Feature Spaces and Machine Learning Regimes for Audio Classification

A Competitive Study

Muhammad M. Al-Maathidi
School of Computing Science and Engineering University of Salford
Salford, United Kingdom
e-mail: M.M.Abd@edu.salford.ac.uk

Francis F. Li
School of Computing Science and Engineering University of Salford
Salford, United Kingdom
e-mail: F.F.Li@salford.ac.uk

Abstract—The rapid development in computer and internet technology resulted in a large amount of audio content. In order to retrieve and access contents an efficient audio classification system is required. This paper compares the performances of two machine learning schemes namely Artificial Neural Networks and Support Vector Machines, in audio classification or clustering. They are both tested using a set of popular feature spaces to find out the best combinations of classifiers and audio feature spaces that will lead to an efficient multi class audio classification system. Such systems provide important pre-processing stages for automated audio metadata generation and content analysis system.

Keywords: audio classification; multimedia indexing; metadata; content descriptor; information retrieval; machine learning; feature space; machine audition.

I. INTRODUCTION

The fast progress in digital audio technology has resulted in a large number of media files, such as audio recording of meetings, phone calls, voice mails, television and radio broadcasts. An efficient automated content analysis and retrieval system is therefore very important since manual indexing is not practical and time-consuming. In this paper, a comparison between two Decision Making Systems (DMS) the first being the Neural Network (NNet.) based system and the second is the Support Vector Machine (SVM) based system. Each one of these two systems will be tested using a set of audio features. The audio signal will be classified into speech, music and other sound events.

Audio content classification is an active research area in which some publications discussed audio segmentation into a set of pre-defined classes such as news, music, advertisement cartoon, and movies [1, 2]. While others emphasized the content types, for example silence, speech, and music [2, 3]. Moreover, other research focused on sport game segmentation such as football, basketball, tennis, hockey, Ping-Pong, and badminton [4]. Other authors concentrated on acoustic homogenously and self-similarity [5, 6]. In addition, some music research placed an emphasis on different areas like music structure segmentation [7], and instrument type [8]. In speech field research, the focus are on speaker clustering and identification [9, 10]. On the other hand, speech recognition and keyword spotting systems achieve notable results; there are some commercial products such as Dragon NaturallySpeaking software in addition to smartphone operating systems that support natural language speech recognition such as Google’s Android and Apple iOS systems.

The current trend in research in this field has encountered several limitations and weaknesses:

- Most of the classification systems are designed to handle specific pre-defined tasks and they cannot be generalized.
- Performances of different classification systems cannot fairly compared due to the use of different test samples and the lack of standard benchmark.
- Most research focused on a specific audio content sometimes they used an idolized sample set to evaluate the system, which may it hard to compare the results of different techniques.
- In current research, not enough attention was given to the detection of overlapped classes, but this is commonplaces in reality.
- Despite all the in the field, there is no available, free or commercial system that claims the ability of automatic audio file content search.

This paper aims to find some optimal combinations of machine learning techniques and audio features to achieve the best possible classification accuracy. The proposed classification system will try to address some of the above shortcomings utilizing the suitable machine learning techniques combined with the best matching feature spaces and test them with the same set of training/ testing database to give unbiased results.

II. AUDIO CLASSIFICATION SYSTEM

The proposed classification system aims to identify the class of the input audio file content. The system will be trained to detect the occurrence of the following audio classes: speech, music, and the event sound.

Such system may be used as a pre-processing stage for the generation of a set of text descriptors by deploying some existing techniques and facilitate search engines to enable audio searching. In addition, this metadata can be integrated with the MPEG-7 audio description standard to describe audio file content. Such a system is urgently needed to utilize the content-based audio retrieval and audio search engines.
Comparison of two classification methods will be considered, namely the SVM and the NNet. The advantage of the SVM is that it is computationally efficient and also it had been proven as a successful audio classification module in [10-12], the NNet has been used successfully for audio classifications [1, 12].

III. AUDIO FEATURE SPACES

A proper selection of audio features is a major step toward a successful audio classification system; a suitable set of features should be able to preserve the significant audio properties in order to be used by the DMS to distinguish between distinct classes. Also, it is expected to be fairly robust against noise and the presence multi-class cases that might interfere with the classification. The system will utilize a set of features that can be categorized according to their domains.

