Golf Impact Detection with Audio Clues

Winston H.-M Hsu

Dept. of Electrical Engineering, Columbia University
winston@ee.columbia.edu

ABSTRACT

In this project, we detect impact sounds from TV golf games with audio clues only. Relevant features are exploited to best represent characteristics of impact sounds. Three classifiers are experimented to match impact sounds with high accuracy. By modeling impact sounds with unimodal multivariate Gaussian featured with Mel-frequency cepstral coefficients (MFCC), an encouraging result was approximately around 70% success rate.

1. INTRODUCTION

TV has been an important information source for long decades. Evolving with the rapid development of Internet and video technologies, Digital TV, streaming videos deliver more visual and audio impacts to home users who, nevertheless, have limited time to browse all available contents. Some highlight extraction technologies are promising with the trends especially those sports programs. There are lots of golf games in TV channels. However, very few researches are mined. A sequence of players planning, addressing, swinging and falling of the ball could be the highlight of the games. The impact video shot, characterized by a clear impact sound and fast motion are the key frames of highlights. In our approach, we focus on detecting impact shot through audio clues only without relying on those time-consuming visual features. It's a big challenge since impact sound is comparably short and often mixed with other signals.

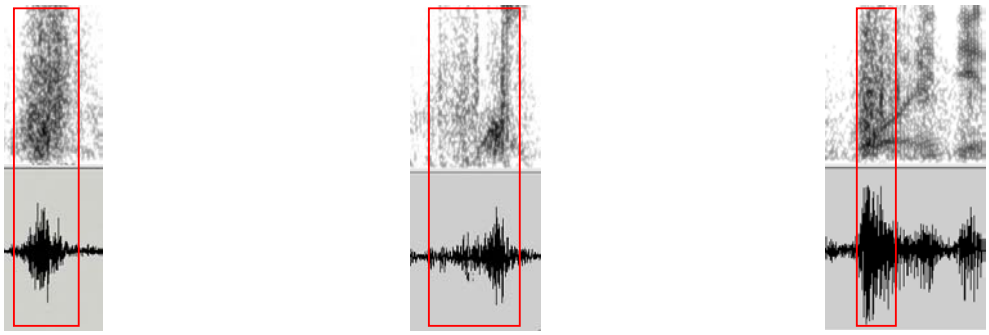
Prior work on golf impact detection is not seen in the literatures as far as we could see. A similar work might be [7], where 4 heuristic hit templates of baseball, described by sub-band energies, are used to detect the baseball hit with some distance measurements. There are also some works [2][3] about speech, music and non-speech classifications. In these tasks, at least a one-second window is required to discriminate audio types or extract some semantic meanings. On the other hand, we might treat impact detection as speaker recognition [4][6] concerned with extracting the identity of the person speaking the utterance. However, the recognition or identification task is completed through seconds of speech within controlled environments. As for golf impacts, with mini-seconds only, it's really significantly challenging to detect those precisely. Gladly with a simple parametric classifier, effective post-processing and features representing the impact sound, an encouraging result is discovered within this project. We had constructed a framework that could detect the impact sound with fair performance and seems invariant to production rules and environmental noises.

Our work differs the others since this might be the only work about golf highlights. Moreover we construct a framework driven by impact characteristics rather than heuristic rules and fused with temporal and spectral criteria of golf impacts.

In section 2 we discuss impact characteristics and feature selections; in section 3 we investigate experimented classifiers: Neural Network, unimodal multivariate Gaussian distribution and Gaussian Mixture Model (GMM); in section 4 we outline the results; in section 5 the future works are briefly addressed; and section 6 concludes this report.

