Proceedings of the 2009 IEEE International Conference on Systems, Man, and Cybernetics
San Antonio, TX, USA - October 2009

Classification of Non-Speech Human Sounds:

Feature Selection and Snoring Sound Analysis

Wen-Hung Liao, Yu-Kai Lin
Department of Computer Science
National Chengchi University
Taipei, TAIWAN
whliao@cs.nccu.edu.tw

***Abstract*—Human sounds can be roughly divided into two categories: speech and non-speech. Traditional audio scene analysis research puts more emphasis on the classification of audio signals into human speech, music, and environmental sounds. We take a different perspective in this paper. We are mainly interested in the analysis of non-speech human sounds, including laugh, scream, sneeze, and snore. Toward this goal, we investigate many commonly used acoustic features and select useful ones for classification using multivariate adaptive regression splines (MARS) and support vector machine (SVM). To evaluate the robustness of the selected features, we also perform extensive simulations to observe the effect of noise on the accuracy of the classification. Finally, for the class of snoring sounds, we propose a robust approach to further categorize them into simple snores and snores of subjects with obstructive sleep apnea (OSA).**

***Keywords*—audio classification, acoustic features, feature selection, snore analysis.**

I. Introduction

In the past decade, spoken language processing has achieved significant advances. Recognition rate of human speech can exceed 95% in many applications [1]. However, the accuracy of emotion recognition from speech is relatively low, ranging from 50% to 60% [2]. This is caused a mismatch between the training data and the test data. Voices purposely articulated professional actors are usually employed as training data. The test data, however, are collected in a quite different setting. To get around this issue, we resort to a different approach in this paper. Instead of trying to infer emotion from speech, we proposed to detect an individual's status using non-speech human sounds.

Non-speech human sounds include many kinds of audio, such as speech, coughing and screaming. In [3], we have developed a hierarchical audio classification scheme to classify audio data into speech, non-speech and environment sounds. In this paper, we focus on the classification of non-speech signals. This is based on the assumption that the mood changes not only appear in speech, but also reveal in non-speech manners. For instance, screaming usually signifies scare and the emotion state can be labeled as *negative*. Laughing, on the other hand, is labeled as *positive*. Additionally, coughing may indicate a negative status, and the temperature should be adjusted accordingly in a smart home environment.

Many feature-based audio classification paradigms have been proposed in recently years. Wang *et al.* [4] presented an environmental sound classification architecture using MPEG-7 descriptors. The classifier utilizes the support vector machine (SVM) and k-nearest neighbor (KNN).

On the other hand, feature analysis and selection has gained increasing attention recently. Rong *et al.* [5] developed an automatic feature selector based on a RF2TREE and the C4.5 algorithm. The selector was applied to find the important acoustic features used in emotion speech recognition. According to their experiment, pitch and energy-related features are the most important.

Jarina *et al.* [6] focused on the detection applause sounds in audio stream with Mel-scale frequency cepstral coefficients (MFCC). A set of useful coefficients were selected from all the MFCC coefficients by genetic algorithm (GA) and simulated annealing (SA). In this paper, we will also study the properties of various bands of MFCCs for the classification of non-speech human sounds.

The hierarchical audio classification scheme proposed and adopted in this paper is depicted in Fig.1 [3]. The main objective is to classify non-speech human sounds into scream, sneeze, snore and laugh, with potential applications in smart living spaces. Toward this end, we investigate two sets of acoustic features and compare their efficacy in the classification of non-speech human sounds. Set 1 includes 7 commonly used audio descriptors that have definite physical properties. Set 2 includes 20 MFCC features. We also perform extensive simulations to observe the effect of noise on the accuracy of the classification.

In a related application, we also consider robust detection and classification of snoring sounds. Simple snores exhibit regular temporal patterns where as waveforms of OSA snores tend to oscillate significantly in neighboring episodes. In [3], we have proposed to use KL-divergence as the distance metric to compare the similarity between two snore signals. In this paper, we further improve the accuracy and robustness of the classification scheme using earth mover's distance (EMD).

