Master's Thesis

석사 학위논문

Acoustic Event Detection Using Double-Layer Classification

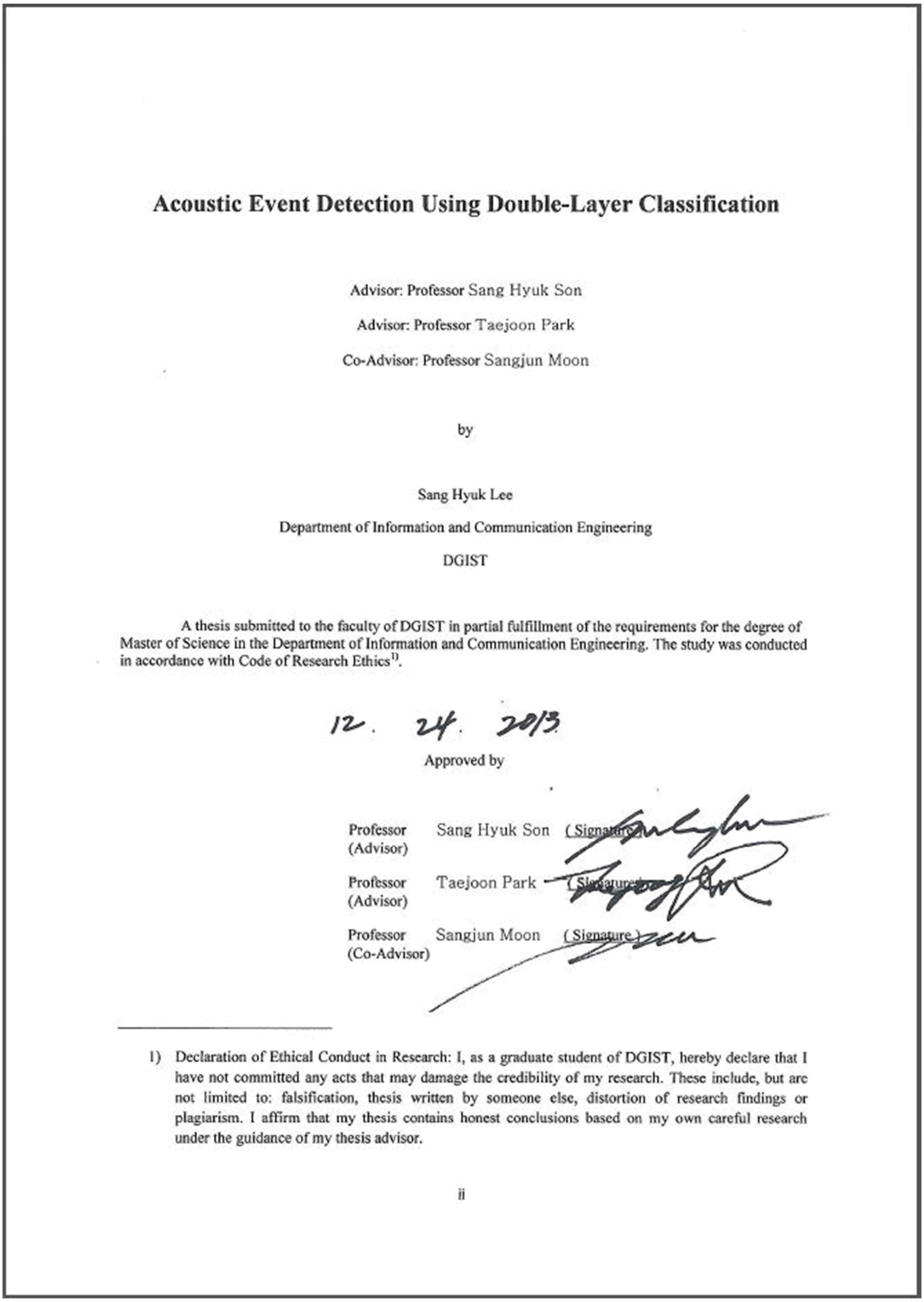
Sang Hyuk Lee (이 상 혁 李 相 赫)

Department of Information and Communication Engineering

정보통신융합전공

DGIST

2014



Acoustic Event Detection Using Double-Layer Classification

Advisor: Professor Sang Hyuk Son

Advisor: Professor Taejoon Park

Co-Advisor: Professor Sangjun Moon

by

Sang Hyuk Lee

Department of Information and Communication Engineering

DGIST

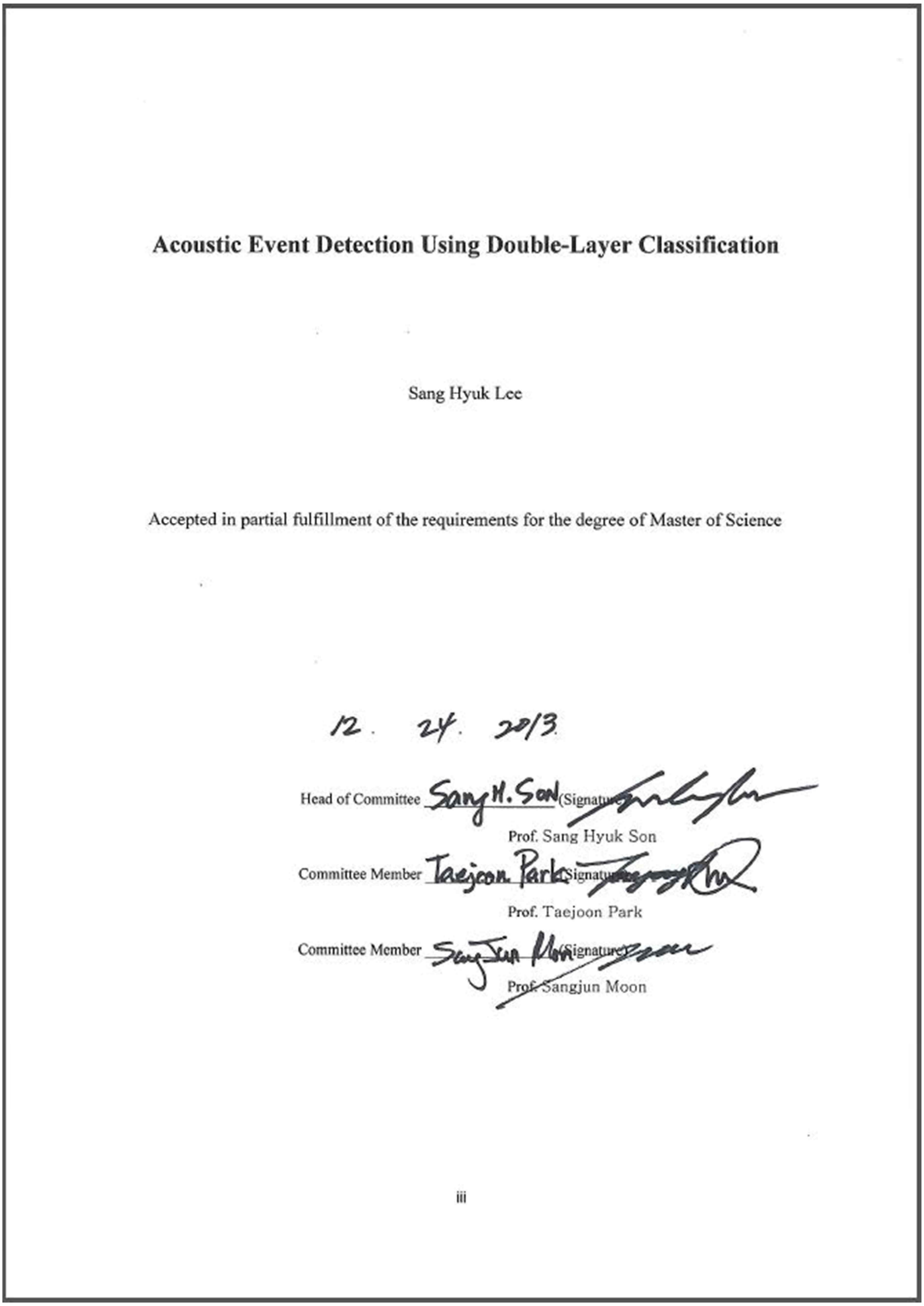
A thesis submitted to the faculty of DGIST in partial fulfillment of the requirements for the degree of Master of Science in the Department of Information and Communication Engineering. The study was conducted in accordance with Code of Research Ethics[1].

