Novel Approach for Detecting Applause in Continuous Meeting Speech

C. Manoj, S. Magesh, Aditya Sriram Sankaran, Dr. M. Sabarimalai Manikandan*
Department of Electronics and Communication Engineering
Amrita School of Engineering, Amrita Vishwa Vidyapeetham
Coimbatore, India
E-mail: *msm.sabari@gmail.com

Abstract— **This paper proposes a robust and automated applause detection algorithm for meeting speech. The features used in the proposed algorithm are the short-time autocorrelation features such as autocorrelation energy decay factor, amplitude and lag values of first local minimum and zero-crossing points extracted from the autocorrelation sequence of a windowed audio signal. We apply decision thresholds for the above acoustic features to identify applause and non-applause segments from the audio stream. The performance of the proposed algorithm is compared with the conventional method using mel frequency cepstral coefficients (MFCC) feature vectors and Gaussian mixture model (GMM) as classifier. We have also analyzed the performance of these algorithms by varying the number of mixtures in GMM (2, 4, 8, 16 and 32) and various thresholds in the proposed method. The methods are tested with a multimedia database of 4 hours 37 minutes of meeting speech and the results are compared. The precision rate, recall rate and F1 score of the proposed method are 94.40%, 90.75% and 92.54% respectively while those of conventional method are 67.47%, 96.13% and 79.29% respectively.**

Keywords- Audio classification, video summarization, sports highlight extraction, semantic video analysis, audio content analysis

I. Introduction

Recently, there has been a great incidence in the processing and study of meeting speech because of the increase in multimedia processing capabilities [1]. In most meetings or speeches, the speakers elicit reactions from other members or the audience by way of applause (sound generated by the clapping of hands). This applause can be regarded as a method by which the speaker and other members of the meeting or the audience communicate. Also, the applause is, in fact, an indication of important parts of the meeting speech as the audiences usually welcome or appreciate the speaker by means of applause. So, applause can be used as a key indicator of highlights in meeting speech. Applause detection also plays an important role in highlights extraction from sports video [1] [2]. Many applause detection methods have been reported in the literature. Xiong et al. [2] used MPEG-7 features and entropic prior HMMs to detect applause for the purpose of highlights extraction from baseball, golf and soccer games. Cai et al. [3] used the 32-dimensional feature vector derived from the perceptual features and mel frequency cepstral coefficients (MFCCs), the hidden Markov models (HMM) to model applause, cheer and laughter in audio stream and a log-likelihood scores based method is used to make final decision. MFCCs are sub-band energy features in mel-scale. Olajec et al. [4] employed MFCCs and Gaussian mixture model (GMM) for applause detection, and then the discriminative properties of various parts of the MFCC feature space for applause detection is studied in [5]. Otsuka et al. [6] used modified discrete cosine transform (MDCT) coefficients and GMMs for applause detection. Cai et al. [7] introduced subband spectral flux and harmonicity prominence features and used HMMs to detect ten key audio elements including applause, cheer, and laughter. In an applause signal, rate of change of the temporal and spectral parameters is determined by the clap density. In such cases, most of the low-level descriptors are incapable of detecting the applause segment. Moreover, in practice, target audio sounds are mixed with various kinds of sound, and thus many non-target sounds can be misclassified as target audio keywords due to inefficient representation of feature vectors of all the non-target sounds that are widely scattered in the feature space [7].

To avoid this misclassification, Xiong Li et al. [8] used some fundamental characteristics like duration, pitch and occurrence locations for finding applause event boundaries in meeting speech. The temporal parameters such as short time energy (STE), zero crossing rate (ZCR) and low energy ratio (LER) and the spectral parameters such as spectral centroid, spectral roll off and spectral flux are inconsistent since the applause signal has the following characteristics: 1) it is very complex non-stationary oscillatory signal and thus may exhibit diverse structures; 2) The signal amplitude and duration vary too much and are inconsistent in practice; and 3) it has varying concentration degree of the harmonic energy. Furthermore, supervised approaches have the following disadvantages: 1) manual segmentation for training prior to the analysis of all the generic and environmental sounds; 2) detection accuracy relies highly on the quality and quantity of the training data and the selection of model parameters for the modeling; 3) establishment of background models to cover all non-target sounds. Moreover, the performance is critical when the training and open test set are significantly different. This paper is organized as follows. The proposed algorithm for applause detection is explained in section II. The experiment setup and the parameters are elaborated in section III. The results of the

978-1-4244-8679-3/11/$26.00 ©2011 IEEE

two methods are also analyzed in section III. Finally, we conclude in section IV.

