South Dakota State University

Open PRAIRIE: Open Public Research Access Institutional Repository and Information Exchange

Electronic Theses and Dissertations

2017

Adaptive Audio Classification Framework for in-Vehicle Environment with Dynamic Noise Characteristics

Haitham Alsaadan

South Dakota State University

Recommended Citation

Alsaadan, Haitham, "Adaptive Audio Classification Framework for in-Vehicle Environment with Dynamic Noise Characteristics" (2017). *Electronic Theses and Dissertations*. 1739.
https://openprairie.sdstate.edu/etd/1739

ADAPTIVE AUDIO CLASSIFICATION FRAMEWORK FOR IN-VEHICLE ENVIRONMENT WITH DYNAMIC NOISE CHARACTERISTICS

BY

HAITHAM ALSAADAN

A thesis submitted in partial fulfillment of the requirements for the

Master of Science

Major in Computer Science

South Dakota State University

2017

Adaptive Audio Classification Framework for in-Vehicle Environment with Dynamic Noise Characteristics

This thesis is approved as a credible and independent investigation by a candidate for the Master of Science degree and is acceptable for meeting the thesis requirements for this degree. Acceptance of this thesis does not imply that the conclusions reached by the candidate are necessarily the conclusions of the major department.

Myounggyu Won, Ph.D.
Thesis Advisor Date

Steven Hietpas, Ph.D.
Head, Department of EECS Date

Kinchel C.Doerner, Ph.D.
Dean, Graduate School Date

ACKNOWLEDGEMENTS

I cannot express enough thanks to my advisor Dr. Myounggyu Won of Department of Computer Science at South Dakota State University. Dr. Won always encourage me and give me advices to increase quality of my thesis. He consistently allowed my research to be my own work while continuously steering me in the right direction.

I would also like to thank the committee members for my thesis especially Dr. Shin and Dr. Liu who gave me valuable feedback during my preliminary presentation. Their constructive comments substantially helped me in improving the quality of my thesis. I am also grateful for Dr. Woodard for taking his time to review my thesis and provide valuable feedback during my final oral exam. Also, I would thank my parents. They always support me to complete learning to get Master degree.

Finally, I would like to thank the Department of Electrical Engineering and Computer Science and all the graduate faculty of the department for their continuous support and encouragement I have received during the course of my degree program. The courses I took from thm helped me greatly to learn about fundamental materials that made my research activities possible.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

ITS	Intelligent Transportation Systems
KNN	k-Nearest Neighbor
SVM	Support Vector Machine
RMS	Root Mean Square
GMM	Gaussian Mixture Model
ANN	Artificial Neural Network
DNN	Deep Neural Networks
ASR	Automatic Speech Recognition
FPGA	Field Programmable Gate Array
CDF	Cumulative Distribution Function

ABSTRACT

ADAPTIVE AUDIO CLASSIFICATION FRAMEWORK FOR IN-VEHICLE ENVIRONMENT WITH DYNAMIC NOISE CHARACTERISTICS

HAITHAM ALSAADAN

2017

With ever-increasing number of car-mounted electric devices that are accessed, managed, and controlled with smartphones, car apps are becoming an important part of the automotive industry. Audio classification is one of the key components of car apps as a front-end technology to enable human-app interactions. Existing approaches for audio classification, however, fall short as the unique and time-varying audio characteristics of car environments are not appropriately taken into account. Leveraging recent advances in mobile sensing technology that allows for effective and accurate driving environment detection, in this thesis, we develop an audio classification framework for mobile apps that categorizes an audio stream into music, speech, speech+music, and noise, adaptably depending on different driving environments. A case study is performed with four different driving environments, *i.e.,* highway, local road, crowded city, and stopped vehicle. More than 420 minutes of audio data are collected including various genres of music, speech, speech+music, and noise from the driving environments. The results demonstrate that the proposed approach improves the average classification accuracy by up to 166%, and 64% for speech, and speech+music, respectively, compared with a non-adaptive approach in

our experimental settings.

Chapter 1

Introduction

1.1 Motivation

Audio classification has been primarily used for automated multimedia content analysis [40] and is rapidly expanding the range of applications into diverse areas. For example, it is a key front-end technology for speech recognition algorithms that are fundamental technological ingredients to help advance the smart city initiative. Also, it is increasingly applied to in-vehicle intelligent transportation system (ITS) applications (apps) for safer and more efficient control of ever increasing car-mounted electronic devices. An example is the automated vehicle audio volume control app that automatically lowers the volume when on-going human speech is detected, and adjusts the volume appropriately depending on the level of vehicle cabin noise (dynamic noise levels as illustrated in Figure 1.1) when music is detected.

1.2 Limitations of Prior Art

Numerous approaches have been proposed for audio classification, especially concentrating on classifying audio data into music and speech. Lu *et al.* presented a k-nearest neighbor (KNN)-based classification algorithm along with some new features like the noise frame ratio and band periodicity [21]. Chen *et al.* utilized support vector machine (SVM) to extract audio data from movies [5]. They selected four major features; silence ratio, and variance of zero-crossing rate from the time domain, and

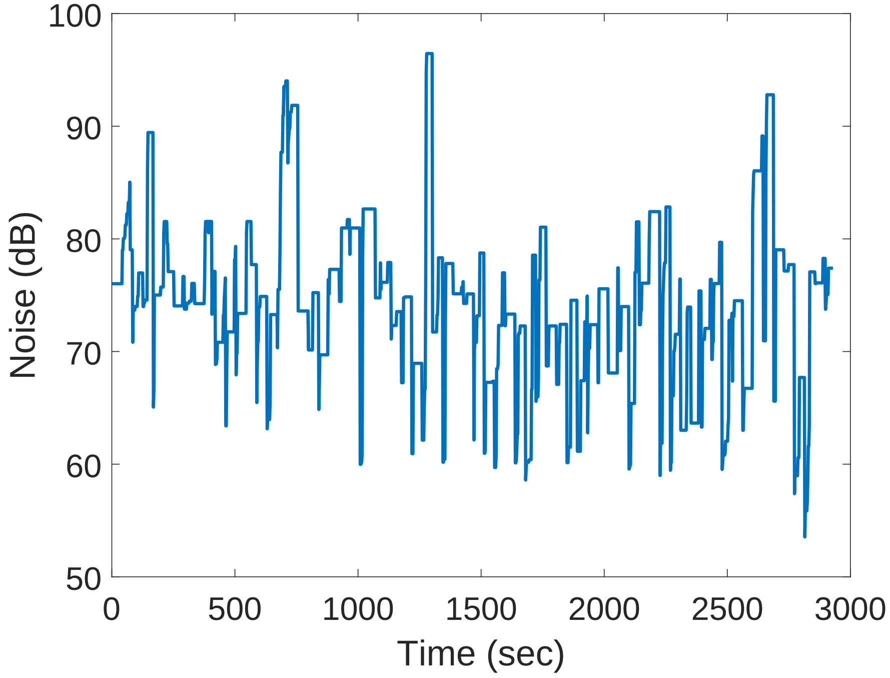

Figure 1.1: Time-varying in-vehicle noise levels in highway.

sub-band energy, and harmonic ratio from the frequency domain. Dogan *et al.* adopted a decision-tree for classification based on MPEG-7 features including audio spectrum centroid, and audio power [6]. Xie *et al.* introduced two new pitch-density-based features namely the average pitch-density and relative tonal power density for more effective classification [40][9]. However, these approaches do not take into account the unique and time-varying audio characteristics of in-vehicle environments, and their effect on the classification accuracy.

