RESEARCH INSTITUTE

AN ANALYSIS OF RHYTHMIC STACCATO-VOCALIZATION BASED ON FREQUENCY DEMODULATION FOR LAUGHTER DETECTION IN CONVERSATIONAL MEETINGS

Sucheta Ghosh   Milos Cernak

AN ANALYSIS OF RHYTHMIC STACCATO-VOCALIZATION BASED ON FREQUENCY DEMODULATION FOR LAUGHTER DETECTION IN CONVERSATIONAL MEETINGS

Sucheta Ghosh[1], Milos Cernak[1], Sarbani Palit[2], and B. B. Chaudhuri[2]

[1]Idiap Research Institute, Switzerland, [2]Indian Statistical Institute, Kolkata, India.
{sucheta.ghosh,milos.cernak}@idiap.ch, {sarbanip,bbc}@isical.ac.in

ABSTRACT

Human laugh is able to convey various kinds of meanings in human communications. There exists various kinds of human laugh signal, for example: vocalized laugh and non vocalized laugh. Following the theories of psychology, among all the vocalized laugh type, rhythmic staccato-vocalization significantly evokes the positive responses in the interactions. In this paper we attempt to exploit this observation to detect human laugh occurrences, i.e., the laughter, in multiparty conversations from the AMI meeting corpus. First, we separate the high energy frames from speech, leaving out the low energy frames through power spectral density estimation. We borrow the algorithm of rhythm detection from the area of music analysis to use that on the high energy frames. Finally, we detect rhythmic laugh frames, analyzing the candidate rhythmic frames using statistics. This novel approach for detection of 'positive' rhythmic human laughter performs better than the standard laughter classification baseline.

Index Terms — laugh signal detection, paralinguistic analysis

1 Introduction

Human laugh is a crucial social signal due to the range of inner meanings it carries. This social signaling event [1] may denote the topical changes, communication synchrony and positive affect; on the other hand, it may also show disagreement or satirist views. Therefore, automatic human laugh occurrence or laughter detection in speech may have many applications in spoken dialog and discourse analysis. In addition, the detection of this speech event may lead to increase in the word accuracies in the spontaneous automatic speech recognition.

Human laugh is developed as an inarticulate utterance to serve as an expressive-communicative social signal. The entire laugh period generally persists from 2 seconds to 8 seconds [2]. There exists many types of laugh. From the acoustical point of view, the sound of laugh can be voiced, as well as it can be unvoiced, resulting into the vocalized and nonvocalized laughter. The whole laugh episode is constituted with a mixture of vocalized and non-vocalized laugh. It was found that the voiced and rhythmic laughs were significantly and more likely to elicit positive responses than the variants such as unvoiced grunts and snort like sounds [3].

The laugh sound or laugh bout can be segmented into three parts, viz.(1) onset: explosive laugh, short and steep, (2) apex: vocalized part of laugh and (3) offset: post-vocalized fading part of laugh. The vocalized apex part is composed of laugh cycles (for example, the laugh sound "ha ha"), each cycle is composed of laugh pulses. The number of pulses depends on the power of the lungs, it can be 4 to 12 for one cycle. These laugh pulses have a rhythmic pattern.

Although it is found that in sustained laughter the apex might be interrupted by inhalations [2], human laughter is easily recognized through the detection of apex part [4]. Therefore, it is clear that the sound of laugh may also be based on the rhythmic breathing resulting in a staccato vocalization, i.e., the vocalization with each sound or note sharply detached or separated from the others. Detection of the apex part plays dominant role to recognize the human laughter.

The majority of the previous works on laugh detection (cf. Section 2) follow the supervised classification paradigm that may face a long extent of a training phase with considerable amount of costly annotated data. We hypothesize that a rhythmic nature of the vocalized laugh can allow us to use existing rhythm-detection signal processing based techniques, e.g., detection the rhythm in music [5], also for the laugh detection. This would lead to an unsupervised and less data-dependent laugh detection, as an alternative to conventional machine learning approaches.

