Gender Classification based on Audio Features

Asst. Prof . Dr.Nidaa F. Hassan **Sarah Qusay Selah Alden**

Computer Science Department, University of Technology/Baghdad

Abstract

Gender audio classification is considered one of the most significant methods in audio processing. In this paper, an algorithm involving audio features (mean, standard deviation, zero crossing, Amplitude) and Support Vector Machine (SVM) is presented to perform speaker gender recognition. For each audio, the highlights vector is utilized as an info vector in the SVM algorithm. An example of 2270 audio, include 1132 female audio with 1138 male audio is analyzed based on this algorithm. With only the four features, the average prediction error is 5%.

Key words: digital audio, mean, standard deviation, amplitude, zero crossing, support vector machine (SVM).

تصنيف نوع الجنس استناداً إلى مميزات الصوت

أ.م.د. نداء فليح حسن **ساره قصي صلاح الدين**

الجامعة التكنولوجية – قسم علوم الحاسبات

المستخلص

إن نوع الجنس بالاعتماد على الصوت يعد في الاونه الاخير من أهم العمليات في معالجة الصوت. في هذا البحث ,تم تقديم خوارزمية تنطوي على مميزات الصوت (الوسط الحسابي ، الانحراف المعياري، عبور الصفر و السعة) و الة دعم التصنيفSVM لتنفيذ التعرف على جنس المتحدث.لكل إشارة، يتم استخدام قيم المميزات كمتجه إدخال في خوارزمية آلة دعم التصنيف. تم تحليل العينات الصوتيه استنادا إلى هذه الخوارزمية والتي يبلغ عددها ٢٢٧٠ إشارة، تحتوي على ١١٣٢ صوت أنثى و ١١٣٨ صوت ذكر،وباستخدام المميزات الأربع فقط , قيمه متوسط خطأ التنبؤ قد بلغ ٥٪.

1- Introduction

Machine Learning community has been developed Kernel-based algorithms, there has been great benefit from the development of these algorithms, SVM is used in handling nonlinear errors and also can solve problems with high dimensions, and this has led to remarkable development in the use of Machine Learning, so it's currently applying different types of problems to be solve successfully, the algorithm is called (SVM), and there is comprehensive literature of SVM in [1], and kernel-based algorithms is in [2].

Machine Learning has been deemed as one of the most effective branches of Kernel technique, utilized in different domains like : SVM algorithm which has been of direct application to standard detection and estimation, prior knowledge of integration in learning process, either by using virtual training samples or by constructing a relevant kernel for the given problem . These applications contain extraction of audio features [3] , speech, audio processing [4], speaker identification [5] , image processing [6], text categorization[7] , audio signal segmentation [8] and these are not all existing problems that can be resolved by kernel methods[9].

2- Related Works

Many researchers have focused on audio classifications; some of these researches are listed below:

1- Stan Z. Li and Guo Dong. G (2003) [8], in this paper, a support vector machines (SVMs) is proposed for content-based audio classification and retrieval. Given a feature set, which is composed of perceptual and spectral feature, optimal class boundaries between classes are learned from training data by using SVMs. Matches are ranked by using distances from boundaries , also many compassion with

different classification methods with feature sets are accomplished to evaluate the performance.

2- Nicolas .S, Daniel .M (2005) [10], this paper describes the algorithm for Audio Genre Classification. This algorithm contains a total of 2929 songs of music: timbre, energy and rhythm. Once features are extracted, a mixture of Support Vector Machines (SVMs) is used for classification into musical genres.

3- Cyril. L, Perfecto. H (2007) [12], they explored how a content-based similarity measure can help to classify by mood a collection of music files. In this algorithm, they train a SVM with many descriptors empirically selected. It uses a set of 133 descriptors and a Support Vector Machine classifier to predict the mood cluster. The features are spectral, temporal and tonal and loudness. The features were selected previously according to experiments on annotated databases.

4- Audio Feature Selection

This step is very important in the selection of audio classification characteristics to get high accuracy of division and classification, determine the good characteristics, including the selection of temporal and spectral characteristics of the audio signal [11].