12. 24. 2013

Approved by

Professor (Advisor)	Sang Hyuk Son	(Signature)
Professor (Advisor)	Taejoon Park	(Signature)
Professor (Co-Advisor)	Sangjun Moon	(Signature)

[1] Declaration of Ethical Conduct in Research: I, as a graduate student of DGIST, hereby declare that I have not committed any acts that may damage the credibility of my research. These include, but are not limited to: falsification, thesis written by someone else, distortion of research findings or plagiarism. I affirm that my thesis contains honest conclusions based on my own careful research under the guidance of my thesis advisor.



Acoustic Event Detection Using Double-Layer Classification

Sang Hyuk Lee

Accepted in partial fulfillment of the requirements for the degree of Master of Science

12. 24. 2013

Head of Committee Sang H. Son (Signature)

Prof. Sang Hyuk Son

Committee Member Taejoon Park (Signature)

Prof. Taejoon Park

Committee Member SangJun Moon (Signature)

Prof. Sangjun Moon

iii

MS/IC
201222009

이 상 혁, Sang Hyuk Lee, Acoustic Event Detection Using Double-Layer Classification. Department of Information and Communication Engineering, 2013, 39 p.
Advisors: Professor Son, Sang Hyuk and Professor Park, Taejoon
Co-Advisor: Professor Moon, Sangjun

ABSTRACT

The world's elderly population is expected to grow by more than triple by 2050. This indicates that detecting human activity is needed to prevent emergency situations for people living alone. Sound is an excellent resource because it has enough information to detect events and it is easy to gather. However, variability is one of the main challenges in research related to sound. To solve variability, most research focuses on selecting sound features because researches related to sound have their own purpose and suitable sound features are different for each research. In this research, we focus on classifiers to solve problem of variability. The Double-layer classification (DLC) is composed with Support Vector Machines (SVM) and the Viterbi search and detects sound events using the Hidden Markov Model (HMM). As a result, unusual sounds which occur at home such as a baby crying, a scream, a breaking glass, and a gunshot are classified by the DLC, which accuracy is 94.4%.

Keywords: Sound event detection, HMM, SVM, Viterbi search, double-layer classification, acoustic event detection, activity of daily living (ADL), context awareness

Contents

List of Figures

I. INTRODUCTION

Today, the population of those who are 65 years or older numbered 40.4 million in 2010 and they represented 13.1% of the U.S. population [1]. Also if dangerous situations that are illegal crimes and accidents were detected lately, it could cause serious problems related life. According to a 2011 report entitled, Crime in the United States, the estimated number of violent crimes reported to law enforcement was 1,203,564 and number of property crimes was 9,063,173 [2]. Because of the increasing number of elderly people, single households living alone, and emergency situations, it is very important to detect human activities or context automatically and accurately for preventing medical emergencies and finding changes of behavior. If there were automatic and stable systems that could detect human activities and specific situations as soon as possible, we can cope with these serious problems rapidly. Furthermore, this useful system can be utilized by many other fields such as home security and the healthcare system.

Recently, several studies have been made to detect human activities and context by using different devices and information. One of method uses the using deployment of a variety of wide-spread sensors. Each different type of sensor measures data from the physical world and transmits them to a main computational system. The main computational system gathers the measured data from the sensors and generates meaningful information using algorithms.

Another approach is using smartphone and sensors to recognize human activities [3]. This approach shows great performance because it combines additional sensing power, computational resources and provides the user-friendly interface of an Android smartphone. However, people cannot always be carrying a smartphone. Moreover, attaching several sensors on the human body is uncomfortable. Yet another approach detects human Activity of

Daily Living (ADL) using a variety of sensors in the home environment [4]. This approach does not require the ground truth to detect specific events since it has a learning technique where the system chooses the subset of training data that needs to be labeled by the user [5]. The ground truth is defined as classification condition based on real data from the implemented and installed system or application in real environment. Also they solve practical problem which is simultaneous activities. For example, it is possible to watch TV during preparing their breakfast. However, when recognizing human activities using wide-spread sensors, they reach one of limitations of sensors, which is energy consumption. Also if one of the sensors in a network is out of order, it is not possible to make an accurate decision.

Additionally, another way to detect events and activities is to use vision image. One of the methods of this uses a video camera system with computer vision algorithms to detect human falls [6]. They measure the bounding box ratio of the person and detect the velocity of the fall from a continuation of images. However, these ways to detect human activities require video cameras to be installed in the home which can raise privacy concerns for some.

Another way to recognize context and human activities is using sound. Sound is a good descriptor for detecting events, context awareness and healthcare. If people hear the sound of a horn, without seeing anything, they can know that a car is nearby. In other words the sound includes enough information to make a decision. Additionally, sound is always present, so we can record sound very easily. Sound event detection focuses on processing the continuous sound signal and converting it into symbolic descriptions. Sound event detection can be utilized in a variety of applications, including context-based indexing and retrieval in multimedia databases, unobtrusive monitoring in health care, and surveillance.

In the studies related to sound event detection, it is very important to select suitable sound features. In the speech recognition and sound event detection [8], [9], and [10], they usually extract Mel-Frequency Cepstral Coefficient (MFCC) features from sound signal. MFCCs have decorrelating property of cepstral analysis and also include some aspects of pronunciation. However, when we want to detect psychological symptoms such as depression, we have to use other sound features. Clinical depression belongs to the group of affective disorders in which emotional disturbances consist of prolonged periods of excessive sadness marked by reduced emotional expression and physical drive [12]. So lots of previous approaches suggest suitable sound features to detect each acoustic event show good performance such as [9], [13], and [14] with using Gaussian mixture models (GMM), Support Vector Machines (SVM) and other classifiers. However, the form of sound waves is variable in the real world. Actually, when we try to detect a sound event with MFCCs and SVM using real recording sound data, baby crying and female scream, the detection accuracy is an unsatisfactory 85.7%. Previous approaches focus on selecting suitable features to solve their own problems. Suitable features cover variability of sound, but not perfectly. Because some parts of sound is very similar with other parts and it causes problems to classify sound events.