In this work, we propose a three stage procedural method to detect human laugh using rhythm, through the laugh apex, which is the most prominent laugh part. This procedural method works in three basic procedural sequences: first we filter out low Power Spectral Density (PSD) frames using an automatic PSD threshold computation based on well-established Otsu's threshold technique [6]. Then we analyze all high-energy PSD frames to detect the rhythmic frames (such as rhythmic speech and/or rhythmic laugh) with a music rhythm detector algorithm [5] based on frequency demodulation. We select higher energy frames because human laugh is predominantly conceptualized as vowel-like high energy bursts [3]. Finally, we compute a statistical threshold to detect only the rhythmic laugh frames. We demonstrate the proposed detection method on naturally occurring conversations, that usually contain plenty of instances of happy and natural human laugh. Therefore, we choose multiparty meeting conversations AMI [7] as a database for evaluation. The recordings of the AMI meeting corpus show a huge variety of spontaneous expressions.

The organisation of this paper is structured as follows: in the next Section 2, we clarify the laugh as a signal and its types as established by the studies, and also describe in details the related works on laughter detection and recognition. In the following Section 3 we illustrate the proposed method. In the next consequent Section 4 we describe used data and experimental set-ups; this is followed by the discussion about the results in the subsection 4.4. Finally, we conclude the findings and possible future works in the Section 5.

2 Background

2.1 The laugh apex

The vocalized laugh can be spontaneous or voluntary. With clinical observation there is a clear distinction between spontaneous and voluntary laugh [8]. It is seen that during spontaneous laugh human self-awareness and self-attention is diminished. On the other hand, in voluntary laugh human produce a laugh sound pattern similar to the spontaneous laugh but still it differs in many aspects like vowel used (viz. the derivative of schwa), pitch, frequencies and amplitudes, voice quality etc. All these differences have effects on the rhythm of spontaneous and voluntary laugh.

Bachorowski et al (2001) found that the vocalized laugh is rhythmic compared to the snort-like laugh or grunts; and also the vocalized laugh elicits positive emotion than the other kinds of laugh [3]. Devillers et al (2007) found that the unvoiced laughs express more negative emotion, whereas the voiced laugh segments are perceived as positive ones [9].

2.2 Laughter Detection

A sizable number of previous works in laugh occurrence detection are already proved to be impressive in terms of techniques, results and large set of intricate features [10]. The majority of the previous works follow the supervised classification paradigm that may face a long extent of training phase with considerable amount of costly annotated data. Many of these works consider the task as a binary (i.e. laugh vs. non-laugh) classification [11, 12, 13] or segmentation problem [14, 15].

The label-type of laughter in the laughter detection tasks may vary from the coarse-grained label to the fine-grained one. The coarse-grained laugh detection generally implies to the binary classification (i.e. laugh vs. non-laugh). There exists some instances of coarse-level multi-class detection viz. the laughter classification along with the other non-laugh classes like silences, fillers etc. [13]. In other works, it detects many kinds of laugh such as polite, mirthful, derisive vs. non-laugh [16].

There also exists a few works on unsupervised classification of the laugh. Some unsupervised techniques depend on the burst detection and classification of the burst as laughter [17]. The affect bursts are defined as short, emotional and non-speech expressions that interrupt speech such as respiration, laughter or unintelligible vocal sound [18]. Therefore it is hard to tag the right meaning of (single or n-tuple) affective bursts without any reference. The non-parametric statistical methods also have been exploited in real-time, training-free framework to detect laughter; still one needs to extract features for this technique [19]. Majority of unsupervised methods are primarily tested on their own collected data.

In this work we attempt to propose a real-time, rhythm-based approach for laughter detection. We attempt to exploit the rhythmic pattern of laughter, following the work of Bachorowski et al (2001) [3], we aim to detect the vocalized laughter through detection of the laugh apex occurrence. We do not aim to detect the unvocalized laughter in this work.