Features are divided into two kinds; thes are:

(i) Mel-frequency Cepstral Coefficients (MFCCs)

(ii) Perceptual Features.

These features can be combined with one advantage vector after adjustment. At first, an audio signal is altered to a genre, which is 8 KHz, 16-bit, and mono-channel. Then it is emphasized with parameter 0.98 to balance the inherent

spectral tilt, and then divided to non-over lapping sub clips. The performance of each interval of the different sub-section is tested with classification in experiments [13]**.**

5- Zero-Crossing Rate (ZCR)

The number of time-domain zero-crossings is within a frame called ZCR. Signal frequency content can be easily measured:

$$ZCR = \frac{1}{2(N-1)} \sum_{m=1}^{N-1} |sgn[x(m+1)] - sgn[x(m)]| \ldots\ldots\ldots (2)$$

Where sgn [.] is a sign function and x (m) is the discrete audio signal, m = 1 ...N.

In general, an audio signal consists of alternating expressive audios and non-paced audios in the section rate, the music signal do not include these types of frame. Hence, for the signal, the zero exchange rate variance is generally greater than the music signals. ZCR is a good distinction between audio and music. Given this, almost systems are used [4, 5, 7-10] button to classify audio [14].

3.1 Mean

The mean refers to an intermediate value between a separate set of numbers, a set of variables divided by total variables. The computation of the arithmetic is almost identical to the calculation of the statistical data of the population and the mean of the sample, with slight variations in the variables used [15]:

$$\bar{x} = \frac{1}{N} \sum_{i=1}^{N} xi \ldots\ldots (3)$$

Where $\bar{x}$ is a mean, N is the numbers of variables, xi is the location of the element in the array.

3.2 Standard Deviation

A standard deviation is a measure that the sum by which each value contrasts inside an informational index from the mean. It viably shows how to limit the qualities in the informational index ushich are assembled around the mean esteem. It is the most powerful and widespread measure of dispersion since then, in contrast to scale and quadrature; it takes into account each variable in the data set. When the values in a data set are grouped together closely, the standard deviation is small, while the values are distributed separately, the standard deviation will be relatively large. The standard deviation is typically displayed in conjunction with the mean and measured in similar units.

For a limited set of numbers, the standard deviation is obtained by taking the square root of the mean of the square deviations of the values from their average value.

The standard deviation is like the average deviation, aside from the average frequency with control instead of sufficiency. This is accomplished by squaring each of the deviations previously taking the mean (recall, power% voltage 2). To complete, the square root is taken to make up for the underlying quadrature. As a condition, the standard deviation [15] is calculated; in the following way.

$$\sigma^2 = \frac{1}{N-1}\sum_{i=0}^{N-1}(xi - \bar{x})^2 \ldots\ldots (4)$$

Where σ^2 is standard deviation, N is the numbers of variables, xi is the location of the element in the array and $\bar{x}$ is the mean.

3.3 Amplitude

Greater changes in the atmospheric pressure from high to low are allowed, when the audio has a large capacitance. Capacity is almost always a comparative measurement, since at the lowest-amplitude end (silence), almost air whit is in movement and at the highest end, the amount of compression and rarefaction, though finite, is the maximum. In electronic circuits, amplitude may be increased by extending the degree of change in fluctuate electrical current. A woodwind player may increase the amplitude of their sound by providing maximal force in the air column i.e. blowing harder.

Acoustic energy or intensity of a sound is directly related by amplitude. Both intensity and amplitude are associated to sound's power.

$$s(t) \times \cos wct = s(t) \cos wct \ldots\ldots (5)$$

Where:
S (t): Is the carrier of the information.
G (t): Is a carrier wave and a sine wave ($\cos wct$).

The **Amplitude** is measured by the force exerted on a particular area. The Newton's per square meter (N/m2) is the most common, unit of measurement of force applied to an area for acoustic study. Figure 2 shows that the amplitude depends to a great extent on estimations of the motions in barometric pressure from one extraordinary (or peak) to the next. The degree of change above or below and imaginary center value is referred to as the peak amplitude or **peak deviation** of that waveform [14].