2. FEATURE SELECTION

Before any detection tasks, we need to approximate impact sounds with right model and features. We have 27 MPEG-I video segments with the length ranging from 12 to 72 seconds. 25 of them are from Mandarin golf programs and 2 of them are Korean ones picked up from MPEG-7 testing suits. In each segment there is exactly an impact shot. The original audio track is 44.1 KHz, 16 bits and further down-sampled to 8 KHz for detection. In general the duration of impact sound is very short and is roughly surrounded by 3 to 27 audio frames (with 256 sample/window and 128 sample/hop). They approximately occupy 0.16% to 1.1% length of sample segments. With such a short duration, it's rather difficult to find semantic clues to identify the impact sound. Moreover, the impact sound is mostly mixed with speech, music, noise and environmental sounds. No doubt, some of them are not salient in the sample segments. In **Figure 1**, some sample impact sounds marked by the rectangle with correspondent spectrograms and waveforms are shown for references. In **Figure 1**-(c), the impact sound is mixed with speech formants. Meanwhile, all these impacts have different spectrograms and waveforms.



(a) (b) (c)

Figure 1. Three impact samples. (c) is mixed with speech.

2.1. Features

Eight audio features have been evaluated for use in this system. The features are calculated within each frame except *spectral flux*. A frame is of windowed samples, with 256-sample window size and 128-sample hop size, or with 128-sample overlap. **Figure 2** presents the system diagram from windowing samples, feature extractions, normalization, classification and post-processing. Actually these tasks are operated with intra-frame only. To get high semantic accuracy, we need consider temporal characteristics of impact sounds as well. It is then applied after classification tasks. The details are described in section 3.5.

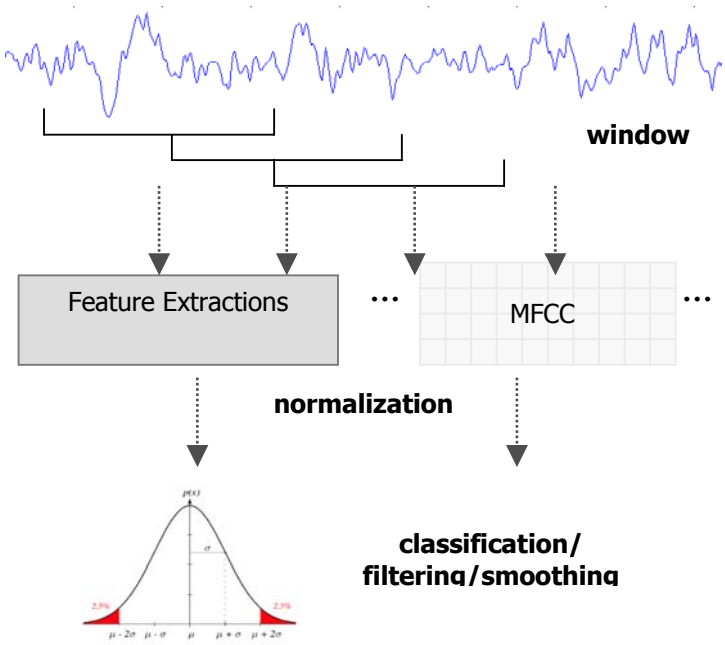


Figure 2. System diagram for golf impact detection with audio clues

The features adopted in this project are (please reference [5] for more details):

- Zero Crossing Rate (ZCR): The number of time-domain zero crossings within a frame.
- Short-Time Energy (STE): The average energy of samples within each frame. In general, frames of an impact would have higher STE.
- High Frequency Short-Time Energy (HFSTE): The average energy within high-frequency sub-bands of samples. Impact frames have high HFSTE than those of noises or minor collision.
- Centroid (CTD): The balancing point of the spectral power distribution.
- Roll-off (85%) (RLF): The 85th percentile of the spectral power distribution.
- Frequency Deviation (FDV): Standard deviation of spectral power distribution.
- Spectral Flux (SFX): The 2-norm of the frame-to-frame spectral amplitude difference vector.
- Mel-frequency Cepstral Coefficients (MFCC): See section 2.2 for more details.
- First Cepstral Coefficient (CP0): The zeroth coefficient of MFCC.
- Second Cepstral Coefficient (CP1): The first coefficient of MFCC.