1.3 Proposed Approach

In this thesis, we design, implement, and evaluate an adaptive audio classification framework for smartphone apps that classifies an input audio stream into four different audio types, *i.e.,* music, speech, speech+music, and noise. The key idea of the proposed framework is that the classification accuracy is substantially improved by generating individual classification models for varying driving environments and adaptively applying the models based on real-time identification of driving environments. This adaptive approach is possible with recent advances in cloud computing and mobile sensing technologies that allow previously unattainable precision in driving environment detection.

1.4 Key Contributions

More specifically, the proposed framework consists of a feature selection module that identifies an optimal feature set, and a classification algorithm based on the support vector machine (SVM) that adapts to varying driving environments. The training phase of the proposed system allows users to build classification models for a wide range of driving environments, *e.g.,* different weather conditions, vehicle models, and road conditions.

The modularized design of the proposed framework increases the extensibility. The driving environment detection module allows for integration of various mobile sensing solutions to efficiently identify the current driving environment. In addition, the feature extraction module of the framework allows for extension with other useful features to improve the classification accuracy. The proposed classification algorithm can be easily integrated with extended modules for better performance.

In this thesis, we perform a case study with four different driving environments;

highway, local road, city, and stopped vehicle. The results indicate that the accuracy is improved by up to 166%, and 64% for classification of speech, and speech+music, respectively, compared with a non-adaptive solution (the accuracy for music was high in both adaptive and non-adaptive approaches). The results also show that the proposed framework out-performs an existing classification algorithm [40].

The contributions of this work are summarized as follows.

- We collected a corpus of more than 420 mins of real-world audio data (*i.e.,* speech, music, speech+music, and noise) collected with a smartphone microphone from various driving environments (*i.e.,* highway, local road, city, and stopped vehicle).

- We identified that non-adaptive audio classification approaches face significant performance degradation in a in-vehicle environment if we do not account for varying driving environments.

- We developed an adaptive multi-class audio classification framework for smartphone apps that adaptively applies classification models depending on driving environments to achieve high classification accuracy.

- We performed experiments with real audio data collected with a smartphone to validate the effectiveness of the proposed approach.

1.5 Thesis Organization

This thesis is organized as follows. In Chapter 2, we review the literature, followed by the audio characteristics of different driving environments and identification of the performance degradation of existing approaches in Chapter 3. We then describe the design of the proposed system in Chapter 4. Experimental results are presented in Chapter 5; The conclusions are included in Chapter 6.

Chapter 2

Literature Review

Audio classification has long been studied. Numerous approaches have been proposed that classify audio input into various types. A large body of the research focused on audio classification into speech and music [9][19][30][33][7]. There were a number of works that classify audio streams into mixed types, *e.g.,* speech, speech+music, music, and environmental sound [5][6][40][43]. Recently, researchers examined the classification of audio data beyond just music and speech. Examples include music genre classification [27][26][39], and classification of acoustic scenes [34][24][14][31]. In this chapter, we briefly review two key aspects of audio classification, *i.e.,* feature extraction, and classification, and specifically focus on research efforts on audio classification for car environments.

2.1 Feature Extraction

In audio classification, features used can be largely classified into four groups: energy features, spectral features, spectral envelope features, and pitch features [40]. Energy features include the root mean square (RMS) [28], low energy ratio [36], 4Hz modulation energy [32], subband energy distribution [23], and noise frame ratio [22]. The spectral flux [20][32], zero crossing rate [35], and spectral kurtosis [20] were used as spectral features. For spectral envelope features, Mel frequency cepstral coefficients [18], linear predictive cepstral coefficients [20], power spectrum deviation [12], and linear spectral pairs [18] have been widely adopted. Pitch features included spectral peak duration, and pitch tunning [45].

2.2 Classification

Classifiers are another important component of audio classification. Various classifiers have been adopted. Most widely used ones were the K-nearest neighbor (KNN) [22], Gaussian mixture model (GMM) [32], artificial neural network (ANN) [3], and support vector machine (SVM) [23]. In particular, it has been reported that SVM outperforms most other classifiers [40]. Recently, deep neural networks (DNN) emerged as a new model for audio classification [17].

2.3 Audio Classification for Car Environments

Having reviewed the fundamental technologies behind audio classification, now we focus on research efforts on audio classification specifically designed for car environments. Audio classification has long been considered as a front-end technology for automatic speech recognition (ASR) for vehicular applications [2]. For example, ASR is performed once the input audio stream is determined to be of the speech type. Unfortunately, to the best of our knowledge, there are very few work on audio classification specifically designed for car environments. Rather, most work were focused on ASR. This section reviews ASR techniques concentrated on vehicular applications.

Collecting in-vehicle acoustic data was of a primary interest for ASR research. Hansel *et al.* collected audio data of a total of 14 noise conditions with +1000 speakers for speech recognition [11]. AVICAR is another widely adopted corpus of acoustic data collected in a car environment [16]. The data were obtained in varying noise conditions, *i.e.,* idle, 35mph, and 55mph with window open/close. However, existing in-vehicle data are concentrated on speech data only, not including other audio types such as speech+music, music, *etc.*

Speech enhancement techniques were proposed to improve the performance of speech recognition in car environments. A speech enhancement algorithm based on the spectral subtraction algorithm was proposed [37]. A specific focus was given to integration of the algorithm into a Virtex-4 field-programmable gate array (FPGA) device [37]. Kleinschmidt considered two methods for speech enhancement, *i.e.*, the likelihood maximization-based Mel-filterbank noise subtraction, and frequency-domain spectrum subtraction [13]. However, these algorithms do not take into account the time-varying characteristics of noise in a vehicle interior.

Noise reduction mechanisms were researched for speech recognition in an in-vehicle environment. Ahn and Ko studied a noise reduction algorithm based on eigendecomposition [1]. A microphone array was used to improve the performance of speech recognition in vehicular environment [42]. Yapanel and Hansen proposed a new feature estimation approach to enable noise-robust speech recognition specifically concentrating on time-varying vehicle noise [41]. However, due to the large overlapping frequency range between noise and music/speech, it is impossible to completely remove in-vehicle noise.

Close to our proposed framework are adaptive approaches. Mporas *et al.*

developed an adaptive speech enhancement method based on varying acoustic environment characteristics in a vehicular environment [25]. Noise models were generated and input speech data were compared with the noise model to make smart decisions in directing subsequent speech processing systems [2]. However, extracting noise information from noisy speech data gets very tricky when background music is being played, which is quite usual in everyday driving experience. In addition, while Akbacak and Hansen [2] used artificially degraded speech data, we collect and utilize real-world audio data of various types including music+speech, and music.