The rest of this paper is organized as follows: Section 2 introduces the approach to the extraction of acoustic features and feature selection methods. Section 3 is concerned with experiment settings and results. The experiment was divided

978-1-4244-2794-9/09/$25.00 ©2009 IEEE

into two parts. The first part deals with the feature selection issue for non-speech human sound classification. In the second part, we will investigate the effect of noise on the accuracy of the classification. Section 4 describes our efforts on analyzing and classifying simple snores and OSA snores using EMD. Section 5 summarizes our findings and outlines directions for future work.

II. Feature extraction, selection and validation

A. Feature Extraction

The acoustic features were divided into two groups based on their physical properties. The 7 features in the first feature set (denoted as AF) are fundamental frequency, spectral centroid, spectral spread, spectral flatness, entropy, and two format frequencies. The physical meanings of these acoustic features and the method to calculate them can be found in [3]. The second feature set is composed of 20 MFCCs. MFCC is a set of vectors which be used to model the human auditory perception system. To obtain MFCC features, the audio signal needs to be transformed from frequency (Hz) scale into the Mel scale according to:

$$M(f) = 1125 \ln\left(1 + \frac{f}{700}\right) \tag{1}$$

After applying discrete cosine transform (DCT) on the 20 log energy values which obtained from 20 triangular band-pass filters in frequency domain, we can get 20 dimensional MFCCs. The 20 triangular filters in the frequency domain are illustrated in Fig. 2. The bands of the triangular filters are linearly divided below 1kHz, but become logarithmic at higher frequencies.

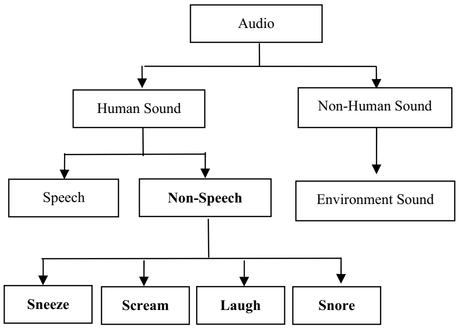


Figure 1. Hierarchical audio classification scheme

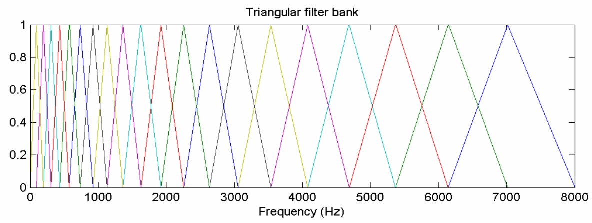


Figure 2. The 20 triangular filters in the frequency domain[7]

B. Feature Selection and Validation

Multivariate adaptive regression spline (MARS) was introduced by Friedman [8]. It is a flexible and multivariate non-parametric regression procedure. The model was made up of the basis functions and approximated to nonlinearity. The MARS model can be expressed as:

$$\hat{f}(x) = a_0 + \sum_{m=1}^{M} a_m \prod_{k=1}^{k_m} \left[s_{km} \cdot \left(x_{v(k,m)} - t_{km} \right) \right]_+ \tag{2}$$

where a_0 、a_m are the coefficients. M is the number of the basis functions. $S_{km} \in \{-1,+1\}$. The $_{v(k,m)}$ label the predictor variables and the t_{km} represent values on the corresponding variables. Using MARS, we can build the optimal model and select the useful variables associated with the model according to the interaction and relationship of the input variables. SVM is then used to examine the features selected by MARS and compare the classification accuracy using different feature sets.

III. Classification of non-speech human sounds

We will describe two experiments conducted in this research and discuss the results in this section. The first is concerned with feature selection. The second experiment investigates the effect of sound quality on the classification task.