The goal of this thesis is to suggest new reliable sound event detection scheme using the Double-Layer Classification (DLC) which consists of SVM [7] and Viterbi search algorithm [21] from Hidden Markov Model (HMM). SVM is a supervised learning model with associated learning algorithms that analyze data and recognize patterns, used for classification and regression analysis. HMM [22] is a Markov chain for which the state is only partially observable. Markov chain models the state of a system with a random variable that changes through time. Observations of HMM are related to the current state, but they are insufficient to precisely determine the state. To determine current state, the Viterbi search algorithm computes the most-likely corresponding sequences of states. The Viterbi search algorithm is a dynamic programming algorithm for finding the most likely sequence of hidden states from observations in HMM.

One of the major challenges to detect sound events is variability. The sound waveform always changes based on age, sex, location and so on. To solve the variability of sound, our approach applies HMM which is generally used in speech recognition.

HMM consists of two components and three parameters: state, observation and emission probability, transition probability and start probability. The observation is detected by SVM which is a machine learning techniques. SVM detects observations continuously and they become the observation sequence. From the observation sequence, emission probability and transition probability are calculated and set the system as parameters of HMM. After the calculated parameters and the observation sequence, the Viterbi search can find the current state from observations.

In this thesis, DLC system classifies unusual sounds in the home environment such as baby crying, a scream, breaking glasses, and gun shot. There are two HMMs in the experiment. First HMM classifies two sounds: sound of a baby crying and a scream. And second HMM classifies four sounds: sound of a baby crying, a scream, a breaking glass, and a gun shot. The total accuracy of first HMM is 96.5% and second is 93.3%.

The main contributions of the DLC system are as in the following.

- More accurate classification of sound events using SVM and the Viterbi search.
- Real-time event detection.

The Viterbi search is global search mechanism because it always considers previous information. One of main issue is solving variability of sound and the Viterbi search is very appropriate to solve variability because of global search mechanism. For example, it is possible that part of sound is similar to other sound. In the sound of a baby crying, amplitude of sound is increasing to maximum. At that time, it is very similar to the sound of a scream.

So a classifier such as SVM can determine that this sound is a scream at that time. However, the Viterbi search always considers previous information so that the DLC system can determine it is not a scream even though the part of sound is similar to a scream. Also the DLC tries to detect sound events in real-time. The DLC has the unique real-time method, called Early Detection (ED), which can make decision early if the DLC system is reading sound wave.

This thesis is composed of five chapters: introduction, related work, system overview, evaluation, and conclusion. Chapter 1, introduction, is summery of this thesis briefly and previous sound event detection research is introduced in chapter 2. In the chapter 3, system overview, The DLC system is explained and basic background knowledge is introduced. And performance analysis is shown in chapter 4 evaluation. Finally, summery of total result and analysis is in chapter 5 conculsion.

II. RELATED WORK

To increase the accuracy of event detection, [9] proposed to consider context-information when detecting a sound event. The context information is generated by the Viterbi algorithm and considered when detect sound event using GMM. Although similar sounds can occur, for example, the sound of a waterfall and white noise from a TV, they rarely occur at the same time. So if context information adds a classification process, it is very helpful to make a decision. However, this approach requires input data that is too long (4 sec, 20 sec and 40 sec) and total accuracy is low when use 4 sec data(41 % ~ 91%). Since many sounds last only a brief amount of time, such as scream and gun fire, a sound event system should classify accurately and in a short time.

Hazardous situations are classified such as gun fire, explosions, and screams accurately using HMM in [18]. However, the difference of length of each labeled sound data is variable. The length difference between a scream and a gunshot is about 28 sec on value.

Specific diseases or psychological symptoms can cause specific vocal characteristics. Depression is detected in [11] focused on symptoms which often belong to mood disorders which consist of prolonged periods of excessive sadness marked by reduced emotional expression and variation of amplitude. In their study, five main features are used: Teager Energy Operator (TEO), cepstral, prosodic, spectral and glottal features with GMM and SVM. In this research, maximum accuracy was 86.64% and minimum accuracy 72.01%. Also [15] detects depression using several sound features and GMM classifier. The total accuracy of their study was 76.33% including gender-independent tasks.

Another approach, [16], proposed to detect Parkinson's disease (PD) that includes reduced loudness, increased vocal tremor, and breathiness in speech, called *dysphonia* (inability to produce normal vocal sounds). Several sound features are using to detect *dysphonia* and classify it using Back propagation learning algorithm based on Levenberg-Marquardt algorithm [17]. Total accuracy of measurements is acceptable, however they need more improvement to be used in real environment.

MFCCs and SVM is used for sound event detection widely. [24], [25]. Their approaches detect unusual sounds at home such as a scream, a gunshot, a baby crying and so on. When noise is removed in test sound samples, false positive rate (if there is a detection of scream at any time in a non-scream clip) is just 5.54%. However, the false negative rate (if no scream is detected at any time in a scream clip) is 27.79% and this rate is very high. When they tested without noise removal, false positive rate is 12.18% and false negative rate is 8.93%. Although they tested sound samples with noisy environment, false negative rate is down than tested with noise removal. The result of [24] is irony and the total accuracy is insufficiency to use in real-environment. [25] also used SVM and MFCC, but the difference of other research with SVM is that they used low-power SVM for mobile devices. Their accuracy of classification could not reach more than 90%.

III. SYSTEM OVERVIEW – DOUBLE LAYER CLASSIFICATION

This section explains the sound event detection with a DLC composed of SVM, Viterbi search, and HMM. The system has four stages shown in Figure 1, i.e., preprocessing, feature extraction, first layer SVM, and second layer Viterbi search. The major advantage of DLC is more accurate classification methods to detect specific sound events than using just one classifier.