Chapter 3

Preliminaries

3.1 Preliminary Results

We examined the performance of a non-adaptive approach for audio classification under varying driving environments to get an insight that justifies the need for an adaptive solution. For this study, classification models for varying audio types, *i.e.,* speech, music, speech+music, and background noise were generated based on supervised learning under an uniform environment, a stopped vehicle. More specifically, an input audio stream was divided into frames, and selected features were extracted from each frame. These features were then labeled according to the audio types in generating the classification models. In the meantime, the test data were collected from different driving environments, *i.e.,* highway, local road, and city. The collected test data were classified based on the generated classification models.

Figure 3.1 depicts the average classification accuracy for the input audio data of different types collected from varying driving environments. The highest accuracy was achieved in the stopped vehicle (labeled in the graph as 'idle') because the test data were collected from the same driving environment. On the other hand, the accuracy degraded for different driving environments. Especially the accuracy for speech and speech+music was significantly impacted while the accuracy for music was kept high regardless of varying driving environments potentially because of the distinctive acoustic characteristics of music.

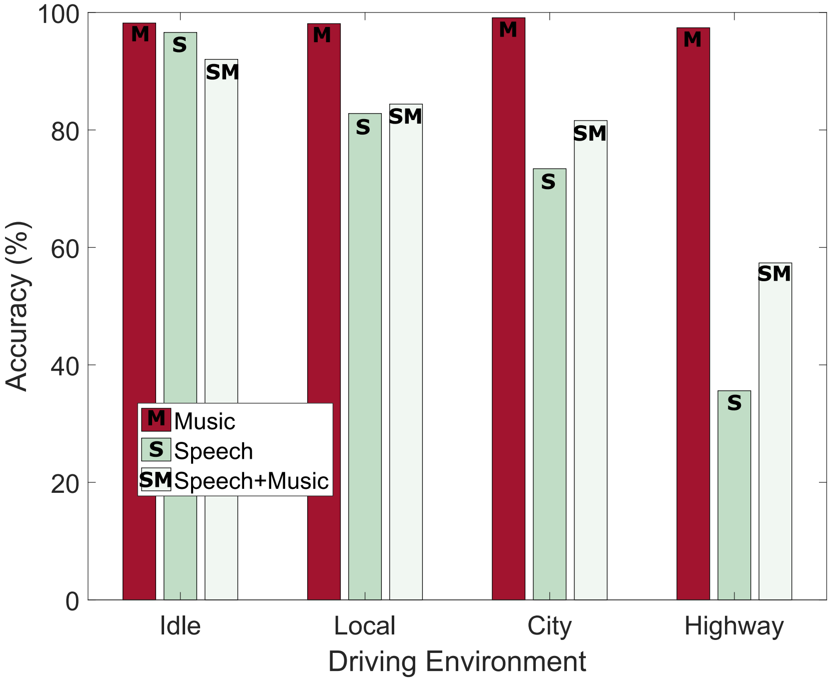

Figure 3.1: The average classification accuracy of a non-adaptive approach for varying driving environments.

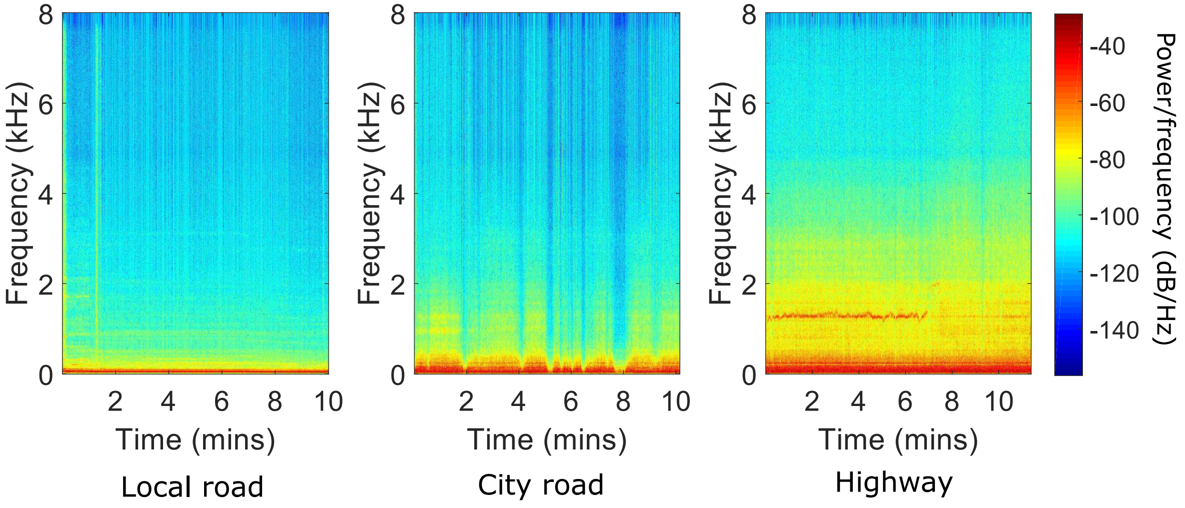

Figure 3.2: Spectrograms of in-vehicle noise for different driving environments.

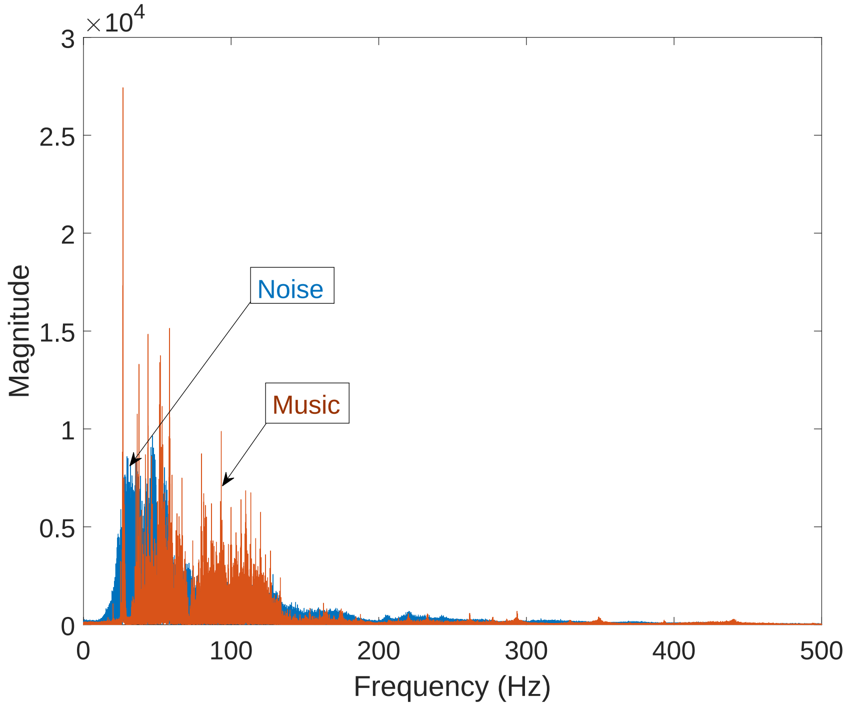

Figure 3.3: Results of FFT for noise and music.