A. Experiment I: Feature Selection

We have gathered four kinds of human sounds, include cough (70), scream (70), laugh (70) and snore (70). There are a total of 280 sound files in our audio database. The feature selection experiment is conducted in two phases: the first phase extracts useful features using MARS, and the second phase validates the efficacy of the selected features using SVM.

In the first phase, we randomly divide 280 sound files into two groups (noted G1 、G2) and extracted feature set AF and MFCC from G1 and G2 respectively. After normalizing the feature values to [0,1], we obtained the feature sets G1$_{\mathrm{AF,}}$ G1$_{\mathrm{MFCC}}$, G2$_{\mathrm{AF}}$ and G2$_{\mathrm{MFCC}}$.The sound class and acoustic features were defined as dependent and independent variables respectively. After inputting the variables to MARS, we can obtain two MARS models: G1$_{\mathrm{Mode}}$ and G2$_{\mathrm{Model.}}$ The number of references to the basis functions by all variables was recorded in the MARS model. The variables were deemed 'useful' when the number of reference is greater than zero. Table 1 is an example of the MARS model generated using the feature set AF.

We only retain the variables whose number of the references to the basis functions is greater than zero. In Table I, the referred variables of G1$_{\mathrm{Model}}$ and G2$_{\mathrm{Model}}$ are different. For example, the number of references to the feature SS is 3 in G1$_{\mathrm{Model}}$, yet SS in not used in generating G2$_{\mathrm{Model.}}$ The reason of such case is that the features are unstable and undistinguishable in their own groups. However, we can observe that certain features have been referenced in both models. Therefore, we take the intersection of the two models (G1$_{\mathrm{Model}}$, G2$_{\mathrm{Model}}$) and its complement to obtain the feature subsets S1 and S2, respectively. We expect the subset S1 to be more effective (i.e., have a higher recognition rate) in performing the classification task than S2.

TABLE I. AN INSTANCE OF MARS MODEL BUILT WITH AF

	$\mathbf{G1}_{model}$	$\mathbf{G2}_{model}$
	References (to Basis Functions)	References (to Basis Functions)
SC	3	3
SS	3	0
SF	1	1
En	0	0
F_0	1	1
F1	0	0
F2	1	0

In the second phase of the experiment, SVM is adopted to examine if the feature subset S1 is indeed more important than the feature subset S2. The feature sets AF, MFCC and feature subsets $S1_{AF}$ 、$S2_{AF}$ and $S1_{MFCC}$ were compared based on 2-fold cross validations in G1 and G2. The feature selection experiment is repeated 12 times to eliminate potential bias using insufficient rounds of sampling.

Table II summarizes the result of 12 rounds of feature selection experiment for the feature set AF. Those features which were checked in Table 2 and Table 3 belong to feature subset S1. They are deemed 'useful' for the classification of the human sounds. The average accuracy obtained with feature subset $S1_{AF}$ is 81.95%., which is significantly higher than the one obtained with feature subset $S2_{AF}$ (56.14%). This result is in accordance with the presumption made earlier. The last column in Table II shows the total counts which were selected in the feature subset S1 after 12 rounds of feature selection process. Based on the number of references, we can identify the most contributive features by order. They are: spectral centroid (SC), fundamental frequency (F_0), spectral spread (SS) and spectral flatness (SF), respectively.

Table III shows the result of 12 rounds of feature selection experiment for the feature set MFCC. From Table 3, we can conclude that not all 20 MFCCs are useful in classifying the human sounds. In particular, MFCC coefficients with lower indices are referenced more frequently in building the MARS. In other words, low frequency components of the spectrogram play more important roles in the classification task. The average accuracy with 20 MFCCs is 89.81%, slightly higher than the one obtained using feature subset $S1_{MFCC}$ (84.82% using only 3-6 MFCC coefficients). The result again suggests that the features in feature subset $S1_{MFCC}$ are indeed more contributive.