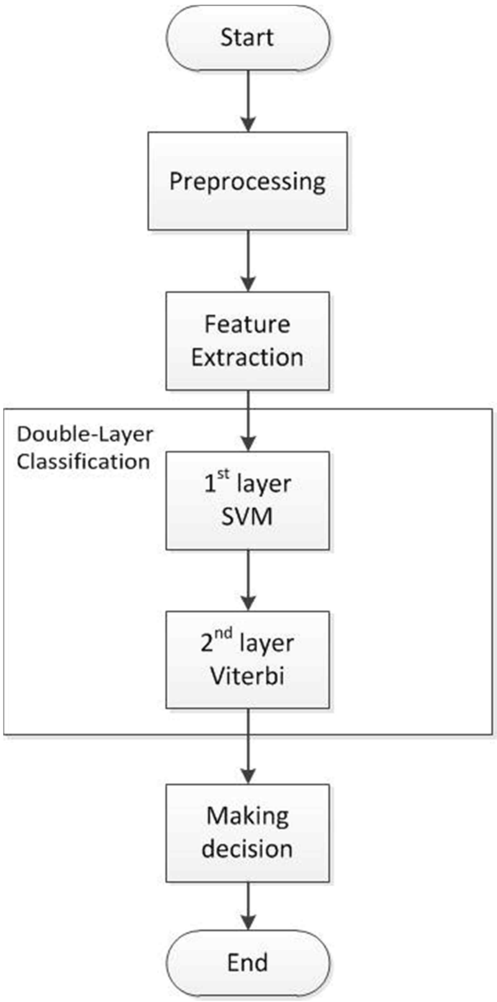


Figure 1. Architecture of the sound event detecting system

3.1. HMM(Hidden Markov Model)

3.1.1. Example of HMM

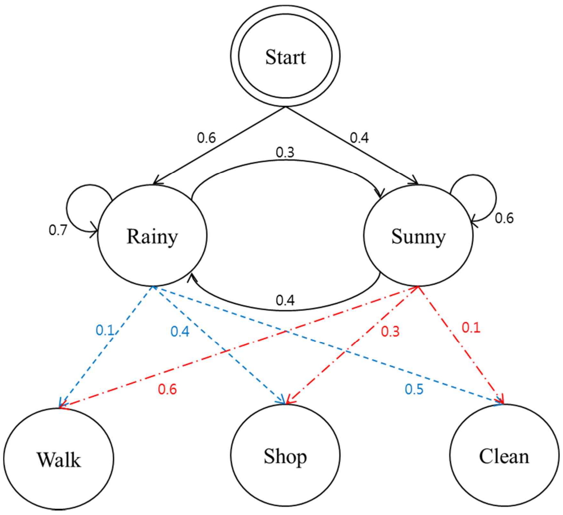


Figure 2. Example of weather HMM. HMM can derive current weather from a person's activity. There are two states: rainy and sunny, and three observations: walk, shop, and clean. Numbers in the example are probability of parameters.

HMM have two components: state and observation, and three parameters: start probability, transition probability, and emission probability. Here is weather example of HMM. As an example, consider a person who is only interested in three activities: walking, shopping, and cleaning, as shown in Figure 2. They are observations in weather HMM. Also the weather 'rainy' and 'sunny' are state of HMM.

Let's assume that we have to know current weather from only person's activity. The weather on a given day is the most influential factor to determine what to do. When today is sunny, there is a 60% chance of walking, 30% of shopping and 10% of cleaning. This probability is emission probability of state 'sunny'. And if today is sunny, tomorrow weather is same with 60% chance and changes to rainy with 40%. It is transition probability that means probability of changes between states. There is no definite weather information, but we can try to guess the current weather from the person's activities. There are two states, "rainy" and "sunny", however they cannot be observed directly which means that they are hidden. However, if the person tells us what he did on that day, we can guess the current weather because activity of the person depends on the weather. In other words, the parameters of the HMM are known. The entire system is that of an HMM.

3.1.2. Define HMM Parameters

In the HMM, there are two components: state and observation, and three parameters: start probability, transition probability, and emission probability. States mean sound events and observations are several phases of the state. Start probability is the first transition method when HMM is starting. Transition probability is the probability of changing state at next time and control transition among states. Particular observations are observed in each state and calculated by probability which is called the emission probability.

To calculate transition probability and emission probability, HMM needs a training process with training algorithms. The Baum-Welch algorithm is an expected maximization algorithm and is used to find the transition parameters of HMM. It calculates maximum likelihood estimates for transition probability of HMM from training data. Emission probability is based on the frequency of observed phases from each sound event and computed by accumulated training data.

3.1.3. HMM for the Sound Event Detection

A HMM is a widely used modeling theory which is a Markov process with unobserved states. In the sound event detection, state means specific sound event such as a baby crying or a scream and so on. Although the state of HMM is not observed directly, observations which is dependent on the state can appear. And each state has a probability distribution for every possible observation. The observations sequence contains some information to generate states. Because of this reason, the accuracy of classifying observations is very important to make the final decision. Figure 2 shows HMM for sound event detection in the DLC system.

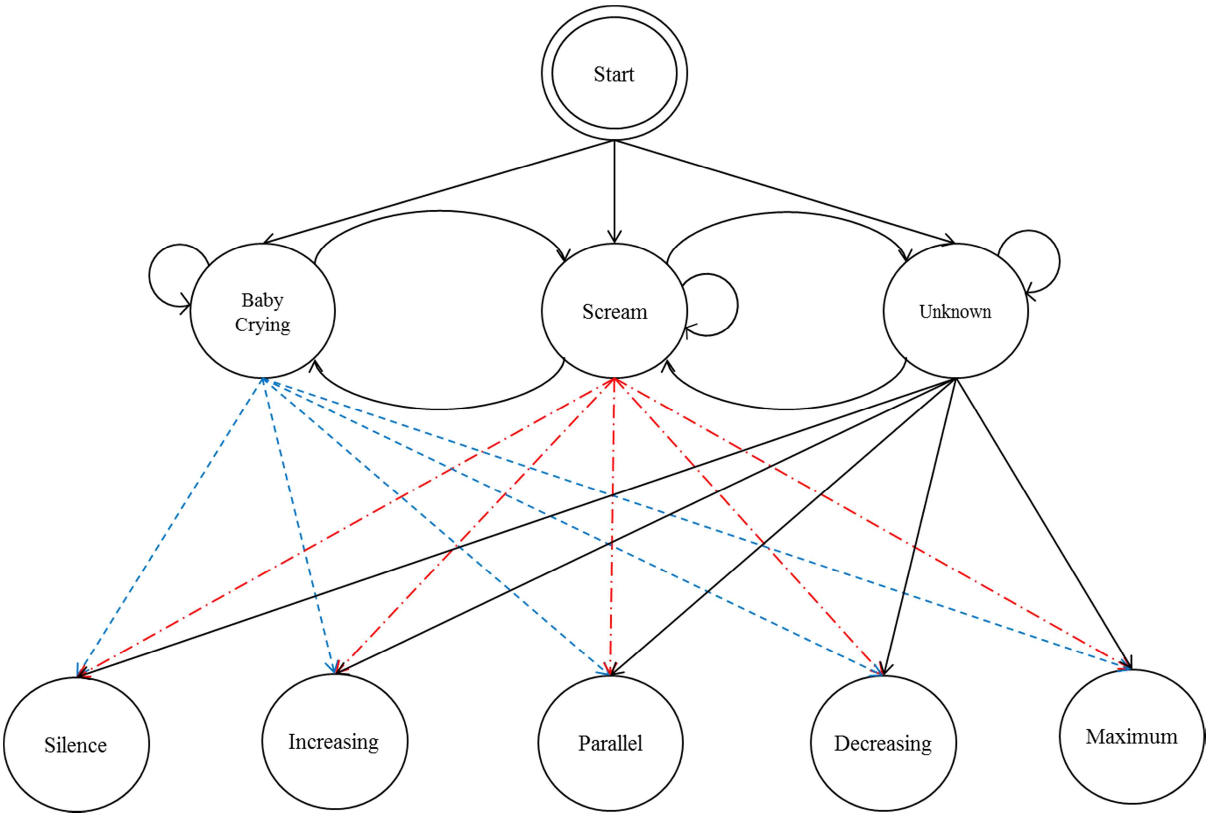


Figure 3. HMM of the DLC system. In the DLC system, there are three states: a baby crying, a scream, and unknown. Unknown is the DLC system can't detect specific sound such as a baby crying and a scream. So silence, laugh, speech and other contexts are included in unknown state. And there are five observations which are divided from baby crying and scream. Each observation is detected not only a baby crying but also a scream. However detecting frequency is not same so that it can be made by probability. The DLC system classifies sound events using difference of probability value with HMM parameters.