Another notable observation was that the degree of performance degradation was different depending on varying driving environments because of unique attributes of noise in each environment. More specifically, the spectrograms of cabin noise show that the frequency range and the intensity of noise were unique in each driving environment (Figure 3.2). Unfortunately, however, the wide frequency range of noise implies that applying a simple bandpass filter is not enough to completely remove the noise due to the large amount of overlapping frequencies compared with speech and music considering that typical human voice and music have frequency ranges of 300Hz $\sim$ 3Khz, and 20Hz $\sim$ 20Khz, respectively [29]. Figure 3.3 displays the frequency distributions of noise and music for highway, which specifies the significant amount of overlapping frequency ranges.

3.2 Implications

These preliminary results suggest that a non-adaptive approach is not a suitable method for audio classification in car environment. We thus propose to reduce the performance degradation by individually training classification models for varying driving environments and adaptively applying the models depending on the real-time identification of driving environments. More specifically, in this paper, we consider four different driving environments for the purpose of case study, *i.e.,* local road, city, highway, and stopped vehicle. In the following chapter, we present the details of the proposed adaptive audio classification framework.

Chapter 4

System Design

This chapter presents an overview of the proposed framework, followed by the details of main components of the framework.

4.1 Overview

An overview of the proposed framework is illustrated in Figure 4.1. The driving environment detection module identifies the current driving environment. Collected audio data are then sent to the feature extraction module that extracts appropriate features from collected audio data. Obtained features are used in two different modes. In the training mode, they are utilized to train classification models corresponding to the current driving environment. In the testing mode, extracted features are provided to the classification algorithm in order to classify the input data into music, speech, speech+music, and noise; once the classification is completed, the input data are used for training to consolidate the classification models.

This framework can be implemented as a smartphone app. When the user starts this app, the user is provided with two options, *i.e.,* training and testing. In the training mode, the user will input his/her vehicle type and tire model to generate customized classification models for varying driving environments. While the user is driving in this training mode, the app will collect audio data (the user will be asked to label the audio type as speech, speech+music, or music), automatically identify the driving environment, and generate appropriate classification models.

These classification models can then be used in the testing mode in which the app will perform classification of incoming audio streams in real time. An important use case of the proposed work is that it can be used to reinforce existing in-vehicle ASR techniques. For example, unlike traditional ASR approaches, the proposed framework allows for detection of various audio types including speech, and speech+music. If speech+music is detected, the volume of music can be temporarily and automatically reduced to significantly improve the ASR performance, while existing non-adaptive in-vehicle ASR solutions assume only speech audio type resulting in degraded performance when there is background noise of diverse characteristics or even music being played.Ta

4.2 Driving Environment Detection

For our case study, four generic driving environments are used. These driving environments can be identified easily using the GPS module of a smartphone. More specifically, the vehicle speed and vehicle location on a map indicate whether a vehicle is currently stopped, or is driving on a highway, *etc.* However, there are numerous other significant factors that affect the level of in-vehicle noise essentially creating different driving environments. We define such major factors as the following: vehicle types, tire conditions, adverse weather conditions, open vehicle windows, and quality of road surface.

Recent advances in mobile sensing and cloud computing technologies enable effective identification of driving environments. For example, the barometer sensor of a smartphone can be used to monitor if the vehicle window is open [38]. Research has shown that road conditions can also be monitored with a smartphone [8]. An app can even be integrated with the noise map to identify areas with exceptionally high noise. Real time weather conditions like wind speed and direction can be easily obtained and used for defining a new classification model [4]. Vehicle-specific factors such as vehicle types and tire conditions can be input by the user since the classification models will be generated typically for their own vehicle.

4.3 Feature Extraction

It is of paramount importance to select appropriate features to achieve high classification accuracy [40]. We consider 16 primary features including well-known and effective features for audio classification into speech, music, and and environmental sound [21]. These features extracted using MirToolbox [15] are summarized in Table 4.1. The wrapper-based feature selection method is then used to find a feature set that optimizes the classification accuracy for each classifier that we use in our classification algorithm [10]. The details of the classification algorithm is described in the following section.

4.4 Classification

Extracted features are used to train classification models for each audio type and driving environment. Support vector machine (SVM) is known to show superior performance compared with other classifiers [22][40]. We adopt SVM to create the

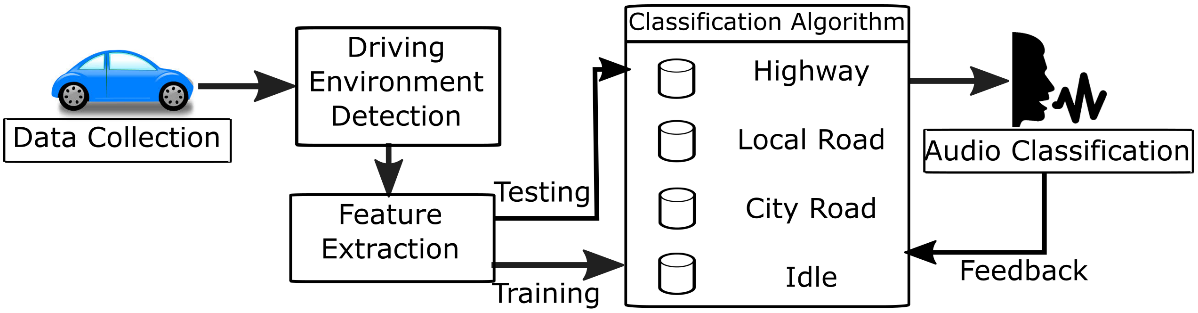

Figure 4.1: Overview of the proposed framework.

Table 4.1: Features considered for the proposed framework.

Type	Feature
Time domain	Root mean square
	Zero-crossing rate
	High zero-crossing rate ratio
	Low short time energy ratio
	Noise frame ratio
	Silence frame ratio
Spectral domain	Spectral centroid
	Spectral spread
	Spectral flux
	Spectral kurtosis
	Spectral roll-off frequency
	Band period
	Subband energy distribution
	Mel frequency cepstral coefficients
	Linear predictive cepstral coefficients
	Linear spectral pairs

models based on supervised learning.