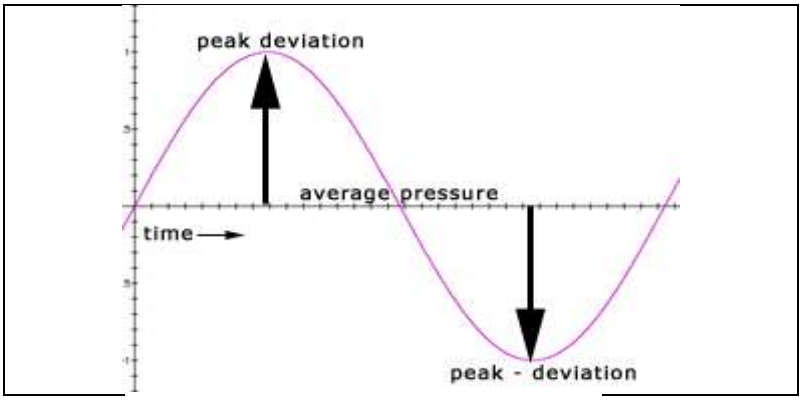

Figure 1: Amplitude of audio signal.

6- Pattern Recognition via SVMs

SVMs idea is defined as a limit between two classes by main separation of the closest observations. In practice, SVMs are useful algorithm on binary classification functions. In figure 2 the SVMs idea is shown in the graph below [14].

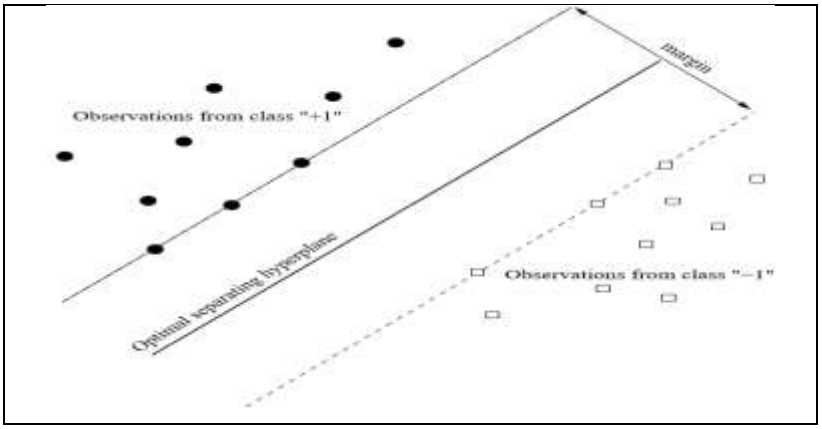

Figure 2: SVMs idea is to maximize the margin.

Consider a dataset D = {(xi, yi), xi ∈Rp, yi∈ {−1, 1}} n i=1 that is linearly separable. The margin is defined as the shortest perpendicular distance between the hyperplane and the observations, referring to the width of the blank region

separating two data clouds. The goal for SVMs is to maximize the margin. Any hyperplane can be written as

$w \cdot x - b = 0$…… (6)

The w is a coefficient vector and b is fixed. While the info can be linearly separable, there exist two hyper planes that can separate the data completely and no points fall in between. That can be defined as:

$w \cdot x - b = -1$ …… (7)
$w \cdot x - b = 1$ ……..(8)

That region in between is the margin. It can be shown that maximizing the margin is equivalent to minimizing k w k. Finally, the classification can be achieved by

$$class\ (xi) = \begin{cases} 1 & w.xi - b > 0 \\ -1 & w.xi - b \leq 0 \end{cases} \text{ …….. (9)}$$

When the data are not linearly separable, kernel functions play an important role by linearizing the data. That is, when the data are not linearly separable, they can be transformed use function K (x, y) into the inner product space in which it is feasible to separate them linearly. Common kernel functions are the radial rule function kernel (RBF kernel) [14].

$$k(x,y) = exp\left(-\frac{\|x-y\|2}{2\sigma^2}\right) \text{ …… (10)}$$

And the polynomial kernel
$k(x,y) = (x.y + c)^d$ …… (11)

7- Precision and Accuracy

These terms (Precision and Accuracy) used to describe systems and methods are Precision and accuracy that are used for measurement, estimation, or prediction. In all these cases, there is a parameter that must be known; which is called the true value, or simply, truth as close as possible to the true value, so the ways of describing the error that can exist between these two values are precision and accuracy [15].