In this approach, we select two kinds of sound event: baby crying and a scream. The reason to select two sound classes is so that when conducting an experiment to classify a baby crying and a scream using SVM, the experimental accuracy is not satisfactory because of similar phases in the two sounds. For example, a crying sound has several phases such as silence, increasing, parallel, decreasing, and maximum phase and a scream has also the same phases.

The difference between crying and screaming is the repetition rate of each phase. This difference is an important factor to classify two sound events in HMM. The observations of HMM are defined in several phases and they are classified by the first layer SVM machine learning algorithm.

3.1.4. Definition of Observations

Every sound has specific characteristics. Some characteristics can be shown in only one kind of sound, but others are shown in several kinds of sounds. These characteristics are defined as observations (phases) in the DLC HMM. Figure 3 shows several phases of the sound of a baby crying.

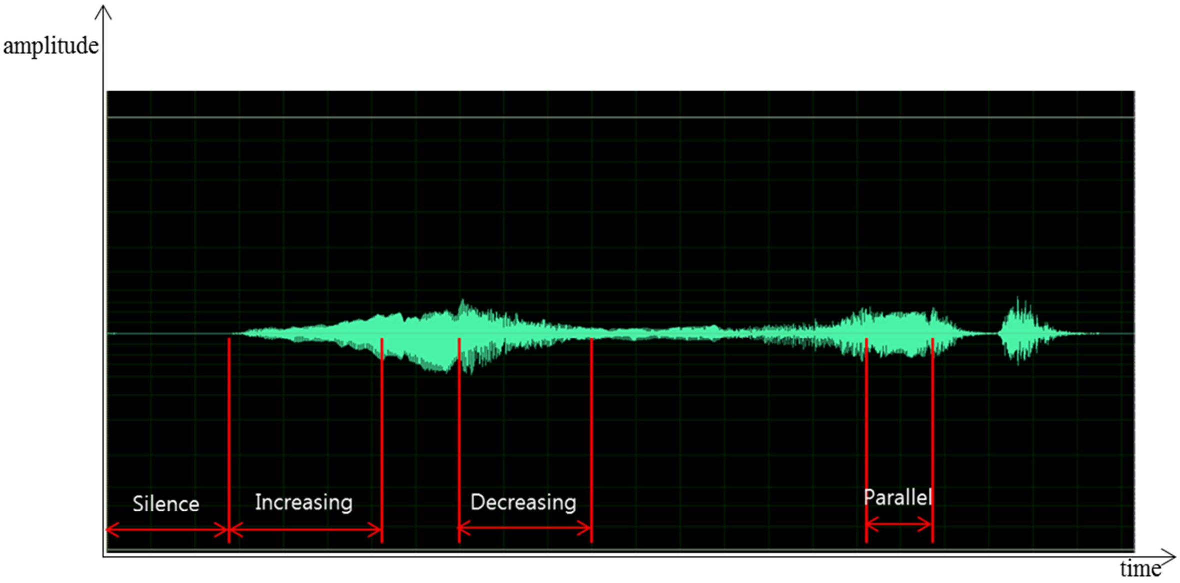


Figure 4. Several phases of the sound of a baby crying

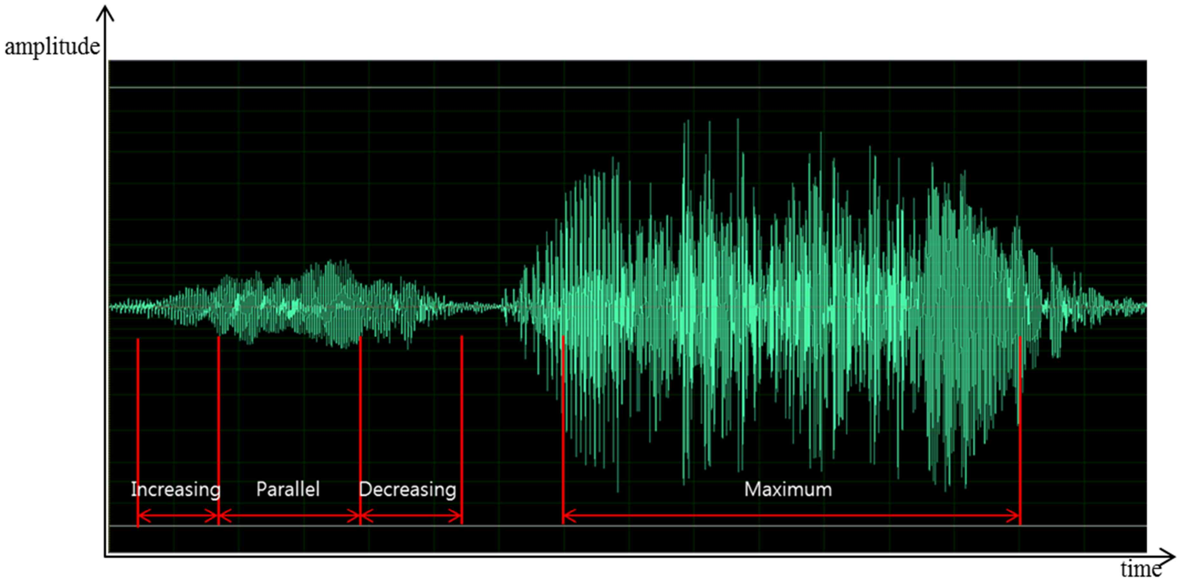


Figure 5. Several phases of the sound of a scream

There are four phases in the baby crying sound. Figure 4 also shows several phases of the sound of a scream. Although Figure 3 cannot present maximum phase, some crying sound have maximum volume of a baby crying phases. These phases are defined as observations in HMM. Every phases are observed from baby crying sound and scream sound. However, number of frequency of observed phases are different and DLC calculates probability with this different characteristic to detect specific sound events. Finally, we can observe that observations of HMM help to solve variability and similarity which are challenges of sound.

3.2.Preprocessing and Feature Extraction

Sound is recorded at a 44.1kHz sampling rate and divided into 100ms frames with 75% overlap. The reason of 100ms frame duration is that DLC system shows the most accurate result when the frame duration is set to 100ms. Also computation time from read sound signal to making decision is the fastest than other frame durations. The detailed result is in the evaluation chapter.

After preprocessing, the system extracts MFCCs, Linear Predictive Coding (LPC), and minimum and maximum amplitude (MinMax) from the sound signal.

The coarse shape of the power spectrum of the sound signal is represented with MFCCs which is popular in acoustic research such as voice recognition. MFCC based technique using cepstrum with a nonlinear frequency domain called mel-frequency [19] was recently demonstrated to provide accurate discriminative performance.

The basic idea of LPC [20] is that a speech signal can be approximated as a linear combination of past speech signals. LPC analysis provides compact representation of vocal tract configuration by relatively simple computation compared to cepstral analysis. The linear prediction method gives us robust, reliable, and accurate values for estimating the parameters that characterize the linear time-varying system representing vocal tract.