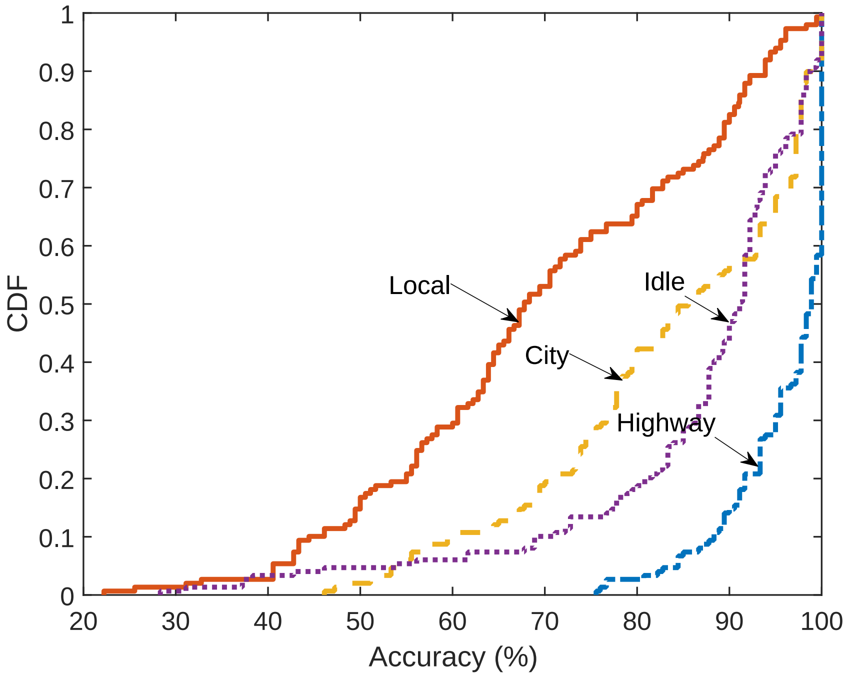

Figure 4.2: CDF of accuracy for music.

A challenge is that audio data need to be classified into multiple audio types using SVM that is inherently a binary classifier. A known approach to implement multi-class audio classification using SVM is to define ‘speech’ and ‘non-speech’ classes and to classify the input data into these two classes first [40]. More specifically, audio types of speech and speech+music are combined to create the ‘speech’ class, and the combination of music, and environmental sound is used to build the ‘non-speech’ class. Once the input data is found to be in the ‘speech’ class, the input data is subsequently classified into speech and speech+music. Similarly if the input data belongs to the ‘non-speech’ class, it is further classified into music and environmental sound (noise). However, when this approach was applied to our data set, the accuracy for music was significantly degraded as shown in the cumulative distribution function (CDF) graph of the music accuracy for different driving environments (Figure 4.2), potentially due to the 'combination effect' of multiple audio data types.

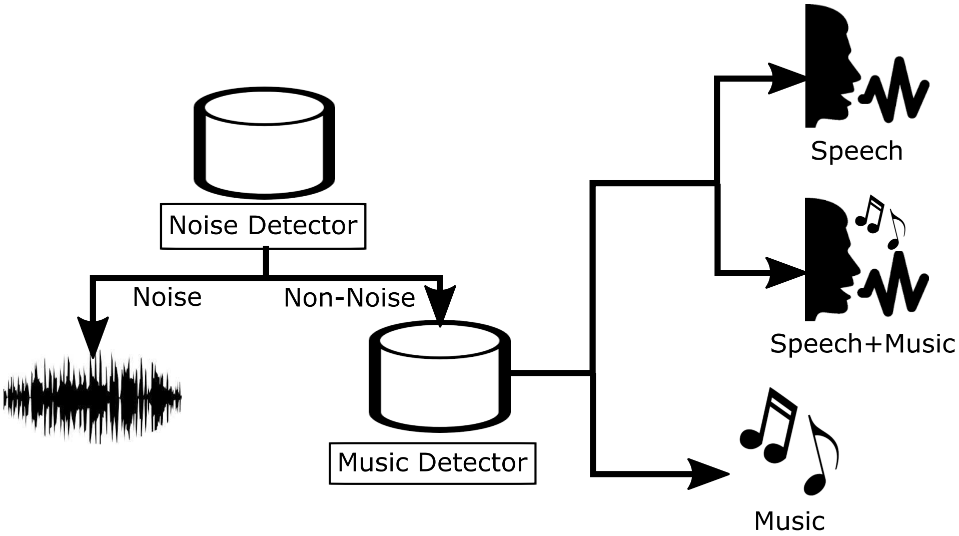

Figure 4.3: Overview of classification algorithm.

In order to address this challenge of degraded accuracy for music, a new classification algorithm is designed. An overview of the proposed algorithm is displayed in Figure 4.3. It consists of two modules: Noise Detector, and Music Detector. The noise detector performs testing of input data using three types of classifiers, *i.e.*, 'speech vs noise', 'music vs noise', and 'speech+music vs noise'. Essentially they are a set of one-versus-one classifiers that are specifically focused on identifying noise. Each classifier determines that the input data is noise if the accuracy is greater than a threshold. Consequently, the noise detector specifies that the input data is noise if all three classifiers indicate that it is noise. This way we eliminate the need for merging audio data of different types into a single class. Once the noise detector finds that the input data is of non-noise type, *i.e.,* either speech, speech+music, or music, the music detector starts to run to determine if the input data is of music type. The design of the music detector is similar to the noise detector; the difference is that it consists of a set of one-versus-one classifiers that are used to discern music. More specifically it tests the input data stream against two types of classifiers 'speech vs music', and 'speech+music vs music' to identify music. Finally, if the result is non-music, a binary classifier 'speech+music vs speech' is used to differentiate speech, and speech+music.

Chapter 5

Experimental Results

In this chapter, we evaluated the performance of the proposed framework. More specifically, we measured the classification accuracy of the proposed framework, a non-adaptive approach, and an adaptive solution with the classification algorithm based on 'speech' vs 'non-speech' [40], where the classification accuracy was defined as the total number of correctly identified frames divided by the total number of input audio frames. The effect of varying music genres on the accuracy was also considered.

5.1 Experimental Setup

Audio data were collected from cities of Brookings (Local), Sioux Falls (City), Interstate-29 (Highway), and a stopped vehicle in South Dakota, United States. More specifically, 10mins of audio data for each type were collected, *i.e.,* speech, music (4 different genres), speech+music, and noise from varying driving environments, totalling 280mins of data. For testing, 5mins of audio data were collected for each audio type and for each driving environment. Consequently, audio data of the total data size of about 420mins were used. The audio streams were 16kHz mono with 16bit per sample. Audio streams were divided into non-overlapping frames of 100ms, and for analysis with longer frames, clips of 1 second were used.