II. Proposed Method

In this paper, we propose a new decision tree based approach using autocorrelation features for detecting applause sounds in meeting speech. The proposed algorithm has three steps: preprocessing, feature extraction and decision making.

A. *Preprocessing*

In realistic environments, the meeting speech signal is corrupted by different kinds of noise sources such as microphone artifacts and power line interference. Therefore, we designed a linear-phase finite impulse response (FIR) high pass filter with a desired amplitude response given in Table I.

The FIR filter coefficients are computed using the Least Square linear-phase FIR filter design method. The original speech signal is fed to the filter for suppressing microphone artifacts and power line interference. The magnitude and phase response of the designed filter is shown in Fig.1. Before passing through the filter, the signal is normalized. Normalization scales the signal so that the signal is within the range [-1 1]. It was observed in [8] that the duration of applause is greater than 3 seconds in 90% of the segments. So, the audio segment is split into segments of duration 2 seconds with no overlap between segments and each segment is processed separately and decision will be made on each 2 second segments. To reduce the computational overhead, each segment is checked whether it is silence or not. Only the segments which are detected as silent are discarded and only the other frames are processed further. Silence detection is based on the energy of the segments. Energy is given by

$$E = \sum_{i=1}^{N} x^2[i] \tag{1}$$

If energy of each frame is less than 0.3, then the frame is detected as silence.

B. *Feature extraction*

The selection of audio features plays an important role in any detection algorithm. To get better accuracy, the features should be able to discriminate between the desired classes of sounds. The features should also be robust to noise and other background sounds so that the system can be used in real-time. Considering the above criteria, the following features based on autocorrelation features have been proposed. The segments detected as not silence is split into frames of 16 ms (256 samples for 16 kHz sampling frequency) and 50% overlap. Hamming window is applied to each frame. The energy of each frame is calculated using (1) and if the energy of the frame is less than 0.01, the frame is discarded. The following features are computed for the frames that have energy greater than 0.01.

TABLE I: Filter Specifications

Frequency (Hz)	Amplitude (Gain)
0	0
300	0
310	1
Fs/2	1

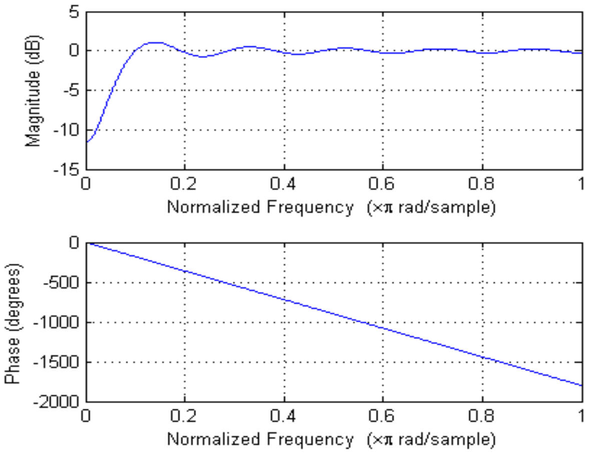

Figure 1. Magnitude and phase response of high pass filter

C. *Decay factor*

The short time autocorrelation function (ACF) of the windowed signal of length N is given by

$$r_i[k] = \frac{1}{N}\sum_{m=0}^{N-1} x_i[m]x_i[m+k] \tag{2}$$

$$r_i[k] = \frac{1}{N}\sum_{m=0}^{N-k-1} (x[i+m]w[m])(x[i+k+m]w[k+m]) \tag{3}$$

The periodicity of the autocorrelation function indicates that the signal is periodic and if the signal is not periodic, the autocorrelation function falls to zero. In the case of speech signals, the autocorrelation function of voiced speech is periodic while the autocorrelation function of noise and unvoiced speech falls to zero rapidly. So, the ratio of the energy in the first L values of ACF to the total energy in ACF can be used to differentiate applause segments from other segments of speech. We have observed that the shape of energy of ACF resembles an exponential signal. Maximum part of the energy of the ACF is concentrated near the zero lag. The autocorrelation energy decay factor estimating the concentration of energy within a specified autocorrelation lag L is given by

$$\eta_i = \frac{\sum_{k=0}^{L} r_i^2[k]}{\sum_{k=0}^{N-1} r_i^2[k]} \tag{4}$$

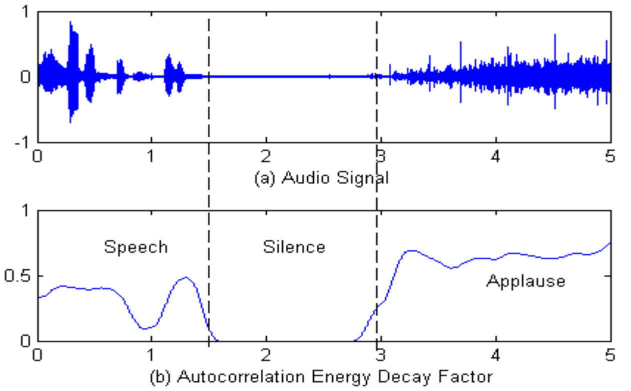

Figure 2. Speech signal and its autocorrelation decay factor.