The MinMax amplitude feature simply selects minimum and maximum amplitude. MinMax feature also uses to recognize audio recognition [23]. Three kind of feature are usually using in sound event detection area and assure accurate performance. MFCCs, LPC, and MinMax are extracted in each frame and transformed into features to be used as the input data for SVM.

3.3. First layer : SVM

SVM is a discriminative classifier and is based on the concept of decision planes that define decision boundaries. Before classification from input data, SVM have to store the same kind of data. This is called *training* data. In other words, the *training* process is gathering data to recognize patterns. Further, *prediction* means comparing input data with training data for classification. Generally, if training data increase, so too does the accuracy of SVM. The DLC system is implemented in Java and use LIBSVM [7] which is an open source library of SVM.

The role of first layer SVM is classification of observations of HMM. To classify observations accurately, MFCCs, LPC, and MinMax sound features are used and trained more than 40 samples for each observation. First layer is a very important part of the DLC system since the final decision is made based on observations. Using several features and several training processes helps SVM to classify reliable predictions.

3.4. Second Layer : Viterbi Search

For a particular HMM, the Viterbi search [21] is used to find the most likely sequence of states given a sequence of observed states. The Viterbi search exploits the time invariance of the probabilities to reduce the complexity of the problem by avoiding the necessity for examining every route through the trellis. The algorithm keeps a backward pointer for each state after the first time ($t > 1$), and stores a probability (*p*) with each state. The probability *p* is the probability of having reached the state following the path indicated by the back pointers. When the algorithm reaches the states on time, the *p*'s for the final states are the probabilities of following the optimal (most probable) route to that state. Thus selecting the largest, and using the implied route, provides the best answer to the problem.

It is possible that we can select the largest probability when moving to the next state and repeat this progress continuously. So we can think that final state is the most optimal state in HMM. However, that answer is not true and we call that kind of progress is a ‘local search’ or ‘greedy search’. A local search assures fast computation, but it is not the most accurate method. The Viterbi search has to select the biggest probability from the current state and store the calculated value before transition. When the search process reaches the final state, the compute back tracking process for finding the optimal state sequence considers the calculated values stored in the previous selecting process.

Observation	Emission p
Silence	0.1441389
Increasing	0.1000145
Parallel	0.4375003
Decreasing	0.0838247
Maximum	0.2345216

Emission probability of a baby crying

Observation	Emission p
Silence	0.1023126
Increasing	0.080577
Parallel	0.3219344
Decreasing	0.0703060
Maximum	0.4548713

Emission probability of a scream

Observation	Emission p
Silence	0.346271
Increasing	0.201536
Parallel	0.124593
Decreasing	0.189835
Maximum	0.137765

Emission probability of a unknown

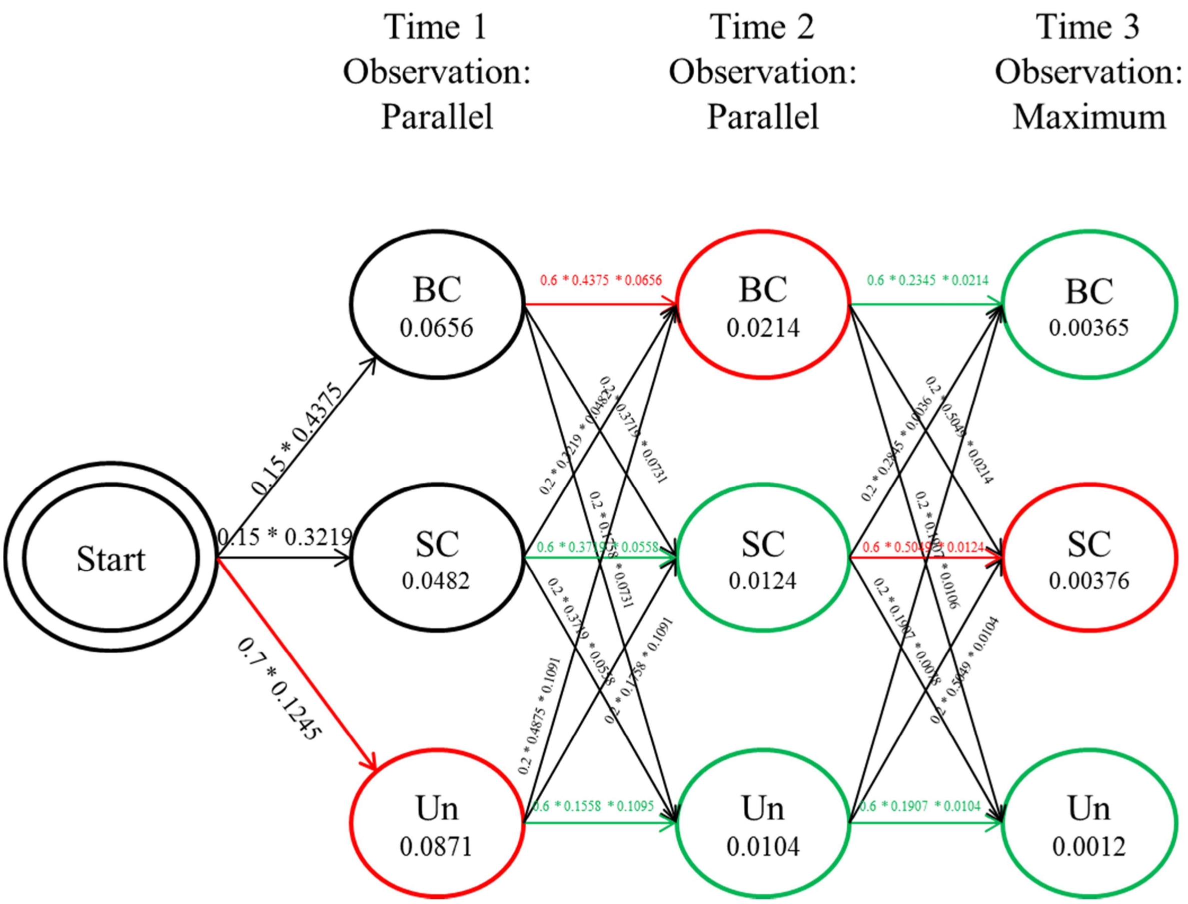


Figure 6. The Viterbi search generates the optimal state sequence from observations. When calculating the probability, the Viterbi search uses the following parameters of HMM: emission probability, transition probability, and start probability.

Figure 6 shows the computation progress of the Viterbi search. When the Viterbi search generates an optimal state sequence, it uses three parameters of HMM: emission probability, transition probability, and start probability. When the Viterbi search calculates the first state, start probability is used with emission probability. After the first stage, start probability is replaced by transition probability. And an additional factor is that the previous state value uses a search algorithm. The Viterbi search generates the current state for each frame from observation and is shown as a red circle in Figure 6. At time 1, parallel observation is detected by first layer SVM and at the second layer the Viterbi search calculates probability with start probability and emission probability.

3.5. Decision Making

The decision making component determines what the current sound event is from DLC. Before making the final decision, it is important to decide when the final decision is made. If the sound event is decided on the end of the sound wave file, it is only useful as a simulation. To apply a sound event detection system in a real environment, the final decision is determined on the reading the sound because sound is always recording by a microphone.