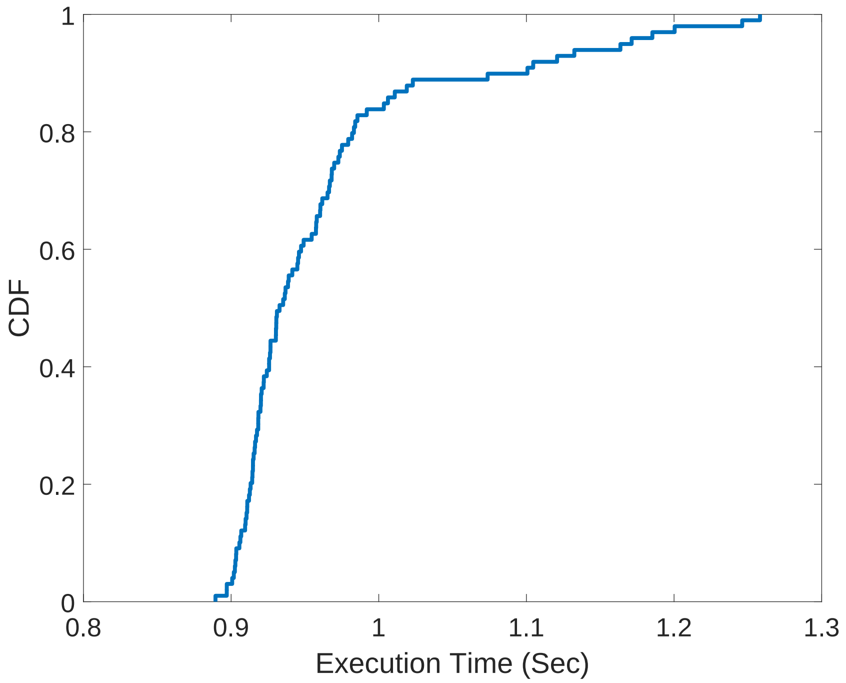

Figure 5.1: CDF of execution time.

5.2 Execution Time

The system execution time is important in applying the proposed system to real-time applications such as mobile apps for smart transportation. We measured the execution time for 1-second clips. The results are depicted in Figure 5.1. The CDF graph indicates that, in most cases, the classification was completed in around 1 second which makes feasible to perform classification approximately every 2 seconds.

5.3 Classification Accuracy

We first analyze the accuracy of the proposed framework in comparison with a non-adaptive approach as presented in Chapter 3. The average accuracy of the proposed adaptive approach and a non-adaptive approach is shown in Tables 5.1

Table 5.1: Accuracy of the proposed framework.

Category	Highway	Local	City	Idle	Average
Speech	92.3%	91.4%	96.5%	96.3%	94.1%
Music	97.1%	92.9%	97.8%	98.2%	96.5%
Speech+Music	93.5%	89.8%	88.7%	92.0%	91.0%
Average	94.3%	91.4%	94.3%	95.5%	

Table 5.2: Accuracy of a non-adaptive approach.

Category	Highway	Local	City	Idle	Average
Speech	35.6%	82.8%	73.4%	96.6%	72.1%
Music	97.4%	98.1%	99.1%	98.2%	98.2%
Speech+Music	57.4%	84.4%	81.6%	92.0%	78.9%
Average	63.5%	88.4%	84.7%	95.6%	

and 5.2 respectively. As shown, the accuracy was significantly improved for all driving environments compared with a non-adaptive approach. The highest improvement in the accuracy was achieved in the highway: the accuracy was improved by 166% and 64% for speech, speech+music, respectively, compared with a non-adaptive approach, whereas the accuracy for music was not improved much because the music type was accurately classified in a non-adaptive approach because of the distinctive acoustic characteristics of music. An interesting observation was that no significant correlation between the noise level of driving environments and the accuracy was found after applying the adaptive method, *e.g.,* the accuracy for local road was smaller than that for the highway. This result indicated that the impact of varying noise levels in different driving environments was successfully relieved by applying adaptive classification models.

Table 5.3: Accuracy of an adaptive approach in comparison with a classification algorithm for in-vehicle environment.

Category	Highway	Local	City	Idle	Average
Speech	94.9%	94.5%	92.8%	96.3%	94.6%
Music	96.0%	69.1%	82.9%	86.2%	83.6%
Speech+Music	89.7%	89.0%	85.7%	94.7%	89.8%
Average	93.5%	84.2%	87.1%	92.4%	

We then used a different classification algorithm [40]. Recall that this Xie's algorithm is based on binary classification into 'speech' and 'non-speech' data types. The accuracy obtained when the Xie's algorithm was applied is shown in Table 5.3. A notable observation was that the Xie's algorithm had significantly low accuracy for music, as previously noted in the CDF graph (Figure 4.2). More specifically, the average accuracy for music was 13.4% smaller than that for the proposed classification algorithm. The reason was that speech and music+speech data were combined as a single 'speech' class, and when music data were compared with the 'speech' class, a lot of music frames were classified into the 'speech' class. This also explained the slight increase of the average accuracy for speech. In contrast, the proposed classification algorithm does not rely on such combined classes, and consequently, it achieved 0.9%, 8.6%, 8.2%, and 3.4% higher accuracy for highway, local, city, and idle, respectively, compared with the Xie's algorithm in our experimental settings.

5.4 Effect of Music Genre

We performed supervised training with only a single music genre, *i.e.,* pop music. In order to further improve the performance, we trained the classification model

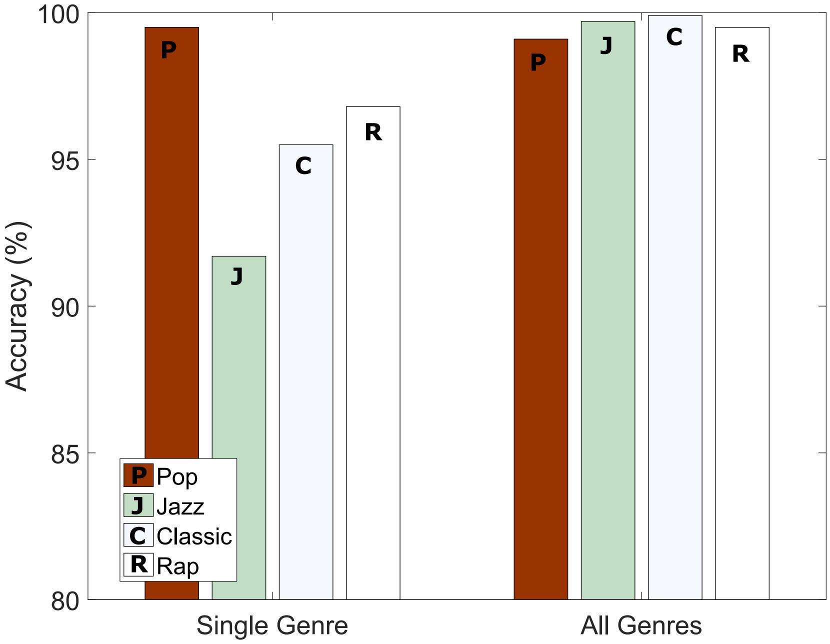

Figure 5.2: Accuracy for music with different genres.