Through the correlation analysis, some of those features correlate highly, as **Table 1**. We found that RLF and FDV are highly correlated with CTD and CP1 such that these two features are not used in the following classifications.

	ZCR	STE	CTD	RLF	FDV	SFX	CP0	CP1
ZCR	1.000	0.058	0.841	0.697	0.602	0.080	0.371	0.536
STE	0.058	1.000	0.049	0.077	0.142	0.779	0.892	0.102
CTD	0.841	0.049	1.000	0.921	0.830	0.096	0.415	0.770
RLF	0.697	0.077	0.921	1.000	0.937	0.013	0.278	0.881

FDV	0.602	0.142	0.830	0.937	1.000	0.092	0.161	0.951
SFX	0.080	0.779	0.096	0.013	0.092	1.000	0.742	0.060
CP0	0.371	0.892	0.415	0.278	0.161	0.742	1.000	0.154
CP1	0.536	0.102	0.770	0.881	0.951	0.060	0.154	1.000

Table 1. The covariance matrix between features

2.2. Mel-frequency Cepstral Coefficients

Here we adopted Mel-frequency cepstral coefficients as the major representation of impact sounds [8]. For each windowed sample, the log of the power spectrum is computed using a discrete Fourier transform. A non-linear map of the frequency scale perceptually weights the log spectral coefficients. This operation, called Mel-scaling, emphasizes mid-frequency bands in proportion to their perceptual importance. Then the Mel-weighted spectrum is transformed into cepstral coefficients. In our experiment the zeroth cepstral coefficient plays a vital part to detect the impact. It might be due to that the normalized energy level is an important characteristic for modeling golf impacts.

Initially, we tried to represent golf impacts with linear predictive coding (LPC) [1] and linear spectral pairs (LSP) [3][4]. LSP is another representation of the inverse filter constructed by LPC and the zeros of the filter are mapped onto the unit circle in the Z-plane through a pair of auxiliary polynomials.

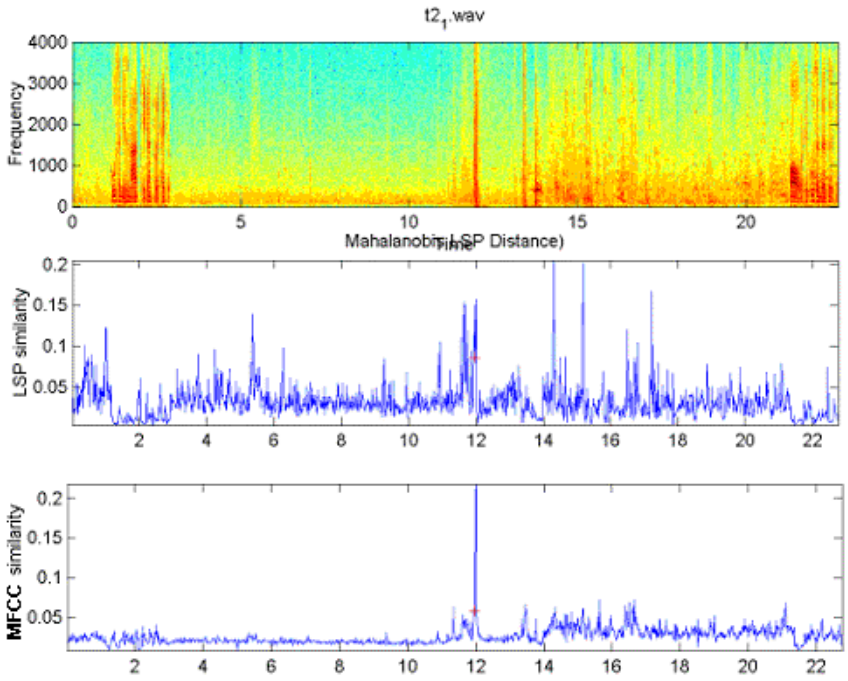


Figure 3. The similarities of the same audio clip measured by LSP and MFCC. The impact templates are extracted directly from the same audio surrounding the impact.

