term energy evolution, is less sensitive to the cross-environmental effects presented in different rooms.

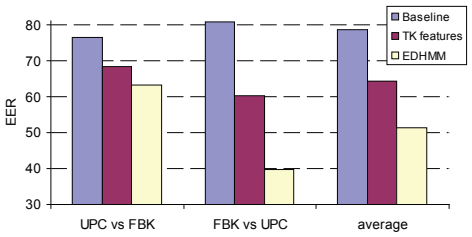

Fig. 8 – *Comparison results for "footsteps" AE.*

5. CONCLUSIONS

In this work we proposed two concepts that arise from temporal analysis of AEs: the repetition of the self similarity of the audio signal and the sense of duration. These concepts are incorporated at feature and signal model level for detection of "footstep" audio signal. These signals do not conform to the assumptions underlying the STFA-HMM paradigm which has its roots in speech recognition. The TKEO operator allowed us to represent the repetition accurately as it calculates energy as a functions of amplitude and frequency while we used EDHMM to model the duration of these repetitions. The results indicate a significant improvement over the state of the art for the "footsteps" signal in cross-environmental detection task. This signal representation and modeling framework is general enough to allow extension to other acoustic event classes, which we plan to pursue in a future work.

6. REFERENCES

[1] R. Plomp, "The ear as a frequency analyzer," *J. Acoust. Soc. Am.*, vol. 36, no. 9, p. 1628, 1964. [Online]. Available: http://asadl.org/jasa/resource/1/jasman/v36/i9/p1628_s1

[2] S. Furui, "Cepstral analysis technique for automatic speaker verification," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 29, no. 2, pp. 254 – 272, 1981.

[3] N. E. Huang, Z. Shen, S. R. Long, M. C. Wu, H. H. Shih, Q. Zheng, N. C. Yen, C. C. Tung, and H. H. Liu, "The empirical mode decomposition and the hilbert spectrum for nonlinear and non-stationary time series analysis," *Royal Society of London Proceedings Series A*, vol. 454, pp. 903–995, 1998.

[4] R. Stiefelhagen, J. G. Fiscus, and R. Bowers, Eds., *Multimodal Technologies for Perception of Humans*. Springer, 2008.

[5] S. Greenberg, "Auditory function," in *Encyclopedia of Acoustics*, M. J. Crocker, Ed. John Wiley & Sons, Inc, 1997, vol. 3, pp. 1301–1323.

[6] E. Lukasik and S. Grocholewski, "Comparison of some time-frequency analysis methods for classification of plosives," in *EUSIPCO*, 1998. [Online]. Available: http://ww.eurasip.org/Proceedings/

[7] R. Stiefelhagen and J. Garofolo, Eds., *Multimodal Technologies for Perception of Humans: First International Workshop on Classification of events, activities and relationships, CLEAR 2006*. Springer, 2007. [Online]. Available: http://www.springer.com/computer/image+processing/book/978-3-540-69567-7

[8] M. Johnson, "Capacity and complexity of HMM duration modeling techniques," *IEEE Signal Processing Letters*, vol. 12, no. 5, pp. 407 – 410, 2005.

[9] J. Bach, J. Anemüller, and B. Kollmeier, "Robust speech detection in real acoustic backgrounds with perceptually motivated features," *Speech Communication*, vol. 53, no. 5, pp. 690–706, 2011.

[10] S. Ravindran, K. Schlemmer, and D. Anderson, "A physiologically inspired method for audio classification," *EURASIP Journal on Advances in Signal Processing*, vol. 2005, no. 9, pp. 1374–1381, 2005.

[11] M. Büchler, S. Allegro, S. Launer, and N. Dillier, "Sound classification in hearing aids inspired by auditory scene analysis," *EURASIP Journal on Applied Signal Processing*, vol. 2005, pp. 2991–3002, 2005.

[12] K. Saifuddin, T. Matsushima, and Y. Ando, "Duration sensation when listening to pure tone and complex tone," *Journal of Temporal Design in Architecture and the Environment*, vol. 2, no. 1, pp. 42–47, 2002.

[13] A. Temko and C. Nadeu, "Acoustic event detection in meeting-room environments," *Pattern Recognition Letters*, vol. 30, no. 14, pp. 1281–1288, 2009. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167865509001603

[14] (2008) UPC-TALP database of isolated meeting-room acoustic events. [Online]. Available: http://catalog.elra.info/product_info.php?products_id=1053

[15] (2008) FBK-Irst database of isolated meeting-room acoustic events. [Online]. Available: http://catalog.elra.info/product_info.php?products_id=1093

[16] L. Cohen, *Time-Frequency Analysis*. Englewood Cliffs N.J: Prentice Hall PTR, 1995.

[17] J. F. Kaiser, "On a simple algorithm to calculate the 'energy' of a signal," in *Proc. International Conference on Acoustics, Speech, and Signal Processing ICASSP*, vol. 1, 1990, pp. 381–384.