The DLC system has two methods of making a decision:

- Early detection (ED)
- Continuous unknown states

The DLC generates the optimal state sequence from each frame and the system knows how many same states are detected repeatedly. Since an unusual sound such as a baby crying and a scream have to be detected rapidly, the DLC have a mechanism, called ED, for solving real-time issues. If the same state is observed continuously, the system makes a decision earlier.

However, it is possible that unusual events happen for a very short time so that the sound event state is generated but not enough to operate ED. In this case, unknown states will arise after detecting an unusual sound event. If a number of unknown states is detected repeatedly, the DLC system reads the final previous state before changing to the unknown state because the DLC system always records the previous state before transition.

IV. Evaluation

4.1. Performance Comparison

To analyze the DLC system performance, we assume that the DLC system is installed in a home environment and detects unusual events such as a baby crying or a scream. However, we cannot know the performance of the DLC, so anther method of the sound event detection [24] that classifies sound of a scream in home environment is implemented. When the DLC compares the accuracy with [24], use only MFCCs features because [24] does not extract other features such as LPC and MinMax.

[24] approach trains the kinds of sounds which can occur at home such as speech, laugh, crying and so on. They make a decision if the sound of a scream is detected for more than 0.3 seconds. If a scream is detected for less than 0.3 seconds, a particular pronunciations, 'ah', can be regarded as a scream. If the condition is 0.5 seconds, the total accuracy was lower than 0.3 seconds because the sound of a scream of less than 0.5 seconds cannot be detected.

	DLC + MFCCs	**SVM + MFCCs**
No.1 Experiment - Scream	92.9% (26/28)	82.1% (23/28)
No.2 Experiment - Baby crying - Scream	94.7% (54/57)	85.7% (48/57)

Figure 7. Accuracy between the DLC and another approach [24]. The other approach uses MFCCs features and SVM to detect a scream at home. The first experiment purposes to detect only a scream and other noises such as speech, laughing, crying and so on. The second experiment classifies among a baby crying, a scream, and other unknown sounds from the home environment.

To compare the DLC, two sound event detection experiments were performed and results are shown in Figure 7. The first experiment detects only the sound of a scream and the second experiment detects the sound of a baby crying and a scream. The two experiments train several kinds of sounds such as speech, laughing, crying, screams, and silence. Fourty sounds in each category are trained by SVM.

A total of 28 screams were tested to detect sound events using two classifiers: DLC and SVM. The difference of accuracy between the two classifiers is 10.8%. The [24] approach is designed to detect a screaming when a scream is maintained for more than 0.3 second. So if the duration of a scream is less than 0.3 seconds, their approach cannot detect the scream sound event. However, this comparison is unfair because a scream of less than 0.3 seconds can be elusive to make a final decision.

The second experiment can show more certain results to compare between the two classifiers. In the sound of a baby crying, some parts are similar to the sound of a scream and can cause wrong detection. Therefore, the sound of a baby crying and a scream are detected in the second experiment and the same kinds of sounds that are used in the first experiment are trained in the second experiment.

The result of the second experiment shows that the DLC is more accurate than only using SVM and the difference of accuracy is 9%. The DLC is more suitable for detecting between the sound of a baby crying and a scream.

4.2. Accuracy of Sound Features

When we study related sounds, it is very important to select appropriate sound features. MFCCs and LPC are common sound features to detect sound events or recognize speech. Generally, many studies [9], [10], [11], [13], [15], and [16] related with sound use several sound features to increase total accuracy. Figure 8 shows accuracy of sound features.

	Accuracy of the DLC
MFCCs	94.72%(54/47)
MFCCs + LPC + MinMax	96.49%(55/57)

Figure 8. Accuracy for using sound features to detect sound of a baby crying and a scream.

When using MFCCs, LPC, and MinMax sound features, the total accuracy is better than using only MFCCs. Same sounds (speech, a baby crying, scream, laughing and silence) are trained for each category and tested 29 baby cries and 28 screams in two simulations. Figure 8 shows that MFCCs is a powerful sound feature to classify sound events. Suitable sound features can help to increase accuracy of detecting sound events.

4.3. Relationship between the Early Detection and Computation Time

Unusual sounds, a baby crying, a scream, a gunshot, breaking glass, and so on need to be detected as soon as possible. To help the real-time issue, ED has been designed. In a test with three sound features (see Figure 8), 49 of 57 (86%) total test sounds are detected by the ED. Since the detected rate by the ED is high, the reliability of accuracy is problematic so we tested the accuracy of the ED. Among 49 test samples which are detected by the ED, only two sound samples failed the DLC. So we tested again without the ED and just one test sample failed because first layer SVM could not classify it.