A sample signal containing speech, silence and applause and its autocorrelation energy decay factor is plotted in Fig. 2. It can be seen that the applause segment has clearly distinct higher decay factor compared to other sounds.

D. *First Local Minimum of autocorrelation function*

The first local minimum of the autocorrelation function has been used to find the pitch frequency of the speech signal. We use this feature to discriminate between speech signal and applause signal. For speech signal, the first local minima occurs around 16 to 20 lag. It was observed that for applause, the first occurs for lag between 4 and 7. For noise, it usually occurs below 4. We use this property of ACF to discriminate between applause and other segments. Noise segments are also removed by the property that their first zero crossing of the ACF occurs below 3 and the first minimum value is greater than -0.3.

A segment of audio signal containing speech, silence and applause segments is shown in Fig. 3(a). The lag corresponding to the first local minima of the signal is shown in Fig. 3(b). It can be inferred from the plot that applause segments can be distinguished from other segments.

Speec

E. *Band Energy Ratio*

The Fourier Transform of the signal gives the frequency content present in the signal. Since different sounds have different distribution of energy over the frequencies, Fourier transform can be used to analyze the nature of the signal. The peak of the spectrum of the signal gives the dominant frequency of the signal. It was observed that for applause signals, this peak was between 1000 and 5000 Hz. To implement this, we take 64 bin fast Fourier transform (FFT) of the signal and take its peak amplitude in the range of 1000 to 5000 Hz i.e. bins 5 to 20 in 64 bins of FFT energy values. A sample speech signal and its BER are plotted in Fig. 4. It can be seen that applause segment has a consistent BER compared to speech and silence segments.

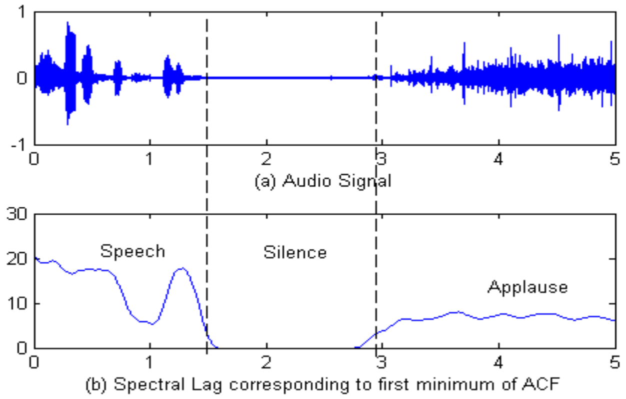

Figure 3. Speech signal and its lag corresponding to first minimum

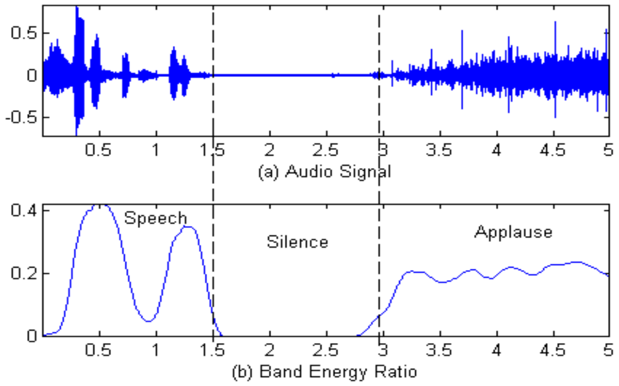

Figure 4. Speech signal and its band energy ratio

F. *Decision Making*

To detect whether a segment is applause or not, thresholds are applied to the above features. The thresholds for the various features are given in Table II.

TABLE II: DECISION THRESHOLDS

Audio Features	**Threshold range**
Decay Factor	0.6 – 0.8
Lag of first minima of ACF	4 – 7
Index of FFT bins for BER (using 64 bin FFT)	5 – 20
Band Energy Ratio	0.08 – 0.31

III. EXPERIMENTS AND EVALUATION

In this section, we compare the performances of supervised MFCC method and the proposed method for applause detection for meeting speech.