Precision (P) is characterized as the number of true positives (T_p) over the number of true positives in addition to the number of false positives (F_p).

$$\boldsymbol{P} = \frac{\boldsymbol{T_p}}{\boldsymbol{T_p + F_p}} \ldots\ldots\ldots (12)$$

Recall (R) is characterized as the number of true positives (T_p) over the number of true positives in addition to the number of false negatives (F_n).

$$\boldsymbol{R} = \frac{\boldsymbol{T_p}}{\boldsymbol{T_p + F_p}} \ldots\ldots\ldots (13)$$

These quantities are also related to the (F_1) score, which is defined as the harmonic mean of precision and recall.

$$\boldsymbol{F1} = \boldsymbol{2} \times \frac{\boldsymbol{P \times R}}{\boldsymbol{P + R}} \ldots\ldots\ldots (14)$$

8- The Proposed Algorithm

In this proposed algorithm the audio signal is passed through a number of processes to classify gender sound (male from female).The gender classification process is applied to CMU_ARCTIC (Carnegie Mellon University) audio dataset with format Wave [16].

The idea of algorithm is focused on the types of extracted features that produced satisfied results from SVM, since extracted features reduce the original data set, and made classification process to distinguish one audio signal from another easier. The SVM depends on four features; Meaning, Standard Deviation, Amplitude and ZCR.

Figure 3 shows the diagram of gender classification process, the audio gender classification includes below steps below:

1. Input of the main digital audio signal.
2. Training and Testing process.
3. Gender classification process based on SVM value.
4. Precision, Recall and F1- Measure.
5. Output gender detection (Male or Female).