Comparing the results in Table II and III, we can conclude that the feature set with 20 MFCCs achieve the best recognition rate. However, it is only 6% better than the one using feature set AF (7-dimensional vector). If we reduce the number of MFCC to 6 (i.e., use subset $S1_{MFCC}$), the accuracy becomes comparable. In other words, when the feature dimension is fixed, neither feature set performs better than the other. Whereas the parameters in feature set AF characterize well-defined physical properties, the MFCC method is related to frequency-domain representation of a signal.

B. Experiment II: Noise Sensitivity

It was discovered in the previous experiment that the most contributive features in non-speech human sounds are spectral centroid (SC), fundamental frequency (F_0), spectral spread (SS) and spectral flatness (SF). Denote this set {SC, F_0,SS,SF} as R, our objective in this experiment is to observe the effect of noise on the accuracy of the classification result. We adopt the signal-to-noise (SNR) to measure the level of audio signal to the level of white noise. It is defined as:

$$SNR(dB) = 10\ \log \frac{E_{signal}}{E_{noise}} \tag{3}$$

where $E_{\ signal}$ and $E_{\ noise}$ are the energy of audio signal and noise respectively. For the experiment, we employ SVM as the classifier. All 280 sound files in our audio database are treated as training data. Each file is then alternated to be the test data by adding white Gaussian noise at different level SNR ratios from 100 dB to 10 dB. As a result, there are a total of 2800 sound files in the test set. The SVM classifier is then utilized to perform the classification for the data at each noise level.

We compare with the effect of noise on the accuracy of the classification using feature set AF (7-dimension) and its subset R (4-dimension). The experiment result is shown in Fig. 3. It can be seen that the recognition rates remain unchanged (around 83%) when the SNR is over 50dB. As noise level increases and the SNR approaches 40dB, the accuracy starts to decrease. It is interesting to note that the result using feature set R is somewhat better than that obtained with feature set AF when the noise becomes more prominent. A possible explanation for this phenomenon is that the four contributive features in set R are also more robust to interferences caused by noise.

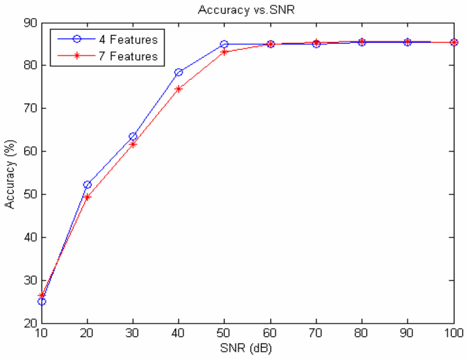


Figure 3. Classification accuracies of feature set AF and R with different SNRs

IV. SNORE SIGNAL ANALYSIS

Using the approach described in the previous section, we can extract segments of snore sounds from audio recorded in all-night sleep studies. It is desirable to further categorize snoring sounds into episodes of different nature to investigate whether the subject suffers from sleep apnea syndrome. For example, Fig. 4 and 5 depict the waveforms and spectrograms of a simple/OSA snore. These two types of snores can be distinguished quite easily by examining the regularities in terms of temporal pattern and energy distribution in the frequency domain.

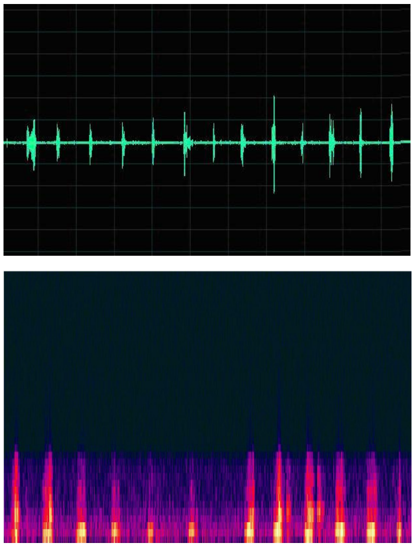


Figure 4. Waveform and spectrogram of a simple snore.