2.3 Rhythm Analysis

Rhythm is defined as the systematic temporal and accentual patterning of sound. In music, rhythm perception is usually studied by using metrical tasks. Metrical structure also plays an organizational function in the phonology of language, via speech prosody or laughter [20]. We attempt to use this metrical structure of the human laugh without analyzing speech prosody. From the earlier studies [21, 5], we see that prosody and musical structure (such as rhythm) borrow or share concepts since long back. This studies with rhythmic patterns lead to the birth of the linguistic theories of stress-timed and syllable-timed languages. Here we do not consider the rhythm in the speech prosody.

Recent studies [22, 23] reveal that "rhythm" in speech should not be equated with isochrony. The absence of isochrony is not the same as the absence of rhythm. In [22], isochrony is defined as the organization of sound into portions perceived as being of equal or unequal duration. Strict isochrony expects the different elements to be of exactly equal duration, whereas weak one claims to have the tendency for the different elements to have the same duration. So, the languages can have rhythmic differences which have nothing to do with isochrony. But the rhythm in human laugh is always isochronous like any music, so we exploit the isochronous behavior of the human laugh in this work, and do not consider the non-isochronous rhythm of languages.

We use an approach of frequency modulation to retrieve this rhythm, following [5, 24]. To detect rhythmic laughter first we segment the whole speech to select the probable laughter segments, then we classify the candidate frames for voiced laughter using a rhythm algorithm based on frequency demodulation; finally we select the rhythmic laughter frames through a statistical process. We do not consider shared laughter captured on a single channel, rather our method is engineered for a solo laughter by the single participant.

3 Proposed Method

3.1 Rhythm based laughter detection

We use an unsupervised algorithm to detect laughter using its rhythmic property. This entire process can be divided into three basic sub-processes: first we filter out low power spectral density frames using an automatic PSD threshold computation. Then we use all high-energy PSD frames to detect all rhythmic segments (such as rhythmic speech and/or rhythmic laughter) with the rhythm detector algorithm. Finally, we compute a statistical threshold to detect only the rhythmic laughter frames.

Based on the detected rhythmic laughter frames, we are able to generate the time boundaries of the laugh segments. Description the three aforementioned sub-processes is following.

3.1.1 PSD threshold computation

We compute the PSD threshold using nonparametric power spectral density (PSD) estimation through Welch's overlapped segment averaging PSD estimator $S(e^{j\omega})/F$, where $F = fs$, i.e. the sampling frequency [25].

We compute the PSD threshold D_{th} following the Otsu method [6]. In this method the computed PSD set (PS) is sorted in ascending order, let us consider the index sets as $[1 \cdots L]$, then the sorted set is divided into two sets randomly, say: $\{1 \cdots k\}$ and $\{k + 1 \cdots L\}$, where $L = n(PS)$. Next, for $1 < k < L$, we iteratively compute $\sigma_B(L)$ then finally we compute D_{th}, given by,

$$D_{th} = \max \sigma_B(L) = \max \left[\max \left(\frac{[\mu(L) * \omega(k) - \mu k]^2}{\omega(k) * (1 - \omega(k))} \right) \right]$$

where, $\omega_k = \sum_{i=1}^{k} prob_i$, $\mu_k = \sum_{i=1}^{k} i * prob_i$, and $\mu_L = \sum_{i=1}^{L} i * prob_i$. Here $prob_i$ denotes the i-th probability considering the elements of the corresponding set in the iteration, further details is in the paper by Otsu (1975)[6].

We attempt to acquire the optimal value PSD threshold through a brute-force optimization process of running our laugh detection algorithm on the development data. In the section 4.4, we experimentally compare the performance of PSD threshold computation with the development data using the unsupervised method by Otsu (1975) [6] and the brute-force optimization method.

3.1.2 Rhythmic frame selection

First we select the high PSD frames using the threshold computed in subsection 3.1.1. Then these high PSD frames are passed through the rhythm calculation, thus we select the rhythmic frames among all the high energy frames. More specifically, we call these rhythmic frames as the candidate laughter frames.