In **Figure 3**, LSP and MFCC are used to measure the similarity between impact sounds and testing data. In this test, the impact frame samples are directly extracted from the same sample except that they are represented in LSP and MFCC. From the result, MFCC has higher performance

as modeling the impact sound. From our experiment, LSP has low discriminability on noise and small impacts caused by other objects. The effect might be probably that LPC or LSP is a model-based representation and could be severely affected by noise [6]. At outdoors, the recordings might full of all kinds of noises.

2.3. Normolization

Since we hope to construct a generic framework for impact detection through audio clues, a confusion of absolute signal level is also an important issue to avoid. At this aspect, STE, HFSTE and each row of MFCC are all applied with online normalization, where each data sequence is locally normalized to have approximately zero mean and unit variance. With our proposed approach, invariant to production rules, extra two Korean golf programs (described in section 4) are detected successful without any framework or parameter modification.

3. CLASSIFICATION

Along with the experiment of this project, several approaches are taken with different outcomes. A nonparametric classifier, Neural Network, is firstly invoked and equipped with 6 features. We also seek to match impact patterns stochastically by modeling impact frames and with parametric classifiers.

3.1. Neural Network (NN)

A simple NN featured with ZCR, STE, CTD, SFX, CP0 and CP1 is firstly experimented to classify the samples, where the impact frames are marked with 1 and the other are labeled with 0. This classifier could somehow discriminate impact frames with other environmental sounds. However, it mixes with speeches that perceptually are easy to differentiate from impact sounds. According to the **Figure 4**, NN with those features does not perform well. The reason is that our positive training samples are very few (<1% of audio frames) and thus overcome by negative samples.

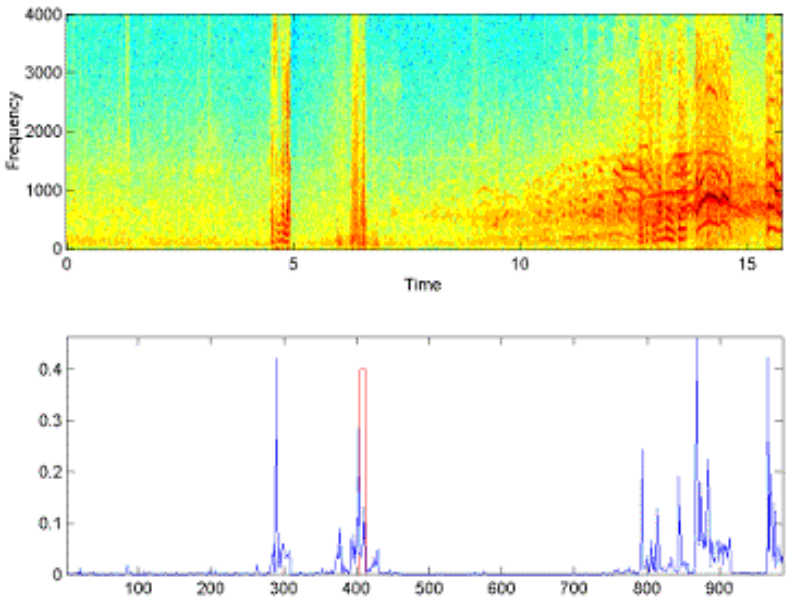


Figure 4. NN classifier featured with ZCR, STE, CTD, SFX, CP0 and CP1. The red line mark impact frames

3.2. Model Impact Sound

We further adopt pattern matching in this project, which requires a measurement of the similarity between the input feature vectors and impact models. Generally, template matching is categorized as *template models* and *stochastic models*. In early stage of speech processing, template matching was used intuitively. Here a stochastic model is used since it has more flexibility and results in a more theoretically meaningful probabilistic likelihood score.