A. *Peformance Metrics*

The performance metrics such as precision rate (PR), recall rate (RR) and F1-Score are evaluated to compare the different methods for applause detection. These metrics are explained below:

Precision rate (PR) is defined as the ratio of number of correctly detected segments as applause to the total number of segments detected as applause. The PR is defined as

$$PR = \frac{N_c}{N_c + N_f} \times 100\% \qquad (6)$$

Recall rate (RR) is defined as the ratio of number of correctly detected segments as applause to the total number of applause segments. The RR is defined as

$$RR = \frac{N_c}{N_c + N_n} \times 100\% \qquad (5)$$

$$F_1 = \frac{2 \times PR \times RR}{PR + RR} \qquad (6)$$

where, N_c is the number of correctly detected frames (true positive), N_f is the number of false positive detections, N_n is the number of false negative detections.

B. *MFCC-Based Approach*

The conventional phase contains two phases, training and testing. In training phase, using some known datasets, the parameters of GMM are estimated. The accuracy of the system is determined in the testing step. The two steps are explained in detail below. Training data is created manually by separating audio segments into two groups 'Applause' and 'Not Applause'. MFCCs are then calculated for these two classes with number of filters as 26. Number of MFCC coefficients we get is 13, from which we eliminate the zeroth coefficient. We also include the first and derivatives of the features to get 36 dimension feature vectors. GMMs of various orders – 2, 4, 8, 16, 32 - are created for these two classes.

TABLE III: PARAMETERS USED FOR EXTRACTION OF MFCCS

Window size	20 ms
Window step size	10 ms
No. of MFCC coefficients	13
No. of filters	26

The input audio is split into frames of 2 seconds each without any overlap. The 2 second frame is again split into frames of 20ms each with 10ms overlap. The 36 dimension MFCC feature vector for each frame is calculated. Average of these vectors is calculated to get a single 36 dimension vector for the 2 seconds frame. Posteriori probabilities of the vector for both classes 'Applause' and 'Not Applause' are calculated and the class for which the probability is high is chosen.

The results obtained for MFCC algorithm is shown in Table IV. It has been observed that there is no theoretical way to estimate the number of mixtures a priori. So the experiment has been done for different number of mixtures in GMM (2, 4, 8, 16, and 32). The computational time increases when the GMM order increases. So it is of practical importance to employ an efficient Model. The number of False Positive detections is very high which implies that this algorithm detects many non applause segments as applause. This degradation is caused because of the environmental noise such as crowd cheer, music, laughter present during the meeting speech. The plot of precision rate, recall rate and F1-score is shown in the figure below.

It can be observed that the recall rate for all GMM orders is high reaching a maximum of 96.41% for GMM order 32. Even though high recall rate is achieved, the overall performance of the system reduces because of very low precision rate. A bar chart showing the F1-Score for different GMM orders is shown in Fig. 6. Maximum F1 Score obtained is 79.29 % for the Gaussian Mixture Model of order 4. It can be observed that the F1 Score of the system does not increase when the GMM order is increased. This is because in the case of large number of mixtures, models for noise will also be created.

TABLE IV: PERFORMANCE EVALUATION OF CONVENTIONAL METHOD

Number of Mixtures in GMM	Precision Rate (%)	Recall Rate (%)	F_1 Score (%)
2	62.35	93.97	74.96
4	67.47	96.13	79.29
8	59.18	95.77	73.15
16	58.30	96.27	72.62
32	60.49	96.41	74.34

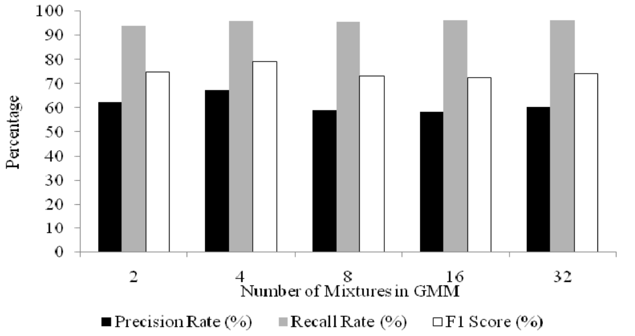

Figure 5. Performance of conventional approach

C. Proposed Approach

The results obtained for the proposed method is shown in the Table V. In this method we use the parameters LFilter, length of filter used for feature smoothing and TDecision, percentage of frames that should satisfy the conditions to be declared as applause segment. These two parameters are varied and the performance measures are obtained. TDecision=0.5 indicates that if more than 50% of frames in a segment satisfy the conditions then that segment is declared as applause segment. It can be observed that the number of false positive detections in all the cases is very less when compared to that of MFCC algorithm. So unlike the MFCC algorithm, the precision rate and recall rate are equally good increasing the overall performance of the system. The maximum F1-Score of 92.54 is obtained for the values LFilter=15 and TDecision = 0.5.