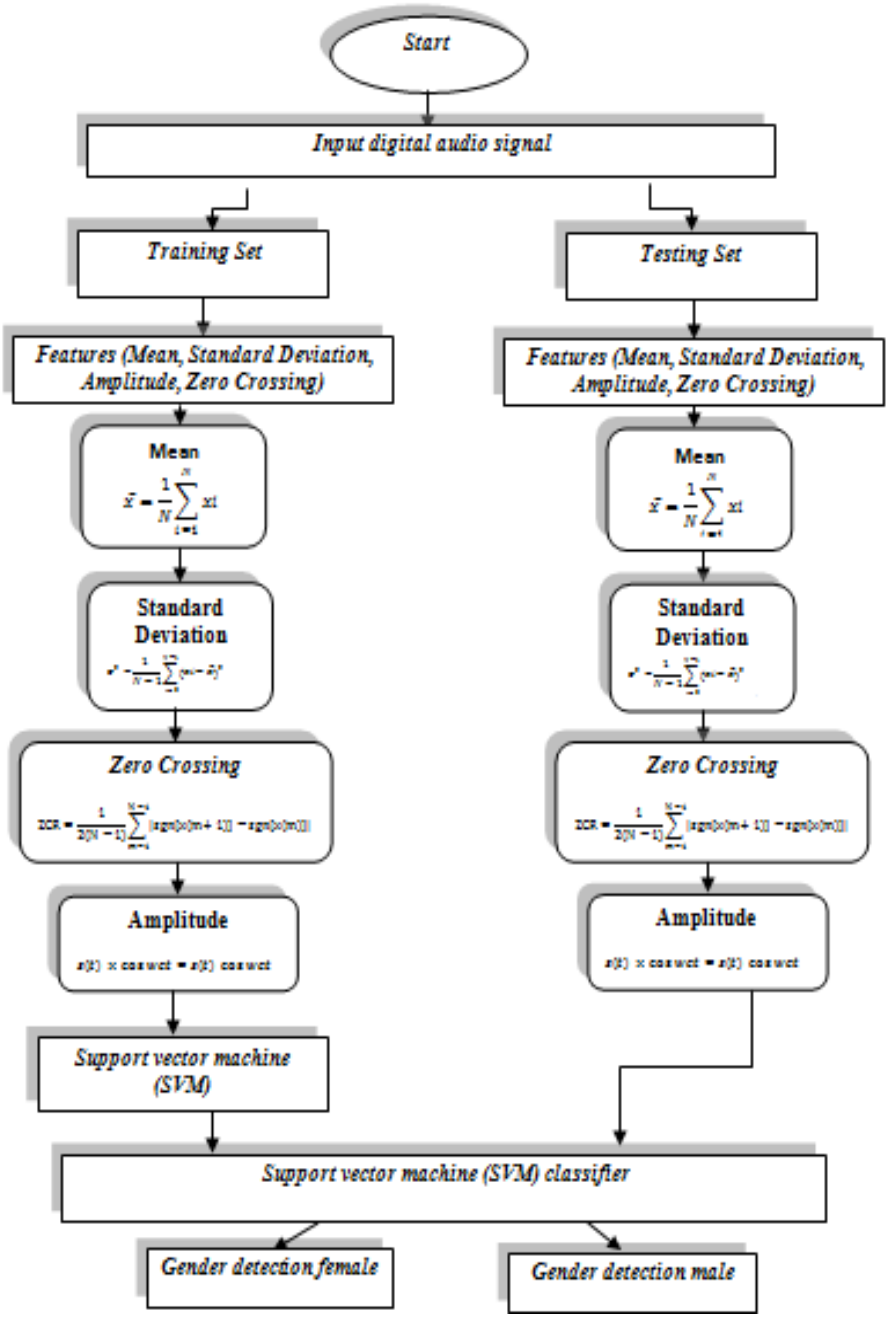

Figure 3: Block diagram of the proposed Gender audio classification.

1. Input the original digital audio signal

The first step in gender audio classification is accomplished by collecting audio signal file as audio samples; these samples are stored in a buffer. This signal is considered to be input signal; figure 4 shows the signal of audio samples.

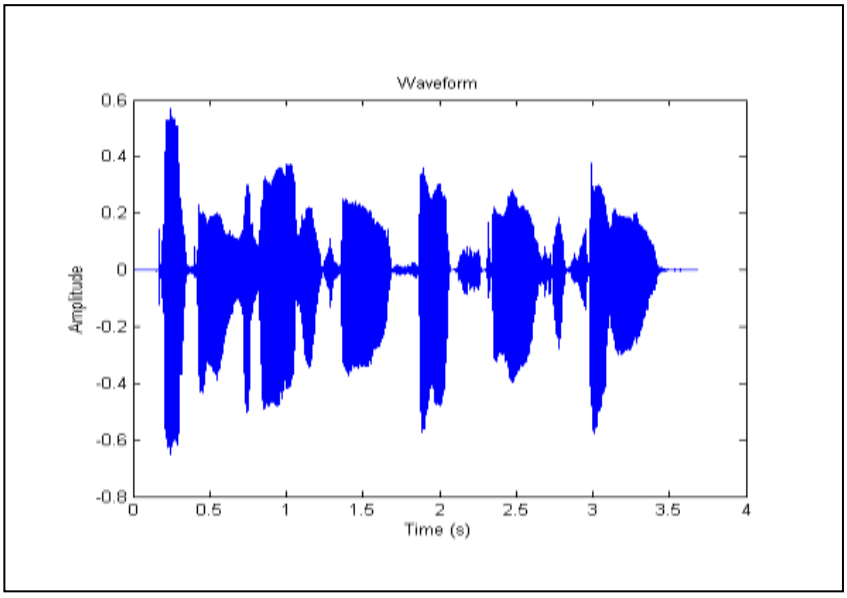

Figure 4: Audio Signal.

2. Training and Testing Process

The proposed algorithm is implemented in two phases (training and testing). These experimental results compare four features. In these steps, the following process is applied to classifying gender:

• The zero, average, standard deviation and capacitance crossing rate are calculated. The zero transit rates are an important parameter for the expressed / non-invoiced

classification; it is often used as part of the front-end processing.