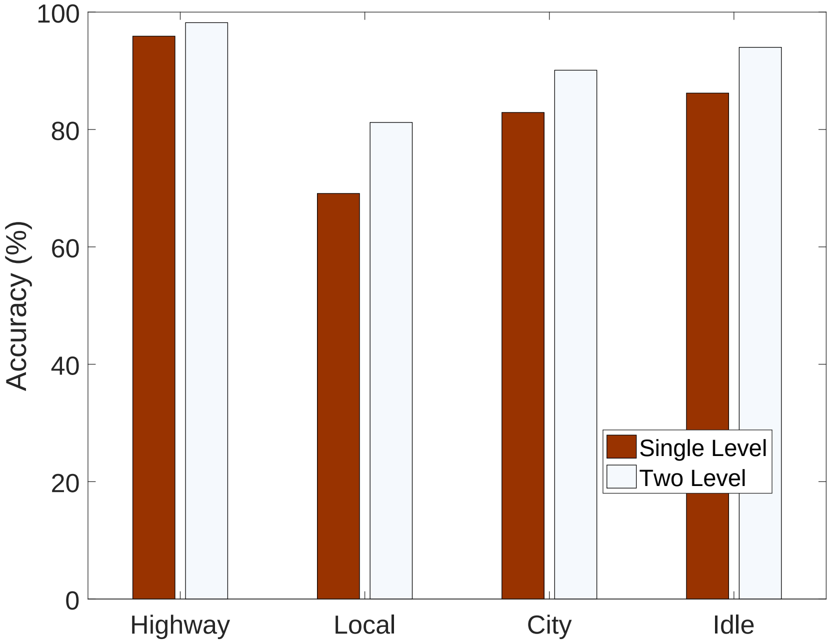

Figure 5.3: Effect of two-level SVM

for music taking into account different music genres (*i.e.,* jazz, classic, and rap) and measured the accuracy. For this set of experiments, we used the highway data. The results are depicted in Figure 5.2. As shown, when it was trained with a single musical genre, the accuracy for other music genres was degraded. It was observed that after training with multiple music genres, the accuracy was increased by 8%, 5%, and 3% for jazz, classic, and rap, respectively.

5.5 Effect of Two-Level SVM

We presumed that the reason for the low accuracy for Music is because Music data are quite similar to Speech+Music. Based on the CDF graph, if the accuracy for Music is lower than 30%, it is highly likely to be speech in all driving environments. Similarly, we accepted it as music if the accuracy is greater than 90%. The problem was the results between 30% and 90%. We need to ensure that the input data was really music by performing a separate classification. The results after applying one more classification are shown in Figure 5.3.

Chapter 6

Conclusion

6.1 Summary

We have presented the design, implementation, and evaluation of an adaptive multi-class music classification framework specifically designed for in-vehicle environments. The proposed framework classifies an input audio stream into four different audio types, *i.e.,* music, speech, speech+music, and noise. The classification accuracy was substantially improved by generating individual classification models for varying driving environments and adaptively applying the models based on real-time identification of driving environments. A novel classification algorithm based on effective feature selection allowed for improved classification accuracy for driving environments with varying noise characteristics.

6.2 Contributions

As outcomes of this work, we collected more than 420mins of real-world audio data (*i.e.,* speech, music, speech+music, and noise) with a smartphone microphone from various driving environments (*i.e.,* highway, local road, city, and stopped vehicle). Additionally, a case study performed with different driving environments demonstrated that the proposed framework substantially improved the mean audio classification accuracy by up to 166%, and 64% for speech, and speech+music, respectively, compared with a non-adaptive approach in our experimental settings.

6.3 Future Work

Future work includes identification of the effectiveness of other features such as pitch-density-based features [40], and bottleneck features extracted using deep neural network [44], analysis of the effect of new features, and incorporation of those features in order to further improve the accuracy of audio classification.

Another important future work is to evaluate the performance of the proposed framework in other test scenarios, *e.g.,* with different vehicle models, and weather conditions.

Finally, the proposed framework will be applied to a variety of smart car apps. Examples include an automatic voice control app which that can identify human voice to automatically adjust the car radio volume.

References

[1] Sungjoo Ahn and Hanseok Ko. Background noise reduction via dual-channel scheme for speech recognition in vehicular environment. *IEEE Transactions on Consumer Electronics*, 51(1):22–27, 2005.

[2] Murat Akbacak and John HL Hansen. Environmental sniffing: noise knowledge estimation for robust speech systems. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(2):465–477, 2007.

[3] Enrique Alexandre, Lucas Cuadra, Lorena Alvarez, and Manuel Rosa-Zurera. Nn-based automatic sound classifier for digital hearing aids. In *Intelligent Signal Processing, 2007. WISP 2007. IEEE International Symposium on*, pages 1–6. IEEE, 2007.

[4] German Castignani, Thierry Derrmann, Raphaël Frank, and Thomas Engel. Driver behavior profiling using smartphones: A low-cost platform for driver monitoring. *IEEE Intelligent Transportation Systems Magazine*, 7(1):91–102, 2015.

[5] Lei Chen, Sule Gunduz, and M Tamer Ozsu. Mixed type audio classification with support vector machine. In *Multimedia and Expo, 2006 IEEE International Conference on*, pages 781–784. IEEE, 2006.

[6] Ebru Dogan, Mustafa Sert, and Adnan Yazici. Content-based classification and segmentation of mixed-type audio by using mpeg-7 features. In *Advances in Multimedia, 2009. MMEDIA'09. First International Conference on*, pages 152–157. IEEE, 2009.

[7] Florian Eyben. *Real-time speech and music classification by large audio feature space extraction*. Springer, 2015.

[8] Lars Forslöf and Hans Jones. Roadroid: continuous road condition monitoring with smart phones. In *IRF 17th World Meeting and Exhibition, Riyadh, Saudi Arabia. Available from http://www. roadroid. com/common/References/IRF*, volume 202013, 2013.

[9] Zhong-Hua Fu, Jhing-Fa Wang, and Lei Xie. Noise robust features for speech/music discrimination in real-time telecommunication. In *Multimedia and Expo, 2009. ICME 2009. IEEE International Conference on*, pages 574–577. IEEE, 2009.

[10] Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *Journal of machine learning research*, 3(Mar):1157–1182, 2003.

[11] John HL Hansen, Pongtep Angkititrakul, Jay P Plucienkowski, Stephen Gallant, Umit H Yapanel, Bryan L Pellom, Wayne H Ward, and Ronald A Cole. " cu-move": analysis & corpus development for interactive in-vehicle speech systems. In *INTERSPEECH*, pages 2023–2026, 2001.

[12] Ji-Soo Keum and Hyon-Soo Lee. Speech/music discrimination using spectral peak feature for speaker indexing. In *Intelligent Signal Processing and Communications, 2006. ISPACS'06. International Symposium on*, pages 323–326. IEEE, 2006.

[13] Tristan Friedrich Kleinschmidt. *Robust speech recognition using speech enhancement*. PhD thesis, Queensland University of Technology, 2010.

[14] Anurag Kumar and Bhiksha Raj. Audio event detection using weakly labeled data. In *Proceedings of the 2016 ACM on Multimedia Conference*, pages 1038–1047. ACM, 2016.