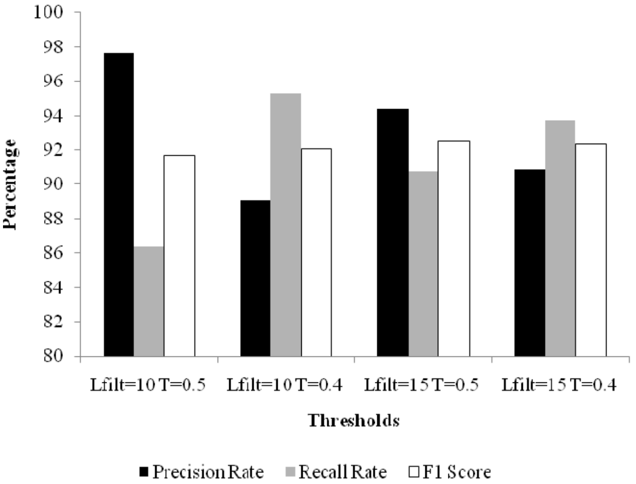

Figure 6. Performance of Proposed Method for Various Thresholds

TABLE V: PERFORMANCE OF PROPOSED METHOD FOR VARIOUS THRESHOLDS

L_{filt}	T	Precision Rate (%)	Recall Rate (%)	F_1 Score (%)
10	0.5	97.65	86.37	91.66
10	0.4	89.07	95.27	92.06
15	0.5	94.4	90.75	92.54
15	0.4	90.88	93.75	92.3

IV. CONCLUSION

In this paper, we proposed a new decision tree based algorithm for applause detection in a continuous meeting speech. The proposed approach is based on the short-time autocorrelation function features – decay factor, first local minimum and band energy ratio. We have evaluated this approach experimentally for 4 hours and 37 minutes of meeting speech and have achieved promising results. The proposed approach gives us an F1 Score of 92.54% (best-case scenario). We have evaluated this proposed framework against an already existing scheme for applause detection in meeting speech. The conventional method using MFCC features and GMM classifier gives us a best-case scenario F1 Score of 79.29% (for GMM with 4 mixtures). Another major improvement over the conventional method is an improvement in the robustness of the system in noisy environments. In the future, we can extend the proposed framework for applause detection to a general system as opposed to its current application for only meeting speech. Also, using the same features, a similar framework can be developed for other applications like laughter detection, whistle-sound detection, etc.

V. REFERENCES

[1] Yan-Xiong Li , Qian-Hua He , Sam Kwong , Tao Li , Ji-Chen Yang, “Characteristics-based effective applause detection for meeting speech,” Signal Processing, Vol. 89 No. 8, pp. 1625-1633, August, 2009.

[2] R. Cai, L. Lu, A. Hanjalic, H.-J. Zhang, and L.-H. Cai, “A flexible framework for key audio effects detection and auditory context inference,” IEEE Transactions on Audio, Speech and Language, Processing vol. 14, no. 3, pp.1026–1039, May 2006.

[3] P. Kathirvel, M. Sabarimalai Manikandan and K. P. Soman, “Automated Referee Whistle Sound Detection for Extraction of Highlights from Sports Video”, International Journal of Computer Applications, Vol. 12, No.11, pp. 0975 – 8887, January 2011.

[4] Z. Xiong, R. Radhakrishnan, A. Divakaran, and T. S. Huang, “Audio events detection based highlight extraction from baseball, golf and soccer games in a United Framework,” in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, vol. 5, pp.632–635, April 2003.

[5] R. Cai, L. Lu, H.-J. Zhang, and L.-H. Cai, “Highlight sound effects detection in audio stream”, Proceedings of the 2003 International Conference on Multimedia and Expo, vol. 3, pp. 37–40, 2003.

[6] J. Olajec, R. Jarina, M. Kuba, “GA-based feature extraction for clapping sound detection,” in Proceedings of 8th Seminar on Neural Network Applications in Electrical Enggineering, pp. 21–25, September 2006.

[7] R. Jarina, J. Olajec, “Discriminative feature selection for applause sounds detection,” in Proc. 8th Int. Workshop on Image Analysis for Multimedia Interactive Service, pp. 13–16, June 2007.

[8] I. Otsuka, R. Radharkishnan, M. Siracusa, A. Divakaran, and H. Mishima, “An enhanced video summarization system using audio features for a personal video recorder,” IEEE Transactions on Consumer Electronics, Vol. 52, No. 1, pp. 168–172, February 2006.