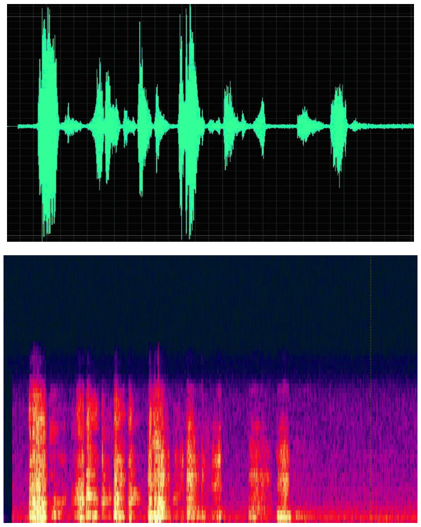


Figure 5. Waveform and spectrogram of an OSA snore.

Two methods for computing the similarity between two segments of snoring signal are evaluated in this paper. Both approaches are based on the results of audio segmentation and spectrogram analysis. In other word, the snoring audio needs to be processed to obtain the individual sound segments, as shown in Fig. 6. Each segment is then transformed into the spectral domain using short-time Fourier transform (refer to Fig 7 (a)). The result further is converted into a probability distribution by choosing a quantization interval q. For example, Fig. 7 (b) is obtained by setting q=500 Hz.

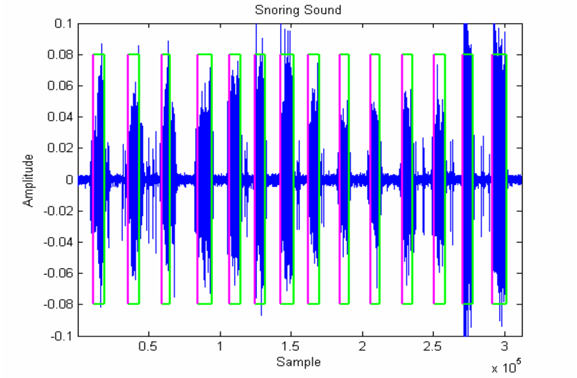
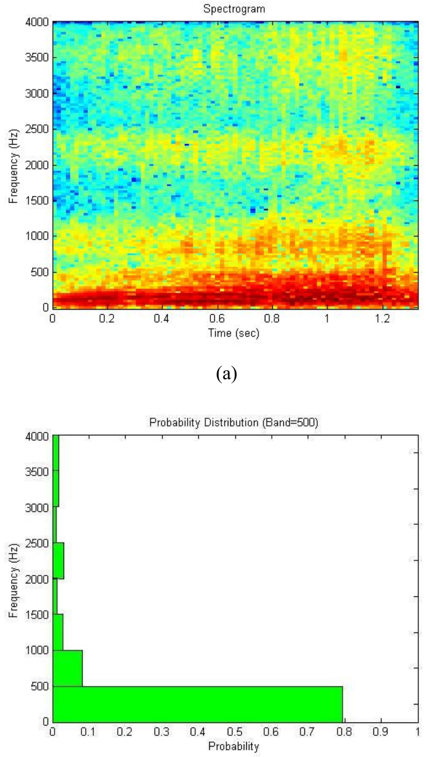


Figure 6. Segmentation of snoring sounds.



(a)

(b)

Figure 7. Converting spectrogram into probability distribution.

Once the input signals are converted into probability distributions, their similarities can be calculated using distance metrics. In [3], we employ a modified KL-divergence to classify simple snores/OSA snores. Even though we have obtained good classification results using this measure, we noticed that the performance varied slightly if the quantization interval q was set differently. Fig. 8 depicts the average KL-divergence between two classes of snores (simple vs. OSA) using quantization interval from 100Hzz to 1000Hz. The largest distance occurs at q=350 Hz.

