Presented at the DLSU Research Congress 2015
De La Salle University, Manila, Philippines
March 2-4, 2015

Ambient Sound Source Classification in the TALA Emphatic Space using Machine Learning

Jeffrey Dy[1], Kevin Adrian Go[1], Mark Benson Ong[1], Rainielle Caezar Yosa[1,*] and Jocelynn Cu[1]
[1] *Center for Emphatic Human-Computer Interactions (CEHCI)*
College of Computer Studies – De La Salle University
Manila, Philippines
**Corresponding Author: rainielle_yosa@dlsu.edu.ph*

Abstract: People working in laboratories, such as the TALA emphatic space are expected to be productive most if not all the time. However, factors such as being exposed to different sound sources could result to distractions that could affect productivity. In order to understand which types of sounds affect the productivity of a person, a system that could classify sound sources should be developed first. Understanding the composition of sound helps determine the state of the environment from which it is heard. In literature, the task of automatic sound classifications has been simplified into a binary classification problem. This is caused by the challenges posed by classifying sound into more than two categories. However, applications in ambient sound control, context-aware computing, among others, require that such multiple sounds classifier be developed. This work is focused on classifying a sound source into chair bump, chair drag, door, and music inside a laboratory. The features extracted from the four sound sources are modelled using decision tree, sequential minimal optimization, and multilayer perceptron. Results showed that for two-way classification, an accuracy of 98.9% is achieved for MLP. Also, 83.9% and 75.37% accuracy for MLP are achevied for three-way and four-way classification ,respectively

Key Words: sound source; classification; machine learning

1. INTRODUCTION

Increased personal control and comfort needs of employees sparked the concern among various organizations. These lead them to come up with an environment and office design that would fulfil their members' needs and even help boost productivity in the workplace. One such environment is the TALA (Cu et al., 2007) empathic space in the Center for Human-Computer Interactions (CEHCI) in De La Salle University.

TALA provides empathic support to its occupants by determining their emotions and adjusting environmental factors such as room temperature and brightness to best meet an occupant's demands. The empathic space is also capable of performing emotion-based interactions through analysing the facial expression, speech, and movement of its occupants. Furthermore, it is capable of playing music that best suits an occupant's

mood. This ability can be utilized when there is sufficient knowledge regarding the types of sounds that can help the occupants' productivity.

Creating a system that has the same mechanism as the human ear has been proven to be a hard task. Finding features or mathematical models that describe the variability of classes is not directly evident. This made it uncommon to encounter works which deal with the classification of sound as speech, noise, or music. Most related works as referred in Table 2 commonly explore on two-class problems such as the classification of speech from music (Fakotakis and Ntalampiras, 2008; Anderson, 2004; Cai, 2013). These works made use of various techniques ranging from machine learning (Anderson, 2004; Cai, 2013) to digital signal processing (Yen, 2011; Fakotakis and Ntalampiras, 2008; Balabko, 1999) which usually yields 62-95% accuracy.

One common problem shared by these works is that the two sound classifications exhibit similar characteristics which often lead to the classification of speech as music or music as speech. Such characteristics exhibited by both sound classifications are strong rhythmic beats (Bugatti et al., 2002) and timbre. In spite of this being the case, there exists a few works (Lu et al., 2012; Montacie and Caraty, 2005; Papaj, 2008) that were able to classify more than two sound classes. However, these multi-class problems are approached using binary classification, a task wherein elements are grouped into two classes.

2. METHODOLOGY

This section serves as the foundation of the research and a guide in choosing the most suitable and distinguishable model that can solve the problem of classifying a sound source into music, noise, or speech.

2.1 Data Gathering

For the entirety of the work, all audio data were recorded in the CEHCI laboratory. The recordings involved a single microphone strategically placed inside the laboratory as shown in Figure 1 such that the amplitude of the sound received by the microphone would not be biased in one part of the room. A total of five hours of audio information was recorded and saved as WAV files with the use of Audacity, an audio editor and recorder.