The most promising result comes as we model MFCC coefficients as a unimodal, multivariate Gaussian distribution. The sample audio frames covered by an impact sound was collected and form the template, which was further used to calculate the similarity by way of Mahalanobis distance. The first coefficient of cepstrum, C0, plays a vital part in discriminating impact with other waves. We did experiment to use coefficients 2 to 13 only. The result is quite poor comparing with those of coefficients of 1 to 13.

3.3. Euclidean Distance Measurements

In pattern matting, we have to define a similarity between the target and testing samples. The similarity is generally interpreted as the inverse of metric distance by some measurements. We adopted the Euclidean distance for the first trial, which turns out to provide poor discriminability, however. The Euclidean distance has an intuitive appeal as it is commonly used to evaluate the proximity of objects in two or three-dimensional space. It works well when a data set has "compact" or "isolated" clusters. The drawback is the tendency of the largest-scaled features to dominate the others such as C0 of this project.

3.4. Unimodal Multivariate Gaussian

For stochastic model, we approximate the impact sound with unimodal multivariate Gaussian distribution, which is characterized by a mean vector μ and covariance matrix Σ. The likelihood of a feature vector and the impact could be represented as the following:

$$p(x \mid \text{impact}) = (2\pi)^{-n/2} \mid \Sigma \mid^{-1/2} \exp\left[-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right].$$

Taking the log of the likelihood, we further got,

$$\log\{p(x \mid \text{impact})\} = -\frac{n}{2}\log 2\pi - \frac{1}{2}\log \mid \Sigma \mid - -\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu).$$

Since the first two terms of the right part of the previous equation are constants, it often takes the remaining part as a measurement of distance between the feature vector x and the mean of the model. This is often referred as the *Mahalanobis distance* d_M^2, where

$$d_M^2 = (x-\mu)^T \Sigma^{-1}(x-\mu).$$

Actually, Σ^{-1} is used to normalize each dimension of features by its correspondent variance in order to prevent from being predominated by specific features with large quantities. Empirical results [9] suggest that covariance matrix Σ could be simplified as a diagonal matrix carrying variance of each dimension and leaving the other parts zero.

The higher d_M^2, the lower is the similarity between a sample with features x and the impact model. Here we further define the similarity of each frame measured with Mahalanobis distance as: $Sim_M(x) = 1/d_M^2$.

3.5. Filtering and Smoothing

The similarity measurement with Mahalanobis distance is then filtered with thresholds of STE or HFSTE. The similarity of a specific audio frame is reduced to zero if its STE or HFSTE value is lower than a specific threshold, meaning that it might not be a right impact but some miscellaneous hits caused by object collisions, cough or other noises. The "open" Morphological operation is further invoked to remove outliers that have just 1, 2 or 3 successive frames and impossibly form a solid impact sound. This operation guarantees the temporal continuity or length of golf impacts. Through **Figure 5**, we could see that HFSTE plus Morphological operations perform very well because HFSTE does remove those noises with high similarity with impact sounds but having low energy or continuity at high frequency sub-bands.