We basically exploit frequency modulation (FM) technique to capture isochronous behavior of rhythm [5]. In this case we use an oscillator to modulate the frequency of a sinusoidal wave. Here the oscillator is the "carrier" and the other one is the "modulator". We attempt to use a sawtooth carrier in this case. Since laugh signal has a periodic nature it is traceable as a sawtooth (or triangular) waveform, therefore we choose the triangular hanning window as the basic oscillator function, which is computed as follows:

$$s = \left[\cos^2\left(\frac{2 * i * \pi}{l}\right) \right]_{i=1\cdots6}$$

here l denotes the hanning window length.

The properties of the "modulator" FM components are defined by the frequency band limit with a set of six harmonics that starts with zero then it reaches the periodicity pitch 200 Hz then all the other four (2×200, 4×200, 8×200 and 16×200) harmonics of that pitch. Here we choose to follow this filterbank implementation method described in Scheirer(1998) [26]. Each harmonics has two band-ranges. Therefore, this also initializes twelve band-range values. These frequency band-limits are used to compute the band-ranges. Since the beginning of the method we were computing data in the time domain. Now the signal is taken from the time domain to the frequency domain with Fourier transform, and we prepare the output using short time windows . Finally, we convolve the inverse fast Fourier transformed window data with a Fourier transformed half-hanning window.

We use a set of six band-limits at this moment: beginning at 200 Hz, increasing this in multiple of two, as the frequency results in a more and more complex multi phonic. The resulting wave is the summation of many different sinusoidal waves; the carrier frequency lies in the middle while the other tones lie above and below it at distances determined by the modulation frequency. When the modulation amplitude rises, the amplitudes of the additional frequencies also rise. However, this increase is difficult to formulate mathematically. The advantage of FM over additive one (the simple addition of sinusoidal waves) is that we need to use only two oscillators to convolute a rich and complex rhythmic human laugh sound. Although currently we use the six modulation frequencies, this number can be changed if needed. The output of this function is basically a six column matrix, each row of the matrix is one frame. We use the median of this output to use it further in the next subsection 3.1.3.

3.1.3 Rhythmic laughter frame(s) detection

We compute the basic statistic functions (namely mean and standard deviation) for all the obtained rhythmic candidate laughter frame with negative gradient (i.e. basically the negative difference between two consecutive points, and this is done to compute all local maxima points). Next, we derive the 95%-confidence bounds through the Student t-test of the standard deviation. Then, we compute a statistical threshold for rhythmic laughter frame selection as the difference between the upper bound of the confidence interval and the estimated population-standard deviation computed through same Student t-test. We compute this threshold on the basis of the hypothesis that the power of laugh is significantly higher than that of rhythmic speech/music. We select the frames as the laughter frames whose standard deviation is equal or higher than the threshold, and we finally compute the time intervals from those selected frames.

Algorithm 1 outlines overall view of the steps involved in the proposed laugh detection.

Algorithm 1 Overall View of Laugh Detection

Input: Wavfile W
read wavfile $(x_i, fs) = waveread(W)$
x_i =data sample; fs =sampling rate
Initialize: [frameSize, frameShift,T: noOfFrame, D_{th}=0]
/***PSD threshold** D_{th} **is calculated – procedure 1***/
D_{th} = threshold(W)   ▷ Otsu threshold computation
for$\{t = 1 \cdots T\}$
    pxx = welch_periodogram(x_t)   ▷ x_t: all data samples in t-th frame.
    /***rhythmic frame detection - procedure 2***/
    if $(pxx \geq D_{th})$
        c=median(FMrhythm(x_t))   ▷ FMrhythm() detects
rhythmic frame in six columns for six bands described in sec. 3.1.2. In c
the median of the six bands stored.
        for (j = 2:length(c))
            if $(c(j) - c(j-1) < 0)$ $flg = 1$;   ▷ local maxima check
        endFor
        σ=std(x_t);   ▷ std: the standard deviation
        if(flg) $\vec{\sigma} = [\sigma]$;   ▷ $\vec{\sigma}$ stores all candidate laughter frame stds σ
    endIf
endFor
/***laughter frame detection - procedure 3***/
if(length($\vec{\sigma}$) > 0)
    $(ci[lb, ub], sd)$ = ttest($\vec{\sigma}$)
        ▷ sd: the estimated population of standard deviation; ci: confidence
interval with lower lb and upper ub bounds; uses 1-sampled t-test
    for(i =1:len($\vec{\sigma}$))
        if($\vec{\sigma}(i) >= (ci[ub] - sd)$)
            Output: Print time interval of frame i.e. matching indices of t
    endFor
endIf