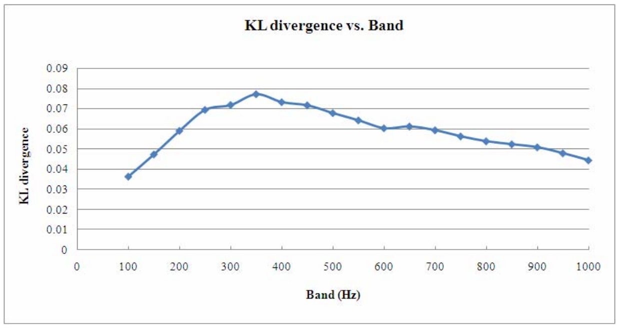


Figure 8. Average KL-divergence using different quantization intervals.

To alleviate the dependency of the distance measure on a specific choice of q, we propose to use another metric: earth mover's distance (EMD) in this study. Suppose $P = \{(p_i, w_{p_i})\}$, $i=1,...,m$, $Q = \{(q_j, w_{q_j})\}$, $j=1,...,n$ are two signatures (or distributions), the EMD between P and Q is defined as [9]:

$$EMD(P,Q) = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} d_{p_i q_j} f_{p_i q_j}}{\sum_{i=1}^{m}\sum_{j=1}^{n} f_{p_i q_j}} \tag{4}$$

where $d_{p_i q_j}$ denotes the distance between feature p_i and q_j and $f_{p_i q_j}$ denotes the flow between p_i and q_{j}., which is obtained by minimizing W:

$$W = \sum_{i=1}^{m}\sum_{j=1}^{n} d_{p_i q_j} f_{p_i q_j} \tag{5}$$

under the following constraints:

$$f_{p_i q_j} \geq 0 \quad 1 \leq i \leq m \quad 1 \leq j \leq n \tag{6}$$

$$\sum_{j=1}^{n} f_{p_i q_j} \leq w_{p_i} \quad 1 \leq i \leq m \tag{7}$$

$$\sum_{i=1}^{m} f_{p_i q_j} \leq w_{q_j} \quad 1 \leq j \leq n \tag{8}$$

$$\sum_{i=1}^{m}\sum_{j=1}^{n} d_{p_i q_j} f_{p_i q_j} = \min\left(\sum_{i=1}^{m} w_{p_i}, \sum_{j=1}^{n} w_{q_j}\right) \tag{9}$$

Fig. 9 presents the average EMD between simple and OSA snores using different settings of q. It can be observed that the distance decreases only slightly with coarser resolution. Consequently, the classification performance remains very much the same with different choices of quantization level q.

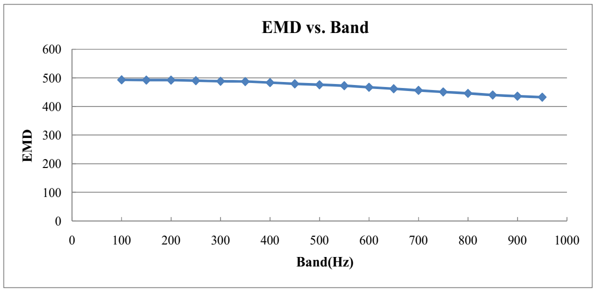


Figure 9. Average EMD using different quantization intervals.

V. CONCLUSIONS

In this paper, we focus on the study of non-speech human sounds. The experimental results show that spectral centroid, fundamental frequency, spectral spread and spectral flatness play important roles in the classification task. Additionally, our simulation indicates that these four features also exhibit better noise insensitivity when the audio signals were acquired under noisy environments.

For comparison, we have also conducted the experiments using MFCC coefficients. The features selected by MARS reveal that low frequency components play significant roles for this particular classification problem.

We have also continued to work on of the classification of snoring sounds. The introduction of earth mover's distance for categorizing simple and OSA snores has increased the robustness and stability of the classification scheme and effectively reduced the need to choose an 'optimal' quantization interval beforehand.