A. Temporal Domain

Temporal domain is the native domain of an audio signal; it represents the sample amplitude against time. Temporal domain features are fast and easy to extract and they have been successfully used as a audio classification feature [13]. However sometimes it will fail to differentiate between distinct audio classes[2], the following features selected to be used in this domain:

- Zero Crossing (ZC): which has been used to discriminate between voiced/unvoiced audio files[14], and for music genre classification [15, 16].
- Root Mean Square Error (RM): approximates the loudness of the signal [17, 18].

B. Frequency Domain "Spectral Domain"

The frequency domain represent an audio signal by its spectral distribution, the frequency domain features characterize the short-time spectrum of the audio signal windows. The following features selected to be used in this domain:

- Spectral Rolloff Frequency (RL): gives the frequency value that below which 85% of the spectrum magnitude is concentrated [15].
- Brightness(BR): Measures the amount of high-frequency content in an audio signal.
- Roughness(RF): Describes the pleasantness reduction in hearing [19].
- Irregularity: Measures the degree of variation in successive spectrum peaks [20].
- Spectral Flux (SF): is the value of average variation in a signal spectrum between adjacent frames. It measures the local spectral change [19]. SF has been used for Speech/Music Discrimination [12, 14].
- Spectral Centroid (SC): Characterize the signal spectrum. It has been designed to discriminate between different musical instrument timbres [18].
- Audio Spectrum Centroid (ASC): Measures the center of gravity of a log-frequency power spectrum. It indicates whether a power spectrum is dominated by high or low frequencies, also it gives an approximation of the signal perceptual sharpness [18].
- Audio Spectrum Spread (ASS): Describes the spectrum distribution around its centroid. It is designed specifically to help differentiation between noise-like and tonal sounds [18].

C. Cepstral Domain

Cepstrum concept was introduced by Bogert et. al. [21]. It is achieved by taking the Fourier transform of the logarithm of the magnitude of the spectrum. The second Fourier transform can be replaced by the IDFT, DCT, and IDCT. But because the DCT decorrelates the data better than the DFT, it is often preferred [22]. Using the cepstrum is a good way for separating the components of complex signals that are made up of several different but simultaneous elements combined together such as speech.

- Mel Frequency Cepstrum Coefficients (MFCC): MFCC is an excellent feature vector that can efficiently used for both speech and music signals [18]. It has been proven to be beneficial in the field of audio classification [8, 10, 12, 23-25]. MFCC is a perceptually motivated representation; it was developed to approximate the responses of the human auditory system, the Mel is a unit of pitch that has been judged by listeners to be equally spaced. To convert from frequency in hertz to the equivalent Mel frequency the following equation is nominated:

\[
Pitch = 1127.0148 \log(1+f(\text{Hz})/700)
\]  

IV. CLASSIFICATION SYSTEM

The proposed system is a supervised multi class audio classification system that classifies input audio streams into the following three classes: speech, music, and event sound. The performance of two versions of the system will be implemented and examined; the first version will utilize the Feed Forward NNet and the second version will utilize SVM with polynomial kernel function. Different combinations of features will be used for system training and testing. The input to the system will be an audio file and the output will be a percentage of one of the predefined audio classes. The system will contain the following units:

A. Framing Unit

The framing unit will split the input audio file into overlapped frames with size equal to 40ms and a hamming window will be utilized if windowing is required.
B. Frame Feature Extraction Unit

Feature extraction is an essential step to reduce the file frame dimension; it aims to extract the unique features to be used by the classifier to discriminate between the distinct target classes.

No Feature will be extracted from the silent frames; that can be identified by their poor Signal to Noise Ratio (SNR)

C. Classification Unit

The supervised classification system is ready now to be trained and tested. The performance of both NNet and SVM based system will be examined. The input to this unit will be one feature vector for each audio frame and the output will be a number between 1 and 3 that represents one of the three pre-defined classes (music, speech and other event sound)

The unit will contain these tow sub-units.
1) Training Sub-Units

In the training sub-units, the frame features of the manually classified input testing files will be used in DMS training. Different DMS unit will used for each one of the pre-defined classes to be trained, these units will be trained to reach the target “+1” if the frame belong to the same class and “-1” if the frame belong to any other class.

2) Classification Sub-Units

The classification sub-unit aims to classify the input testing frames to one of the pre-defined classes, each frame will be examined by a set of trained DMS units “Equal to the number of classes”, the DMS will produce the output of “+1” if the frame belongs to the same class and “-1” if the frame belongs to any other class. To check the classification accuracy the classification results will be compared with the file manual classification and the percentage of the truly classified frames will represent classification accuracy.