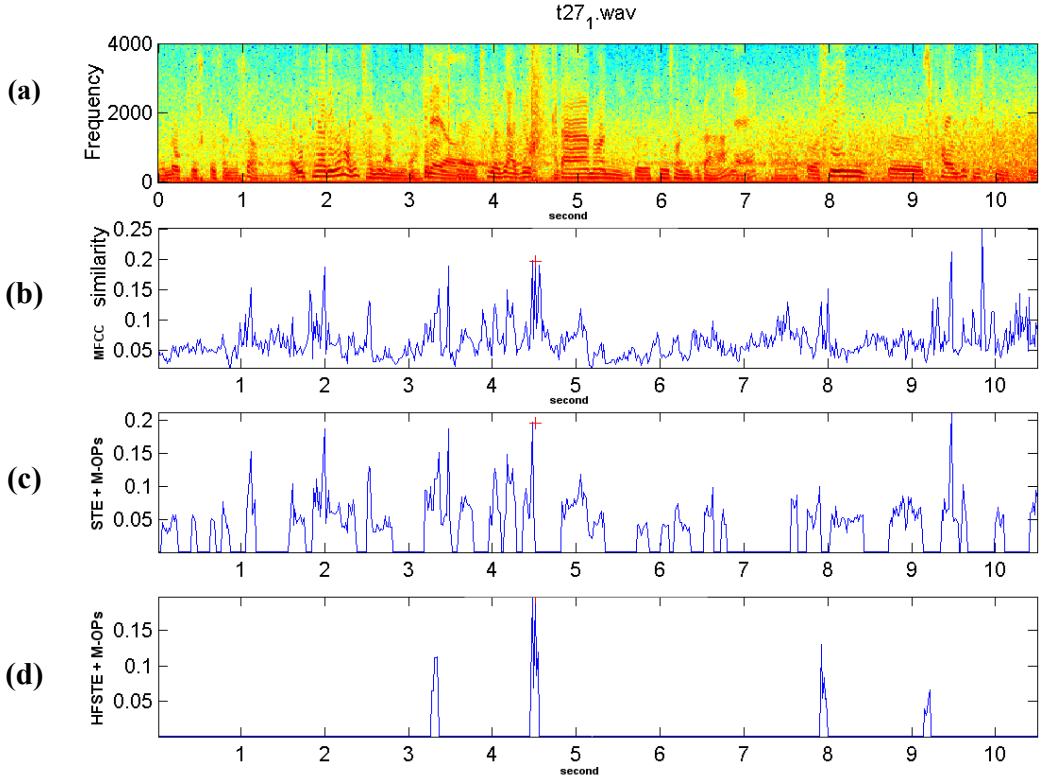


Figure 5. (a) Spectrogram of a Korean golf segment. (b) MFCC similarity measured with Mahalanobis distance. (c) Filter with STE threshold and smooth with Morphological operations. (d) Filter with HFSTE threshold and smooth with Morphological operations.

3.6. GMM

In [6], a Gaussian mixture model is suggested and derived from weighted sum of M Gaussian components and given by the equation:

$$p(x \mid \text{impact}) = \sum_{i=1}^{M} p_i b_i(x) \quad ,$$

where $b_i(x)$ is the Gaussian component density; p_i is the mixture weight with $\sum_{i=1}^{M} p_i = 1$ and

$$b_i(x) = (2\pi)^{-n/2} \mid \Sigma_i \mid^{-1/2} \exp\left[-\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i)\right],$$

with mean vector μ_i and covariance matrix Σ_i. The mean vectors, covariance matrices and mixture weights from all component densities parameterize the complete Gaussian mixture model. These parameters are collectively represented by the notation,

$$\Theta = \{p_i, \mu_i, \Sigma_i \mid i = 1,...,M\}.$$

There are some advantages of using GMM to model impact sounds. Firstly it has the ability to form smooth approximations to arbitrarily shaped densities. Meanwhile the impact samples might be caused by varieties of clubs, swinging power, weather, noise…etc. We might have more precise result by modeling impact sounds with this model. However, it requires a comparable amount of samples to train the GMM. Meanwhile, during the training, the singularities of the covariance matrix might arise when there is not enough data to sufficiently train a component's variance vector or when using noise-corrupted data. Due to time limitation and sample counts, this part is not completed but believed to bring great performance improvement.

4. RESULTS AND DISCUSSIONS

To measure the performance, we define two measurements, *rigid hit* and *fuzzy hit:*

- *Rigid hit*: the impact frames have the highest similarity with training samples. The result is apparently a hit.
- *Fuzzy hit*: the impact frames are not necessarily a rigid hit but not 30% smaller than highest similarity frames. This measurement approaches non-optimal solutions. It provides candidates of prospective impacts which is much more few than raw audio frames. If intending to apply expensive visual classifications, we might work on these fuzzy hits only.