---

4 Experiment

4.1 Experimental Setup

In this framework, the method takes a raw speech signal as input and it outputs the time intervals of the laughter segments in the signal. We consistently apply a standard short-time analysis using a frame window of 2.5 sec (following the study of [2]) with 50% overlaps. We used part of AMI corpus as our test (5 meetings) and development (2 meetings) data.

4.2 Data

We used Augmented Multiparty Interaction (AMI) meeting corpus in this work. AMI meeting corpus [27] consists of 100 hours of meeting recordings. The recordings use a

Data	Baseline	Our Approach
AMI Data	81.1	84.5

Table 1. Percentage F1-measure comparison between baseline & proposed approach for test dataset split of our corpus

range of signals synchronized to a common timeline. These include close-talking and far-field microphones, individual and room-view video cameras, and output from a slide projector and an electronic whiteboard. During the meetings, the participants also have unsynchronized pens available to them that record what is written. The meetings were recorded in English using three different rooms with different acoustic properties, and include mostly non-native speakers. Following Petridis and Pantic (2011) [15] we used only close-talk headset audio (16kHz) recordings. We used the same data, which is used by Petridis and Pantic (2011) [15] (i.e. the seven meeting recordings recordings of eight participants consisting 6 young males and 2 young females of around 210 mins of recordings). We split the whole data set into two parts: the development data consists of two meeting recordings (i.e. IB4010 and IB4011 sets); we present the final result shown in the Table 1 using our test data of five meeting recordings (i.e. IB4001 and IB4005 sets). The challenge of AMI meeting corpus is that the data has a large amount of overlapping speech.

4.3 Baseline

We follow the same baseline protocol using the same feature set like [13, 28]. We establish a baseline for the (general laugh vs non-laugh) classification of human laugh using Interspeech 2013 Paralinguistic feature set. This feature set consists of 141 features [10]. We use support vector machine classifier with 5-fold cross-validation. We extract the features using the OpenSmile tool [29]. We use LibSVM [30] classifier for SVM training and prediction. This is a supervised binary sequential classification task. We use the data segments of 20 msec window at the rate of 10 msec. The baseline is achieved in a speaker dependent scenario. We select this baseline because it is the best performing supervised method for laugh detection.

4.4 Result & Discussion

Table 1 compares the results of our proposed approach with the supervised baseline approach. While the proposed rhythm based algorithm can be used for the detection of positive vocalized laughter using rhythm, the baseline has been designed to classify all kinds of laugh without distinction. We evaluate the results in the percentage F1-measures [31]. The performance of our approach on AMI meeting database shows better performance than the corresponding baseline. We notice that it performs in a balanced way in terms of precision and recall.

Figure 1 presents the ROC (Receiver Operating Curve) comparison of the PSD threshold computation over the development data. We compare the performance of threshold computation using the Otsu method [6] in against to that of the threshold computation using the brute-force optimization method. The ROC is computed using the true positive and false positive percentages. We see from the Figure 1 that the ROC computed with brute-force optimization thresholding is performing marginally better than the ROC computed by the method of Otsu (1975) [6] with the development data.
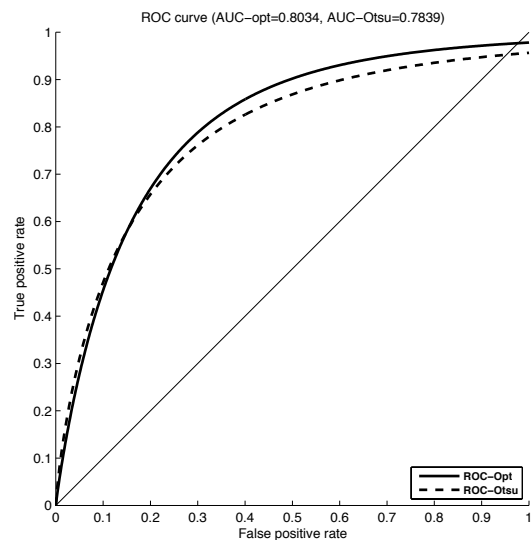

Fig. 1. PSD thresholding: ROC curve for Otsu thresholding and optimized thresholding with the development data

Therefore we use the optimized threshold with the development data to present the result in the Table 1.