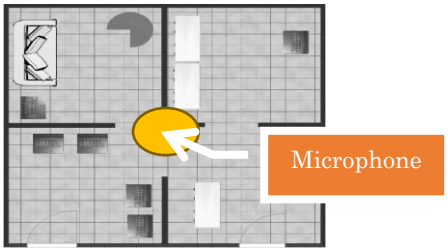

Fig. 1. CEHCI setup

For the training data, actual sound produced by the occupants of the CEHCI was recorded. Any recordings, where an instance involved more than one classification of sound occurring concurrently, were removed from the training data. One example would be when a person spoke during the recording of the raw data. The recordings were then labeled into categories namely: *music*, *bell*, *chatter*, *footsteps*, *telephone ring*, *chair bump*, *chair drag*, and *door*.

All recordings labeled as chair drag, door, chair bump, or music were chosen for the preliminary work. The recordings were sampled at 44.1 kHz and were later segmented into 2048 samples (46 ms) using a Hamming window function which is usually performed by related work (Balabko, 1999).

2.2 Feature Selection

An initial set of 62 features were used for modelling.

JAudio, an audio feature extraction tool, was used to produce the CSV files containing the feature value of the audio recordings. These files were modified in order to create a data set which does not include features with missing values so that the computational speed would be faster. The number of instances for each classification was balanced and the instances where some features cannot be derived were removed. This produced a dataset with a size described in Table 3.

2.3 Modelling

The extracted feature values were fed into WEKA (Waikato Environment for Knowledge and Analysis), an open source classification and data mining software, for the construction of different models. The modelling algorithms used includes J48, Sequential Minimal Optimization (SMO), and Multi-Layer Perceptron (MLP) which were also applied by similar works (Yen, 2011; Lu et al., 2012).

The models were then validated using a number of performance metrics namely: Cohen's kappa coefficient, precision, recall, and accuracy.

3. RESULTS AND DISCUSSION

The initial experiment made use of three different machine learning techniques to establish the optimal modelling technique among the three: decision tree (J48), sequential minimal optimization (SMO), and multilayer perceptron (MLP).

An experiment was conducted using two different sets of features namely: full-featured and feature-selected as shown in Tables 3-6. This was performed to further understand the capabilities of the models.

Features commonly used by different literature (Lu et al., 2012; Fakotakis and Ntalampiras, 2008) and are available in JAudio were included in an initial set. To decrease the dimensionality and computational complexity of the system, the features were reduced to make up the feature-selected set.

The features are then used to extract attributes from these sound sources namely: chair bump, door, chair drag, and music. These were some of the most commonly occurring sound sources in the CEHCI.

For the 4-class classification, as shown in Table 7, a maximum classification rate of 99.9% was achieved in the J48 model. It was observed that the decision rules of J48 only made use of Area Method of Moments (AMM). In order to better understand the models, AMM was removed from the feature sets. This yielded results with an average classification rate of 75.3%. It can also be seen that the full feature set without AMM performed significantly better than feature selected with an average increase of 4.82%.

Since problems may arise from doing multiclass classification, two and three class classification problems were also considered. This will help in understanding the difficulties that may arise once the number of classifications is increased.

For the 2-class classification, it shows that "Bump-Others" performed the worst among the rest. It achieved a classification rate of an average of 69.45% compared to 85.44% of the rest. It can be observed in the model that a bump is usually misclassified as a door. The aperiodic nature of both the door and the bump may have caused this confusion.

Lastly, for the 3-class classification, it shows that "Door+Music+Others" performed the best among the others. It successfully achieved a classification rate of 85.422%. Since the features used were evident in music, it can produce a high accuracy rate. Also, the kappa are high which shows that classification was consistent.

4. CONCLUSIONS

Among the machine learning techniques, MLP achieved the highest classification rate among all multi-class problems. The models also had a difficulty in classifying "chair bump" which led to several models getting low accuracy.

Using the tested models, other sound classifications such as speech and other types of music and noise can be applied. In these cases, machine learning would be a highly suitable approach since a model is easier to customize compared to statistic digital processing. This contributes to the scalability of the system where users may have different profiles and preferences with regards to ambient sound. This eliminates the threshold requirements that usually come along with sound source classifiers. A machine learning approach would also open avenues where multiple dimensions of attributes and large number of features are manageable.

5. FUTURE WORKS

To be able to further describe and understand the environment, it is needed to further select properly the feature set that will be used. Overlapping instances is also something planned to be further explored and experimented. Lastly, being able to classify the sound sources into the three general classes- music, speech, and noise using the feature set prepared will be tested.