V. Audio Samples Database

A real-life, non-biased, high quality, miscellaneous audio database is an important step to a successful DMS training/testing and performance evaluation. We had created and used this database in our previous work in audio classification[26]. The audio samples have audio CD quality “44.1K sampling rate and 16 bit depth”. They are saved in an uncompressed wave file format that will facilitate faster manipulation and avoid quality degradation, each file in the database contains homogenous content that belongs to one single class; these files have been classified manually into one of the following classes:

1. Speech: this class contains a variety of voices such as males, females, kids and a group of people, also it contain voices of lectures, conversations, shouting, and narration.
2. Music: contains different types, genre, modes and musical instruments.
3. Other: contains some event sounds do not fit in the previous two classes, such as sounds of rain, storms, thunders, screaming, helicopter, crashing, busy road, school yard, and many others samples.

The average sample length is 16 seconds, and the contents of each sample are selected to be acoustically homogenous. Table I shows the number of samples in each class and its average length [24].

<table>
<thead>
<tr>
<th>Sample Class</th>
<th>Samples Count</th>
<th>Average Sample Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speech</td>
<td>200</td>
<td>12.3 Sec</td>
</tr>
<tr>
<td>Music</td>
<td>200</td>
<td>18.4 Sec</td>
</tr>
<tr>
<td>Other</td>
<td>200</td>
<td>17.0 Sec</td>
</tr>
</tbody>
</table>

VI. Classification System Testing and Evaluation

System performance evaluation, discussion about results and comparison will be presented in the following sections.

Each class of the audio database will be split into two groups, one for training and the other for testing, and as a classification technique a one-against-all technique will be adopted. In our previous work [26] the selection of some system parameter were discussed and we will continue with these parameter values in the current work. These parameter values include frame size of 40ms with 50% overlap. Minimum frame energy of 10% will be used to differentiate between silent frames and the non-silent frames. Also a smoothing window of size 15 will be utilized to smooth the classification result and all the results between -0.3 and 0.3 will be ignored in order to discard the highly fluctuating adjacent frames classification result.

The trained classifier will produce “+1” or “-1”. In either case, the classifier may reach the correct decision result therefore it will be referred to as a “true classification,” if it misses the correct decision, result will be referred to as “false classification”. The four possible combinations of these results are positive-true, negative-true, positive-false, and negative-false. The aim is to achieve the highest percentage of truly classified frames for each audio file in order to attain the optimum classification performance.

VII. System Testing Results

In order to get the best possible performance from the DMS, a different combination of features will be tested. The MFCC is used as the main classification feature and it had been supplemented with some other features to improve the classification accuracy. Both of NNet and SVM are tests for each one of these three classes: speech, music, and other event sound.

The results will be presented in two tables “Table II” shows the NNet classification results, and “Table III” shows the SVM classification results. Both tables contain the a column that list the tested classification features; followed three groups of columns; the first list the result of speech-against-all, the second list the result of music-against-all, and the third list the result of other-against-all; each one of these three columns have two sub-columns; one of them list the percentage of the positively-true classified frames, and the other list the percent of the negatively-true classified frames. The positive true represents the percentage of
frames that has been truly classified as a same class, and the negative true represents the percentage of frames that has been truly classified as a different class.

VIII. DISCUSSION

It is clear from the tests results that the MFCC is the best feature that leads to a high accurate classification results, and the rest of the features, “especially the spectrum description features,” can improve the results slightly if two to three or four of them are combined to the MFCC. At the same time, combining more than four features may lead to difficulties in the system training and decrease classification accuracy.