5 Conclusions & Future Works

In this work we have outlined a novel algorithm for positive vocalized laugh detection using rhythm. This is a real-time, training-free approach in comparison to the existing supervised approaches of the laugh detection. The algorithm is based on the rhythmic transforms in laughter. The rhythm is analysed through frequency modulations using a modulator and a sawtooth carrier. All the six carriers are fixed, beginning at 200 Hz and the other four multiples of 200Hz. The strength of this technique also resides in that: we do not need to extract pitch or other intricate feature set to analyze rhythm; since it does not involve any complex process of computation, or intermediate file or memory handling, the time and space complexity of this method is low. We used AMI-role based meeting dataset to evaluate the proposed algorithm. The proposed laugh detection approach works well in comparison to the supervised baseline.

In this work we do not detect all kinds of vocalized laugh, we focus on the detection of rhythmic vocalized human laughter. This method is capable to work incrementally for further recognition of different laugh types or other recognition for laugh and rhythmic speech or music.This algorithm is sensitive to speech signal clipping; it may fail to detect the human laugh recorded in a noisy environment, specifically, it will fail to work with human speech data along with a background score of rhythmic music.

The Matlab code of the algorithm is available as open-source code at the following address: `https://github.com/sghoshidiap/LaughDet`.

6 Acknowledgements

7 References

[1] Alessandro Vinciarelli, Maja Pantic, and Hervé Bourlard, "Social signal processing: Survey of an emerging domain," *Image and Vision Computing*, vol. 27, no. 12, pp. 1743–1759, 2009.

[2] Willibald Ruch and Paul Ekman, "The expressive pattern of laughter," *Emotion, qualia, and consciousness*, pp. 426–433, 2001.

[3] Jo-Anne Bachorowski and Michael J. Owren, "Not all laughs are alike: Voiced but not unvoiced laughter readily elicits positive affect," *Psychological Science*, vol. 12, no. 3, pp. 252–257, 2001.

[4] P Drack, T Huber, and Willibald Ruch, "The apex of happy laughter: A facs-study with actors," *Current and Future Perspectives in Facial Expression Research: Topics and Methodical Questions*, 2009.

[5] William Arthur Sethares, *Rhythm and transforms PP. 1-10*, vol. 37, Springer Science and Business Media, 2007.

[6] Nobuyuki Otsu, "A threshold selection method from gray-level histograms," *Automatica 11, no. 285-296 (1975): 23-27.*, 1975.

[7] I. McCowan, J. Carletta, W. Kraaij, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, and V. Karaiskos, "The ami meeting corpus," in *International Conference on Methods and Techniques in Behavioral Research*, 2005, pp. 137–140.

[8] K.E. Bright, T.J. Hixon, and J.D. Hoit, "Respiration as a laughing matter," *WHIMSY IV, D.L.F. Nilsen, ed.*, 1986.

[9] L. Devillers and L. Vidrascu, "Positive and negative emotional states behind the laugh in spontaneous spoken dialogs," in *ICPHS 2007 workshop Laughter, Saarbrücken, Germany, 2007.*, 2007, pp. 37–40.

[10] Bjöjrn Schuller, Stefan Steidl, Anton Batliner, Alessandro Vinciarelli, Klaus Scherer, Fabien Ringeval, and Mohamed Chetouani, "The INTERSPEECH 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism," in *INTERSPEECH 2013 – 16th Annual Conference of the International Speech Communication Association, Proceedings*, 2013.

[11] L. Kennedy and D. Ellis, "Laughter detection in meetings," in *NIST ICASSP 2004 Meeting Recognition Workshop Montreal, QC, Canada*, 2004, number 118-121.