• The number of zero transit is the indicator of the frequency at which energy is concentrated in the signal spectrum. Speech that is expressed due to the excitation of the acoustic tract is produced by the periodic flow of air into the oyster and usually shows a zero-crossing low, while sudden speech produces a constriction of the vocal tract narrow enough to cause turbulent airflow which leads to noise and shows a zero height crossing count. The following figure shows the implementation of the zero crossing rating.

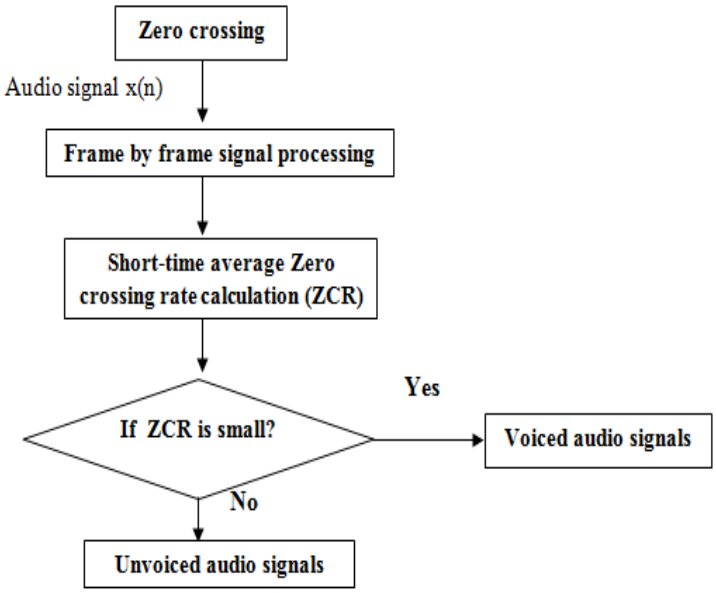

Figure 5: Block diagram of the ZCR classification.

- Dataset load in new file and input in training and testing set using features (Mean, Standard Deviation and Amplitude).
- Input audios training in "Classifier SVM"; using 75% audio in training set input audios testing in "SVM", using 25%

audio in testing set. The size uses full frame (16 bit frame).Table 1shows the results values of applying (Mean, Standard deviation, Zero crossing and Amplitude) for 6 audio samples.

Table 1: Explains the information features of samples waves.

Sample of Wave	*Mean*	*Standard deviation*	*Zero crossing*	*Amplitude*
Sample_1	*-3.623186005342605E-9*	*0.12618067542570324*	*2532.7001953125*	*0.014791988129531674*
Sample_2	*-3.436664203265766E-9*	*0.12206839535617431*	*2052.509765625*	*0.014291381835 9375*
Sample_3	*-3.593927437331763E-9*	*0.14360648109816335*	*2165.814208984375*	*0.016927880842 94562*
Sample_4	*-1.8024444580078125E-7*	*0.08314673759720534*	*1433.5*	*0.013346852302 55127*
Sample_5	*1.2775421142578124E-6*	*0.057249432419751854*	*1006.7999877929688*	*0.007084475708 007813*
Sample_6	*3.37982177734375E-7*	*0.06396277733290304*	*1010.0*	*0.007222145843 505859*

3. Precision, Recall and F1- Measure

The basic idea is to compute all precision and recall of all the classes, then average them to get a single real number measurement.

Table 2 shows the result of using four features in support vector machine (Mean, Standard deviation, Zero crossing and Amplitude) signals. This table contains information of classification (Training Dataset, Testing Dataset, Hit, Miss, Precision, Recall, and F1- Measure).

Table 2: Explains the information to samples of waves uses the measures in (SVM).

Training Dataset	***Testing dataset***	***Hit***	***Miss***	***Precision Rate***	***Recall Rate***	***F1-measure Rate***
1703	**567**	**537**	**30**	**95%**	**32%**	**48%**

The time taken for the process when using these features (mean, standard deviation, zero crossing, amplitude); in SVM (support vector machine) to show the result is 11205millisecond.