### TABLE II. NEURAL NETWORK CLASSIFICATION RESULTS

<table>
<thead>
<tr>
<th>Test No.</th>
<th>Classification Features</th>
<th>Speech Pos. True%</th>
<th>Speech Neg. True%</th>
<th>Music Pos. True%</th>
<th>Music Neg. True%</th>
<th>Other Pos. True%</th>
<th>Other Neg. True%</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>MFCC</td>
<td>90</td>
<td>88</td>
<td>86</td>
<td>87</td>
<td>66</td>
<td>89</td>
</tr>
<tr>
<td>2</td>
<td>MFCC+ZC</td>
<td>92</td>
<td>91</td>
<td>85</td>
<td>88</td>
<td>76</td>
<td>86</td>
</tr>
<tr>
<td>3</td>
<td>MFCC+RM</td>
<td>89</td>
<td>89</td>
<td>86</td>
<td>93</td>
<td>59</td>
<td>94</td>
</tr>
<tr>
<td>4</td>
<td>MFCC+RL</td>
<td>90</td>
<td>88</td>
<td>84</td>
<td>85</td>
<td>74</td>
<td>89</td>
</tr>
<tr>
<td>5</td>
<td>MFCC+BR</td>
<td>95</td>
<td>91</td>
<td>92</td>
<td>20</td>
<td>78</td>
<td>88</td>
</tr>
<tr>
<td>6</td>
<td>MFCC+RF</td>
<td>91</td>
<td>86</td>
<td>92</td>
<td>85</td>
<td>53</td>
<td>93</td>
</tr>
<tr>
<td>7</td>
<td>MFCC+RE</td>
<td>90</td>
<td>91</td>
<td>94</td>
<td>88</td>
<td>64</td>
<td>90</td>
</tr>
<tr>
<td>8</td>
<td>MFCC+CN</td>
<td>94</td>
<td>91</td>
<td>87</td>
<td>92</td>
<td>77</td>
<td>89</td>
</tr>
<tr>
<td>9</td>
<td>MFCC+SP</td>
<td>95</td>
<td>90</td>
<td>96</td>
<td>94</td>
<td>63</td>
<td>92</td>
</tr>
<tr>
<td>10</td>
<td>MFCC+EN</td>
<td>91</td>
<td>93</td>
<td>91</td>
<td>93</td>
<td>68</td>
<td>94</td>
</tr>
<tr>
<td>11</td>
<td>MFCC+FL</td>
<td>89</td>
<td>92</td>
<td>90</td>
<td>88</td>
<td>76</td>
<td>85</td>
</tr>
<tr>
<td>12</td>
<td>MFCC+RM+EN</td>
<td>93</td>
<td>91</td>
<td>90</td>
<td>93</td>
<td>81</td>
<td>91</td>
</tr>
<tr>
<td>13</td>
<td>MFCC+BR+RE</td>
<td>94</td>
<td>92</td>
<td>89</td>
<td>93</td>
<td>70</td>
<td>93</td>
</tr>
<tr>
<td>14</td>
<td>MFCC+BR+EN</td>
<td>95</td>
<td>91</td>
<td>87</td>
<td>92</td>
<td>77</td>
<td>89</td>
</tr>
<tr>
<td>15</td>
<td>MFCC+CN+SP</td>
<td>96</td>
<td>92</td>
<td>93</td>
<td>94</td>
<td>78</td>
<td>87</td>
</tr>
<tr>
<td>16</td>
<td>MFCC+EN+FL</td>
<td>94</td>
<td>92</td>
<td>91</td>
<td>92</td>
<td>78</td>
<td>89</td>
</tr>
<tr>
<td>17</td>
<td>MFCC+RL+SP</td>
<td>93</td>
<td>94</td>
<td>97</td>
<td>93</td>
<td>79</td>
<td>89</td>
</tr>
<tr>
<td>18</td>
<td>MFCC+RL+SP+RF+FL</td>
<td>95</td>
<td>93</td>
<td>89</td>
<td>93</td>
<td>77</td>
<td>89</td>
</tr>
<tr>
<td>19</td>
<td>MFCC+RM+BR+SP+EN</td>
<td>94</td>
<td>94</td>
<td>92</td>
<td>94</td>
<td>76</td>
<td>85</td>
</tr>
<tr>
<td>20</td>
<td>MFCC+ZC+BR+SP+FL</td>
<td>96</td>
<td>93</td>
<td>93</td>
<td>94</td>
<td>81</td>
<td>91</td>
</tr>
</tbody>
</table>