[15] Olivier Lartillot, Petri Toiviainen, and Tuomas Eerola. A matlab toolbox for music information retrieval. In *Data analysis, machine learning and applications*, pages 261–268. Springer, 2008.

[16] Bowon Lee, Mark Hasegawa-Johnson, Camille Goudeseune, Suketu Kamdar, Sarah Borys, Ming Liu, and Thomas S Huang. Avicar: audio-visual speech corpus in a car environment. In *INTERSPEECH*, pages 2489–2492, 2004.

[17] Honglak Lee, Peter Pham, Yan Largman, and Andrew Y Ng. Unsupervised feature learning for audio classification using convolutional deep belief networks. In *Advances in neural information processing systems*, pages 1096–1104, 2009.

[18] Dongge Li, Ishwar K Sethi, Nevenka Dimitrova, and Tom McGee. Classification of general audio data for content-based retrieval. *Pattern recognition letters*, 22(5):533–544, 2001.

[19] Chungsoo Lim, Seong-Ro Lee, and Joon-Hyuk Chang. Efficient implementation of an svm-based speech/music classifier by enhancing temporal locality in support vector references. *IEEE Transactions on Consumer Electronics*, 58(3), 2012.

[20] Chuan Liu, Lei Xie, Helen Meng, et al. Classification of music and speech in mandarin news broadcasts. In *Proc. of the 9th Nat. Conf. on Man-Machine Speech Communication (NCMMSC), Huangshan, Anhui, China*, 2007.

[21] Lie Lu, Hao Jiang, and HongJiang Zhang. A robust audio classification and segmentation method. In *Proceedings of the ninth ACM international conference on Multimedia*, pages 203–211. ACM, 2001.

[22] Lie Lu, Hong-Jiang Zhang, and Hao Jiang. Content analysis for audio classification and segmentation. *IEEE Transactions on speech and audio processing*, 10(7):504–516, 2002.

[23] Lie Lu, Hong-Jiang Zhang, and Stan Z Li. Content-based audio classification and segmentation by using support vector machines. *Multimedia systems*, 8(6):482–492, 2003.

[24] Annamaria Mesaros, Toni Heittola, and Tuomas Virtanen. Tut database for acoustic scene classification and sound event detection. In *Signal Processing Conference (EUSIPCO), 2016 24th European*, pages 1128–1132. IEEE, 2016.

[25] Iosif Mporas, Todor Ganchev, Otilia Kocsis, and Nikos Fakotakis. Dynamic selection of a speech enhancement method for robust speech recognition in moving motorcycle environment. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pages 5176–5179. IEEE, 2011.

[26] Loris Nanni, Yandre MG Costa, Alessandra Lumini, Moo Young Kim, and

Seung Ryul Baek. Combining visual and acoustic features for music genre classification. *Expert Systems with Applications*, 45:108–117, 2016.

[27] Yannis Panagakis, Constantine L Kotropoulos, and Gonzalo R Arce. Music genre classification via joint sparse low-rank representation of audio features. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, 22(12):1905–1917, 2014.

[28] Costas Panagiotakis and Georgios Tziritas. A speech/music discriminator based on rms and zero-crossings. *IEEE Transactions on multimedia*, 7(1):155–166, 2005.

[29] Max Power. Psychology of language. 2003.

[30] Thiruvengatanadhan Ramalingam and P Dhanalakshmi. Speech/music classification using wavelet based feature extraction techniques. *Journal of Computer Science*, 10(1):34–44, 2014.

[31] Jianfeng Ren, Xudong Jiang, Junsong Yuan, and Nadia Magnenat-Thalmann. Sound-event classification using robust texture features for robot hearing. *IEEE Transactions on Multimedia*, 19(3):447–458, 2017.

[32] Eric Scheirer and Malcolm Slaney. Construction and evaluation of a robust multifeature speech/music discriminator. In *Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on*, volume 2, pages 1331–1334. IEEE, 1997.

[33] Gregory Sell and Pascal Clark. Music tonality features for speech/music

discrimination. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pages 2489–2493. IEEE, 2014.

[34] Dan Stowell, Dimitrios Giannoulis, Emmanouil Benetos, Mathieu Lagrange, and Mark D Plumbley. Detection and classification of acoustic scenes and events. *IEEE Transactions on Multimedia*, 17(10):1733–1746, 2015.

[35] Jun Wang, Qiong Wu, Haojiang Deng, and Qin Yan. Real-time speech/music classification with a hierarchical oblique decision tree. In *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, pages 2033–2036. IEEE, 2008.

[36] WQ Wang, W Gao, and DW Ying. A fast and robust speech/music discrimination approach. In *Information, Communications and Signal Processing, 2003 and Fourth Pacific Rim Conference on Multimedia. Proceedings of the 2003 Joint Conference of the Fourth International Conference on*, volume 3, pages 1325–1329. IEEE, 2003.

[37] Jim Whittington, Kapeel Deo, Tristan Kleinschmidt, and Michael Mason. Fpga implementation of spectral subtraction for in-car speech enhancement and recognition. In *Signal Processing and Communication Systems, 2008. ICSPCS 2008. 2nd International Conference on*, pages 1–8. IEEE, 2008.

[38] Myounggyu Won, Shaohu Zhang, Appala Chekuri, and Sang H Son. Enabling energy-efficient driving route detection using built-in smartphone barometer

sensor. In *Intelligent Transportation Systems (ITSC), 2016 IEEE 19th International Conference on*, pages 2378–2385. IEEE, 2016.

[39] Ming-Ju Wu and Jyh-Shing R Jang. Combining acoustic and multilevel visual features for music genre classification. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 12(1):10, 2015.

[40] Lei Xie, Zhong-Hua Fu, Wei Feng, and Yong Luo. Pitch-density-based features and an svm binary tree approach for multi-class audio classification in broadcast news. *Multimedia systems*, 17(2):101–112, 2011.

[41] Umit H Yapanel and John HL Hansen. A new perspective on feature extraction for robust in-vehicle speech recognition. In *INTERSPEECH*, 2003.

[42] Tao Yu and John HL Hansen. An efficient microphone array based voice activity detector for driver's speech in noise and music rich in-vehicle environments. In *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pages 2834–2837. IEEE, 2010.

[43] Saadia Zahid, Fawad Hussain, Muhammad Rashid, Muhammad Haroon Yousaf, and Hafiz Adnan Habib. Optimized audio classification and segmentation algorithm by using ensemble methods. *Mathematical Problems in Engineering*, 2015, 2015.

[44] Bihong Zhang, Lei Xie, Yougen Yuan, Huaiping Ming, Dongyan Huang, and Mingli Song. Deep neural network derived bottleneck features for accurate audio

classification. In *Multimedia & Expo Workshops (ICMEW), 2016 IEEE International Conference on*, pages 1–6. IEEE, 2016.

[45] Yongwei Zhu, Qibin Sun, and Susanto Rahardja. Detecting musical sounds in broadcast audio based on pitch tuning analysis. In *Multimedia and Expo, 2006 IEEE International Conference on*, pages 13–16. IEEE, 2006.