Table 3 shows the difference between the results of this Paper with other papers by Precision Rate and Recall Rate to distinguish the Papers.

Table 3: Explains the difference between the proposed algorithm and other papers.

Papers	***Features***	***Precision Rate***	***Recall Rate***
"Automatic Extraction of Musical Structure Using Pitch Class Distribution Features"[15]	Constant-Q Profile, Pitch Class Profile (PCP) and Harmonic Pitch Class Profile (HPCP)	82%	84%
The proposed algorithm	mean, standard deviation, zero crossing, Amplitude	95%	32%

9- Conclusion

In this paper, audio signal classification to determine gender class (male or female) is proposed. This algorithm depends on four extracted features . Mean, Standard Deviation, Amplitude and Zero Crossing rate. The values of these properties are categorized into two categories using one of the automated algorithm supported by the support vector machine (SVM). The accuracy results indicate that the rated capacity of four attributes is (95%), which is better than accuracy rated (84%) achieved by using HPCP.

References:

[1] V. Vapnik, **"Statistical learning theory"**. NY: Wiley, 1998.

[2] B. Sch¨olkopf and A. Smola, **"Learning with Kernels. Cambridge",** USA: MIT Press, 2002.

[3] C. Burges, J. Platt, and S. Jana, **"Extracting noise-robust features from audio data",** in ICASSP, Orlando, FL, 2002.

[4] N. Smith and M. Gales, **"Speech Recognition using SVMs,"** in T.G. Dietterich, S. Becker, and Z. Ghahramani, editors, Advances in Neural Information Processing Systems 14. MIT Press, pp. 1–8, 2002.

[5] V. Wan and S. Rentals, **"Evaluation of kernel methods for speaker verification and identification,"** in ICASSP, Orlando, FL, 2002.

[6] A .Rabaoui, H. Kadri, Z. Lachiri and N. Ellouze ,**"One-class SVMs challenges in audio detection and classification applications"**, 2008.

[7] Lie Lu, Hong-Jiang Zhang, Stan Z. Li, **"Content-based audio classification and segmentation by using support vector machines"**, Springer-Verlag 2003.

[8] M. Davy and S. God sill, "**Detection of Abrupt Spectral Changes using Support Vector Machines. An Application to Audio Signal Segmentation,"** in IEEE ICASSP, vol. 2, Orlando, USA, May 2002, pp. 1313–1316.

[9] Asma. R, Hachem. K, Zied. L and Noureddine .E, "**One-class SVMs challenges in audio detection and classification applications**", 2008.

[10] Nicolas .S, Daniel .M, "**A Mixture of Support Vector Machines for Audio Classification**", Signal Processing Institute (ITS-LTS-3), Swiss Federal Institute of Technology,(EPFL)Lausanne, CH-1015 Switzerland, 2005.

[11] Cyril. L, Perfecto. H, " **Automatic Detection of Emotion in Music: Interaction with Emotionally Sensitive Machines**", Music Technology Group Universitat Pompeu Fabra Barcelona, Spain, 2007.

[12] Amanita .Ch, and HimadriNath .M, "**Segmentation to Sound Conversion**", IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 16, Issue 3, Ver. V (May-Jun. 2014), PP 44-48.

[13] Bachu .R, Kopparthi. S, Adapa. B, Barkana. B, **"Separation of Voiced and Unvoiced using Zero crossing rate and Energy of the**

Speech Signal", Electrical Engineering Department School of Engineering, University of Bridgeport, 2015.

[14] Steven W. Smith, "**The Scientist and Engineer's Guide to Digital Signal Processing**", California Technical Publishing San Diego, California, 1999.

[15] Fokoue, Ernest and Ma, Zichen, "**Speaker Gender Recognition via MFCCs and SVMs**", Rochester Institute of Technology RIT Scholar Works, 2013.

[16] Index of /cmu_arctic/cmu_us_bdl_arctic/wav, **http://www.speech.cs.cmu.edu/cmu_arctic/cmu_us_bdl_arctic/wav/?C=D;O=A**.