### TABLE III. SUPPORT VECTOR MACHINE CLASSIFICATION RESULTS

<table>
<thead>
<tr>
<th>Test No.</th>
<th>Classification Features</th>
<th>Speech Pos. True%</th>
<th>Speech Neg. True%</th>
<th>Music Pos. True%</th>
<th>Music Neg. True%</th>
<th>Other Pos. True%</th>
<th>Other Neg. True%</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>MFCC</td>
<td>90</td>
<td>80</td>
<td>89</td>
<td>85</td>
<td>48</td>
<td>94</td>
</tr>
<tr>
<td>2</td>
<td>MFCC+RM</td>
<td>90</td>
<td>82</td>
<td>89</td>
<td>86</td>
<td>48</td>
<td>94</td>
</tr>
<tr>
<td>3</td>
<td>MFCC+BR</td>
<td>93</td>
<td>86</td>
<td>87</td>
<td>82</td>
<td>71</td>
<td>93</td>
</tr>
<tr>
<td>4</td>
<td>MFCC+RE</td>
<td>91</td>
<td>85</td>
<td>90</td>
<td>85</td>
<td>71</td>
<td>93</td>
</tr>
<tr>
<td>5</td>
<td>MFCC+EN</td>
<td>93</td>
<td>86</td>
<td>87</td>
<td>82</td>
<td>71</td>
<td>93</td>
</tr>
<tr>
<td>6</td>
<td>MFCC+RM+BR</td>
<td>94</td>
<td>86</td>
<td>87</td>
<td>83</td>
<td>71</td>
<td>93</td>
</tr>
<tr>
<td>7</td>
<td>MFCC+RM+EN</td>
<td>94</td>
<td>86</td>
<td>87</td>
<td>82</td>
<td>71</td>
<td>93</td>
</tr>
<tr>
<td>8</td>
<td>MFCC+BR+RE</td>
<td>94</td>
<td>86</td>
<td>88</td>
<td>83</td>
<td>72</td>
<td>90</td>
</tr>
<tr>
<td>9</td>
<td>MFCC+BR+EN</td>
<td>94</td>
<td>86</td>
<td>87</td>
<td>84</td>
<td>72</td>
<td>91</td>
</tr>
<tr>
<td>10</td>
<td>MFCC+RM+BR+EN</td>
<td>94</td>
<td>87</td>
<td>87</td>
<td>84</td>
<td>71</td>
<td>91</td>
</tr>
<tr>
<td>11</td>
<td>MFCC+RM+BR+EN+FL</td>
<td>94</td>
<td>87</td>
<td>88</td>
<td>84</td>
<td>72</td>
<td>91</td>
</tr>
</tbody>
</table>

### A. NNet Results Discussion

The average of all NNet tests in true classification achieved the accuracy of 87.2%, the average “for both positive and negative classification” in speech-against-all test is 92%; in music-against-all test is 88.6%, and in others-against-all test is 81.1%. The best achieved true classification accuracy “for both positive and negative classification” in speech-against-all was 94.5% achieved in test 20; in music-against-all it was 95% achieved in test 9; in others-against-all it was 86% achieved in test 20. The NNet training went smooth, in the test other-against-all, the positively true classification result was relatively low with average of 72.6%, but the negatively true on the same test was 94%; even though the result look problematic but it is actually make sense because the NNet was able to detect the feature pattern in of both of both speech and music, but for the class others was not that easy to find such recognize pattern because of the variety of class content, even though it achieved the result of 81% for positively-true, and the result of 91% of negatively true in tests number 12 and 20.

### B. SVM Results Discussion

The average of all SVM tests in true classification achieved the accuracy of 84.7%, the average “for both positive and negative classification” in speech-against-all test is 89%; in music-against-all test is 85.6%, and in others-against-all test is 79.5%. The best achieved true classification accuracy “for both positive and negative classification” in speech-against-all was 90.5% achieved in tests 10 and 11, in music-against-all it was 85.6% achieved in tests 2 and 4; in others-against-all it was 82% achieved in test 7. In the SVM the polynomial kernel function was used the these test, but it was not able to converge to a valid target in the tests that combined MFCC and any one of ZC, RL, CN, SP and FL; that’s means that the SVM was not able to linearly separate the training class feature from the other classes features, the radial based kernel was tested for these vectors but the result was not that good, was 74.7%. Even if the result was better than this the radial based kernel is not a suitable kernel for our classification because if the SVM wasn’t able to separate the features laniary for sure it will not be able to do an sufficient radial base function especially with the a relatively high dimension of feature vector.

IX. CONCLUSIONS AND FUTURE WORK

From the above empirical comparisons, we found a NNet based DMS for a multi class audio classification is preferable. Regarding the features we advocate the use the MFCC as the main feature spaces with four spectrum descriptors auxiliary feature spaces, this has been found empirically to give excellent results. For the other-against-all classification we recommend to use the negative classification results only in order to achieve better classification accuracy.

For the future work, the focus will be on the following topics:

1. Improving the classification results by introducing onset/offset class boundaries identification.
2. Each one of the three classes can be further classified into relevant sub classes by utilizing more dedicated algorithms that have been listed in the literature.

3. Combine the results of the three classification modules: speech, music, and others, in order to attain a mixed class classification system.

4. The utilization of the learning vector quantization algorithm will enable the DMS system to automatically select the best set of classification features.

5. Integrate system output with the MPEG-7 multimedia content description standard to enable a fast and easy search for the analyzed audio files content.

REFERENCES