In our future work, we will investigate other acoustic features such as Log Frequency Power Coefficients (LPCC). We will also include more types of human sounds in our audio database, including burp, yawn and cough.

REFERENCES

[1] X. Huang, A. Acero and H. W. Hon, Spoken Language Processing: A Guide to Theory, Algorithm and System Development, Prentice Hall, 2001..

[2] M. Pantic and L.J.M. Rothkrantz, "Toward an affect-sensitive multimodal human-computer interaction," Proceedings of the IEEE, Vol.91, Issue 9, pp.1370 – 1390, 2003.

[3] W. Liao and Y. Su, "Classification of Audio Signals in All-Night Sleep Studies,"18th International Conference on Pattern Recognition, Vol.4, pp.302-305, 2006.

[4] J. Wang, J. Wang, K. He and C. Hsu, "Environmental Sound Classification using Hybrid SVM/KNN Classifier and MPEG-7 Audio

Low-Level Descriptor," International Joint Conference on Neural Networks, 2006.

[5] J. Rong, Y. Chen, M. Chowdhury and G. Li, "Acoustic Features Extraction for Emotion Recognition," 6th IEEE/ACIS International Conference on Computer and Information Science, pp. 419-424, 2007.

[6] R. Jarina and J. Olajec, "Discriminative Feature Selection for Applause Sounds Detection," Image Analysis for Multimedia Interactive Services, Vol., Issue 6-8, pp.13 -16, 2007.

[7] J. S. Jang , Audio Signal Processing and Recognition, http://neural.cs.nthu.edu.tw/jang/books/audioSignalProcessing/ [retrieved March 2009]

[8] J. H. Friedman, "Multivariate Adaptive Regression Splines," Department of Statistics, Stanford University, Technical Report 102 Rev, 1990.

[9] Yossi Rubner; Carlo Tomasi, Leonidas J. Guibas, "A Metric for Distributions with Applications to Image Databases". Proceedings of the International Conference on Computer Vision, pp.59-66, 1998.

TABLE II. FEATURE SELECTION FOR FEATURE SET AF

	AF												
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**Total**
SC	√	√	√	√	√	√	√	√	√	√	√	√	12
SS		√	√		√	√	√	√	√	√	√		9
SF	√	√			√		√	√			√		6
En				√									1
F_0	√	√	√	√	√			√	√	√	√	√	10
F1													0
F2							√	√					2
Feature Set	**SVM Classifier**												**AVG.**
AF (%)	83.86	84.05	84.93	83.68	84.76	82.79	82.80	82.62	82.62	83.15	83.52	83.87	**83.56**
$S1_{AF}$ %)	82.07	83.16	82.43	80.81	83.51	77.40	81.73	82.44	82.44	82.97	82.80	81.55	**81.95**
$S2_{AF}$(%)	60.91	61.99	62.17	63.26	40.50	68.45	46.41	34.05	64.87	64.34	43.38	63.25	**56.14**

TABLE III. FEATURE SELECTION FOR MFCC

	MFCC												
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**Total**
1	√	√	√	√	√	√	√	√	√	√	√	√	12
2	√	√	√	√	√	√	√	√	√	√	√	√	12
3	√	√	√	√	√	√		√	√	√	√	√	11
4				√			√		√		√		4
5		√		√			√		√		√	√	6
6		√		√					√		√		4
7													0
8													0
9	√					√	√						3
10													0
11					√								1
12													0
13	√	√			√	√	√						5
14													0
15													0
16													0
17													0
18													0
19													0
20													0
Feature Set	**SVM Classifier**												**AVG.**
MFCC (%)	89.78	90.68	91.39	89.06	89.6	89.78	89.6	90.85	88.53	89.59	88.89	89.96	**89.81**
$S1_{MFCC}$ (%)	82.78	84.04	83.5	88.71	82.43	82.78	88.52	82.96	87.99	82.25	87.09	84.75	**84.82**