6. REFERENCES

Bugatti, A. & et al. (2002). Audio classification in speech and music: A comparison between a statistical and a neural approach.

Montacie, C. & Caraty, M.J. (2005). A silence /noise/ music/speech splitting algorithm.

Cu, J. & et al. (2007). The TALA empathic space: Integrating affect and activity recognition into a smart space.

Yen, J. (2011). Wavelet for acoustics.

Lu, L. & et al. (2012). Content analysis for audio classification and segmentation.

Papaj, M. (2008). Silence/Noise detection for speech and music signals.

Fakotakis, N. & Ntalampiras, S. (2008). Speech/music discrimination based on discrete wavelet transform.

Balabko, P. (1999). Speech and music discrimination based on signal modulation spectrum.

Khoa, P. (2012). Noise robust voice activity detection.

Anderson, T. (2004). Audio classification and content description.

Cai, W. (2013). Analysis of acoustic feature extraction algorithms in noisy environments.

TABLE 1. FEATURES

Feature	Description
Spectral Centroid	The center of mass of the power spectrum.
Spectral Rolloff Point	The fraction of bins in the power spectrum at which 85% of the power is at lower frequencies. This is a measure of the right-skwedness of the power spectrum.
Spectral Flux	A measure of the amount of spectral change in a signal. Found by calculating the change in the magnitude spectrum from frame to frame.
Compactness	A measure of the noisiness of a signal. Found by comparing the components of a window's magnitude spectrum with the magnitude spectrum of its neighbouring windows.
Spectral Variability	The standard deviation of the magnitude spectrum. This is a measure of the variance of a signal's magnitude spectrum.
Root Mean Square	A measure of the power of a signal.
Zero Crossings	The number of times the waveform changed sign. An indication of frequency as well as noisiness.
Strongest Frequency Via Zero Crossings	The strongest frequency component of a signal, in Hz, found via the number of zero-crossings.
Strongest Frequency Via Spectral Centroid	The strongest frequency component of a signal, in Hz, found via the spectral centroid.
Strongest Frequency Via FFT Maximum	The strongest frequency component of a signal, in Hz, found via finding the FFT bin with the highest power.
MFCC	MFCC calculations based upon Orange Cow code
LPC	Linear Prediction Coeffecients calculated using autocorrelation and Levinson-Durbin recursion.
Method of Moments	Statistical Method of Moments of the Magnitude Spectrum.

Table 2. References

Reference	Description	Approach	Results
Yen, J. (2011).	Classified sound into speech and music	Fourier transform, time-frequency analysis, continuous wavelet transform	78% accuracy
Lu, L. & et al., (2012).	Classified sound into speech/music/silence/ environmental sound	K-nearest-neighbor, spectral pairs-vector quantization	95% music and 88% environmental sound
Fakotakis, N. & Ntalampiras, S. (2008).	Classified sound into speech and music	Wavelet transform, Gaussian mixture model, ten-fold cross validation	Success rate of 91.8%
Balabko, P. (1999).	Classified sound into speech and music	Short-time Fouriere transform, Hamming window, Mel-scal transformation, Gaussian estimation	Error-rate 62% to 98%
Khoa, P. (2012).	Classified speech from silence and noise	Feature extraction	Use spectral local harmonicity feature
Anderson, T. (2004).	Classified sound into speech and music	k-NN, Gaussian mixture model, HMM	Recognition rate of 98%
Cai, W. (2013).	Classified sound into speech and music	Hamming window, pitch estimation algorithm, speaking rate estimation algorithm	BaNa for highest pitch estimation accuracy, sub-band correlation for speaking rate estimation algorithm

TABLE 3. INSTANCE COUNT PER CLASSIFICATION

Class	Number of Instances										
	2-class				*3-class*						*4-class*
Chair Bump	1500	500	500	500	1500	1500	1500	750	750	750	1500
Chair Drag	500	1500	500	500	1500	750	750	1500	1500	750	1500
Music	500	500	1500	500	750	1500	750	1500	750	1500	1500
Door	500	500	500	1500	750	750	1500	750	1500	1500	1500