[12] Khiet P. Truong and David A. Van Leeuwen, "Automatic discrimination between laughter and speech," *Speech Communication*, vol. 49, no. 2, pp. 144–158, 2007.

[13] Hugues Salamin, Anna Polychroniou, and Alessandro Vinciarelli, "Automatic detection of laughter and fillers in spontaneous mobile phone conversations," in *Systems, Man, and Cybernetics (SMC), 2013 IEEE International Conference on. IEEE*, 2013, pp. 4282–4287.

[14] M. Knox and N. Mirghafori, "Automatic laughter detection using neural networks," in *Proc. of Interspeech. pp. 2973-2976. ISCA*, 2007.

[15] S. Petridis and M. Pantic, "Audiovisual discrimination between speech and laughter: why and when visual information might help," *Multimedia, IEEE Transactions on*, pp. 5117–5120, 2011.

[16] Hiroki Tanaka and Nick Campbell, "Classification of social laughter in natural conversational speech," *Computer Speech and Language*, vol. 28, no. 1, pp. 314–325, 2014.

[17] Oren Peleg and Moshe Wasserblat, "Method and system for laughter detection," U.S. Patent 8,571,853, October 2013.

[18] Marc Schröder, "Experimental study of affect bursts," *Speech communication*, vol. 40, no. 1-2, pp. 99–116, 2003.

[19] Chih-Hung Chou and et al, "A real-time training-free laughter detection system based on novel syllable segmentation and correlation methods," in *Awareness Science and Technology (iCAST) 4th International Conference on. IEEE*, 2012, pp. 294 – 297.

[20] Manfred Clynes and Janice Walker, "Neurobiologic functions of rhythm, time, and pulse in music," *Music, mind, and brain*, , no. 2, pp. 171–216, 1982.

[21] Aniruddh D. Patel, John R. Iversen, and Jason C. Rosenberg, "Comparing the rhythm and melody of speech and music: The case of british english and french," *The Journal of the Acoustical Society of America*, vol. 119, pp. 3034–47, 2006.

[22] Caroline Bouzon and Daniel Hirst, "Isochrony and prosodic structure in British English," in *In Speech Prosody 2004 International Conference*, 2004, pp. 223–227.

[23] Martina Huss, John P. Verney, Tim Fosker, Natasha Mead, and Usha Goswami, "Music, rhythm, rise time perception and developmental dyslexia: perception of musical meter predicts reading and phonology," *Cortex*, vol. 47, no. 6, pp. 674–89, 2011.

[24] Radoslaw Niewiadomski, Maurizio Mancini, Yu Ding, Catherine Pelachaud, and Gualtiero Volpe, "Rhythmic Body Movements of Laughter," in *In Proceedings of the 16th International Conference on Multimodal Interaction*, 2014, pp. 299–306.

[25] John G. Proakis and Dimitris G. Manolakis, *Digital Signal Processing PP. 910–913*, Prentice Hall, 1996.

[26] Eric D. Scheirer, "Tempo and beat analysis of acoustic musical signals," *The Journal of the Acoustical Society of America*, vol. 103, no. 1, pp. 588–601, 1998.

[27] S. Renals, T. Hain, and H. Bourlard, "Recognition and understanding of meetings the AMI and AMIDA projects," in *In Automatic Speech Recognition and Understanding ASRU IEEE Workshop*, 2007, pp. 238–247.

[28] L. Kaushik, A. Sangwan, and J. H.L. Hansen, "Laughter and filler detection in naturalistic audio," in *Proceedings of Interspeech Germany*, 2015.

[29] Florian Eyben, Martin Wöllmer, and Bjöjrn Schuller, "Opensmile: the Munich versatile and fast open-source audio feature extractor," in *Proceedings of the international conference on Multimedia. ACM*, 2010, pp. 1459–1462.

[30] Chih-Chung Chang and Chih-Jen Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, pp. 27–54, 2011.

[31] David M W Powers, "Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation," *Journal of Machine Learning Technologies*, vol. 2, no. 1, pp. 37–63, 2011.