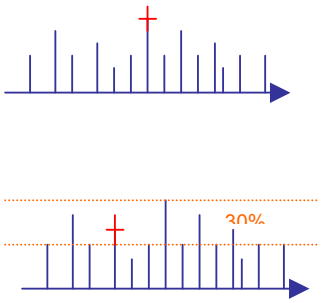


Figure 6. Rigid hit v.s. Fuzzy hit

In our evaluation phase, 7 video segments are used for impact modeling; 21 video segments are collected for testing, two of which are from Korean golf programs. Then the performance is as the following:

- *Rigid hit*: **14** hit among **21** video segments (67%), where Korean games are successfully hit.
- *Fuzzy hit*: **16** hit among **21** video segments (76%).

The result is quite encouraging and much better than we expected as initializing the project. More promisingly, it approaches the other work that detects baseball impacts. Till now, our proposed approach still has large rooms to improve by increasing testing samples, modeling with GMM or investigating other smoothing filters and thresholds.

Those missed segments all have short duration of impact time or with very low energy level. 3 of the impacts are hardly recognizable if listening to audio tracks only. Those low energy impact frames are truncated after HFSTE filtering or Morphological operations characterized with spectral and temporal thresholds. This might be the major reason why some impacts not detected by our approach.

More interestingly, those two Korean video segments are "*rigid hit*" meaning that those impact frames are with the highest similarity. Our golf samples and thresholds are all trained from the other set of video segments with different production rules and energy levels. However, we still get the promising result suggesting that our classification framework might really catch true characteristics of golf impacts but based on few assumptions.

5. FUTURE WORKS

Our current result is encouraging and leads to the right way that models golf impact with MFCC features and applies pattern matching with stochastic models. The result is believed to improve a lot if more testing samples are drawn. Furthermore, with more samples, we could decide better thresholds and filtering rules. Moreover, a GMM could approximate golf impact sounds mixed with environment sounds and caused by different clubs. This approach is also subjected to few training samples. Besides, further incorporated with visual clues by applying on those small amounts of candidate frames of *fuzzy hit* could also increase the precision and recall rate, however, still preserve the simplicity and reduce cost of computing power.

6. CONCLUSION

Modeling and detecting a specific audio event with short duration and mixed with noise is challenging. In this work, golf impact detection with audio clues, we had constructed a simple but not poor classification framework that could detect golf impact and further used as relevant clues for highlight detection in TV golf programs. It's computation-inexpensive. Moreover, by exploiting characteristics of golf impacts, this approach is almost invariant to product rules and insensitive to interferences of noise, environmental sounds and speeches.

REFERENCES:

[1] L. Rabiner and B.-H. Juang, "Fundamentals of Speech Recognition," Prenticl Hall, 1993

[2] T. Zhang and C.-C. Jay Kuo , "Content-Based Classification and Retrieval of Audio," SPIE's 43'rd Annual Meeting, Conference on Advanced Signal Processing Algorithms, Architectures, and Implementations VIII, San Diego, July 1998

[3] L. Lu, H. Jiang, and H, J. Zhang, "A Robust Audio Classification and Segmentation Method," Microsoft Research, China

[4] J. P. Campbell, "Speaker Recognition: A Tutorial," Proceedings of the IEEE, Vol. 85, No. 9, Sept. 1997

[5] E. Scheirer, M. Slaney, "Construction and Evaluation of a Robust Multifeature Speech/Music Discriminator," Proc. ICASSP-97, April 21-24, Munich, Germany

[6] D. A. Reynolds and R. C. Rose, "Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models," IEEE Trans. On Speech and Audio Processing, Vol. 3, No. 1, Jan. 1995

[7] Y. Rui, A. Gupta and A. Acero, "Automatically Extracting Highlights for TV Baseball Programs," Microsoft Research, US

[8] Malcolm Slaney, "Auditory Toolbox," http://rvl4.ecn.purdue.edu/~malcolm/interval/1998-010/

[9] C. Bechhitti and L. P. Ricotti, "Speech Recognition: Theory and C++ Implementation," John Wiley & Sons