TABLE 4. FULL FEATURE 2-CLASS MODEL PERFORMANCE METRIC

Machine Learning	*Chair Bump vs Others*		*Chair Drag vs Others*		*Door vs Others*		*Music vs Others*	
	Accuracy (%)	*Kappa*	*Accuracy (%)*	*Kappa*	*Accuracy (%)*	*Kappa*	*Accuracy (%)*	*Kappa*
J48	70.0667	0.4013	80.2	0.604	78.4667	0.5693	96.7333	0.9347
SMO	71.1	0.422	80.7333	0.6147	79.1333	0.5827	98.1333	0.9627
MLP	73.2667	0.4653	84.4	0.688	81.1333	0.6227	98.9	0.978

TABLE 5. FEATURE SELECTED 2-CLASS MODEL PERFORMANCE METRIC

Machine Learning	*Chair Bump vs Others*		*Chair Drag vs Others*		*Door vs Others*		*Music vs Others*	
	Accuracy (%)	*Kappa*	*Accuracy (%)*	*Kappa*	*Accuracy (%)*	*Kappa*	*Accuracy (%)*	*Kappa*
J48	69.533	0.3907	77.8	0.556	82.1	0.642	94.5667	0.8913
SMO	62.733	0.2547	74.6	0.492	79.6333	0.5927	91.0333	0.8207
MLP	70.0333	0.4007	80.4667	0.6093	83.2667	0.6653	96.6667	0.9333

TABLE 6. FULL FEATURE 3-CLASS MODEL PERFORMANCE METRIC

Machine Learning	*Chair Bump vs Door + Others*		*Chair Bump + Chair Drag + Others*		*Chair Bump + Music + Others*		*Chair Drag + Door + Others*		*Door + Music + Others*		*Chair Drag + Music + Others*	
	Accuracy (%)	*Kappa*	*Accuracy (%)*	*Kappa*	*Accuracy (%)*	*Kappa*	*Accuracy (%)*	*Kappa*	*Accuracy (%)*	*Kappa*	*Accuracy (%)*	*Kappa*
J48	65.8221	0.4873	66.6444	0.4997	73.6222	0.6043	68.5333	0.528	82.2444	0.7337	80.5778	0.7087
SMO	67.5111	0.5127	66.7333	0.501	74.0222	0.6103	73.2222	0.5983	84.1333	0.762	82.8444	0.7427
MLP	71.7111	0.5757	71.5111	0.5727	76.1556	0.6423	76.5111	0.6477	85.4222	0.7813	84.8222	0.7723

TABLE 7. FEATURE SELECTED 3-CLASS MODEL PERFORMANCE METRIC

Machine Learning	*Chair Bump vs Door + Others*		*Chair Bump + Chair Drag + Others*		*Chair Bump + Music + Others*		*Chair Drag + Door + Others*		*Door + Music + Others*		*Chair Drag + Music + Others*	
	Accuracy (%)	*Kappa*	*Accuracy (%)*	*Kappa*	*Accuracy (%)*	*Kappa*	*Accuracy (%)*	*Kappa*	*Accuracy (%)*	*Kappa*	*Accuracy (%)*	*Kappa*
J48	65.667	0.485	63.8667	0.458	70.0222	0.5503	69.2667	0.539	81.1778	0.7177	77.8889	0.6683
SMO	59.5333	0.393	59.3778	0.3907	65.3111	0.4797	69.4667	0.542	77.578	0.6637	75.8444	0.6377
MLP	67.3556	0.5103	66.3111	0.4947	71.4	0.571	74.3111	0.6147	83.9333	0.759	81.6667	0.725

TABLE 8. 4-CLASS MODEL PERFORMANCE METRIC

Machine Learning	*Full Feature Set*		*Recommended Feature Set*		*Full Feature Set Without AMM*		*Feature Selected*	
	Accuracy	*Kappa*	*Accuracy*	*Kappa*	*Accuracy*	*Kappa*	*Accuracy*	*Kappa*
J48	99.9333	0.9991	99.9667	0.9996	69.15	0.5887	67.35	0.5647
SMO	98.9167	0.9856	97.7	0.9693	71.3667	0.6182	63.6833	0.5158
MLP	99.5833	0.9944	99.2667	0.9902	75.3667	0.6716	70.3833	0.6051