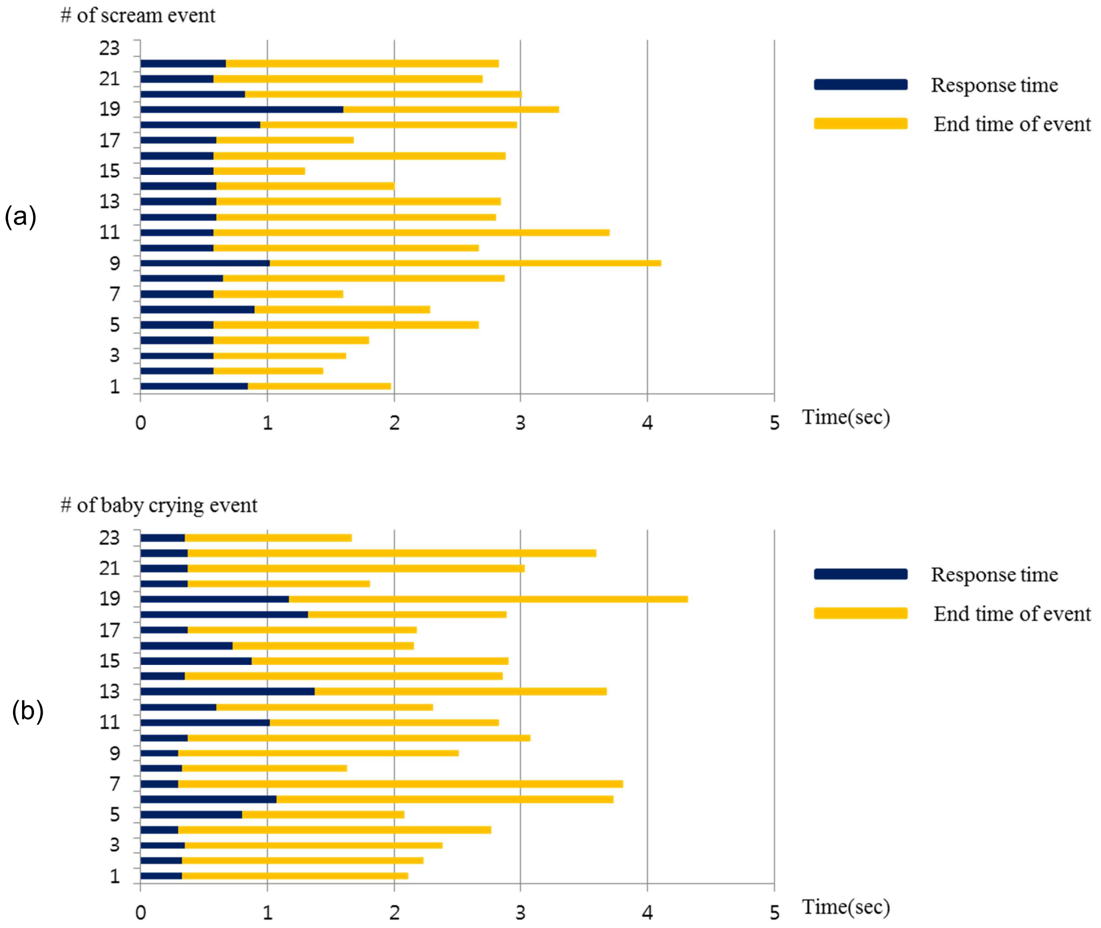


Figure 9. Measurement response time and end time of sound event. The DLC detects sound events early using the ED. (a) shows the average of difference between response time and end time of event is 1.0784 sec in the scream test. Also, (b) shows the average difference of a baby crying is 1.5295 sec.

Figure 9 shows the performance of the ED for solving real-time issues. The response time is the point of the detected sound event by the ED so that the response time is earlier than the end time of the event. The average of difference between the response time and the end time of the event in the test of scream is 1.0784 second and average difference in the baby crying test is 1.5295 seconds. The ED mechanism is very helpful to detect urgent sound events.

4.4. Result of Classification Four Kinds of Unusual Sound Event

Previous experiments classified unusual sound events in homes in only two categories: a baby crying and a scream. However, there are other dangerous situations in a home. In this experiment, we add two unusual sound events: breaking glass and a gunshot. So the DLC system will detect four sound events: a baby crying, a scream, breaking glass, and a gunshot. Also, new observations are added such as breaking moment, splashing glasses, firing, and the echo of shooting. As a result, the number of total observations is nine and they are classified by first layer SVM.

The sound of a breaking glass and a gunshot are finished within a very short time, the ED cannot be operated when using the same condition with a baby crying or a scream. So we adjusted the condition of the ED of breaking glass and a gunshot to increase the occurrence frequency of the ED. A total of 90 sound samples were tested by the DLC and accuracy is 93.3% (84/90) in Figure 10.

Sound	# of test samples	# of successes	# of failures
Screaming	28	27	1
Baby Crying	29	28	1
Breaking Glass	18	16	2
Gunshot	15	13	2
Total Result	90	84	6

Figure 10. Result of detecting four kinds of unusual sound event. Total accuracy is 93.3%(84/90)

4.5. Accuracy of Frame Duration and Computation Time

In general audio processing research, a sound wave is divided as multi frames and this process is called as preprocessing. Before preprocessing, the system has to decide the frame length, overlapping rate, sampling frequency, and so on. Most studies have a different configuration. 20ms of frame length with 50% overlapping is used in [9], [15], and 25ms with 50% overlap is set by [11]. Also 128ms [10], 200ms [18], and 500ms [14] is used as frame length. Frame duration gives more influence to accuracy and computing time in the DLC system. If frame length is lower, the total number of frames decreases and it causes a reduction in execution time but to affects the accuracy of the sound event detection. Next Figure 11 shows the accuracy and computation time for each frame duration.

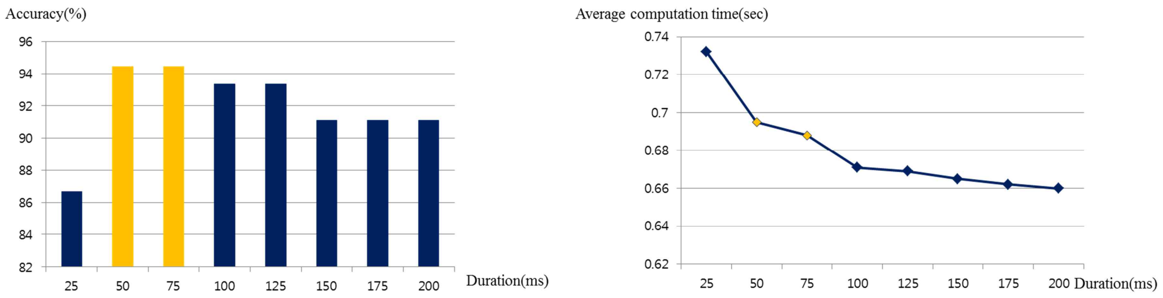


Figure 11. Accuracy and computation time for frame duration. When frame duration is 50ms and 75ms, total accuracy of the DLC is higher than other frame durations. And if frame duration is shorter, computation time is reduced.

A total of eight experiments are tested with different frame durations: 25ms, 50ms, 75ms, 100ms, 125ms, 150ms, 175ms, and 200ms. Every frame is overlapped with 75% and four categories of unusual sounds (a baby crying, a scream, breaking glass, and a gunshot) are trained and tested by the DLC. When frame length is 25ms, first layer SVM could not detect sound phases because 25ms is too short to classify among observations. On the one hand, short frame duration can classify specific observations well such as firing observation which occur within a very short time. So the DLC, which classifies four unusual sounds (a baby crying, a scream, breaking glass, and a gunshot) shows the most accurate performance when using 50ms and 75ms (94.4%).

The computation time is measured based on whole tests result and it means average execution time. When frame duration is longer, the number of computed frames decreases. As a result, the computation time of the DLC becomes lower when using short duration of frame. Therefore, it is more efficient that the DLC, using 75ms duration in the preprocessing because the computation time is faster than 50ms.

4.6. Classification in a Noisy Environment

Sound event detection is very sensitive to noise and other overlapping sounds. Classification of sound events can fail because of very low noise. Reducing noise and distinguishing overlapping sound are important in the sound field.

In this experiment, we test the sensitivity of the DLC system with overlapping sounds and SNR noise. In the first experiment, the DLC detects a baby crying and a scream in three different environments: conversation, washing dishes and street. A total of 55 sound samples were tested by the DLC which succeeded to detect sound event in Figure 8 using MFCCs, LPC, and MinMax sound features.

Overlapped Sound	# of test sample	# of success	# of failure	Accuracy
Conversation	55	48	7	87.27%
Washing dishes	55	31	24	55.36%
Street noise	55	48	7	87.27%

Figure 12. Accuracy with different overlapped sounds.

Among the three sounds, the most sensitive environment is washing dishes. Because of overlapping sound, the first layer SVM could not classify observations of the sound event correctly and made many errors. The results of SVM affect the second layer Viterbi search algorithm. There were too many errors to generate a correct state sequence.

We also added Adaptive White Gaussian Noise (AWGN) to the sound wave file using Matlab and tested it with the DLC. The signal to Noise Rate (SNR) of the AWGN was 20 dB to 75 dB, and we compared the accuracy of the DLC with each SNR. Two experiments were used to compare accuracy because influence of SNR is different for each sound event. In the case of a gunshot, it is affected by SNR the most seriously than other unusual sounds such as a baby crying, a scream and breaking glass. So when SNR is 75, the accuracy of total sound sample is the same when the DLC classifies original sound samples.

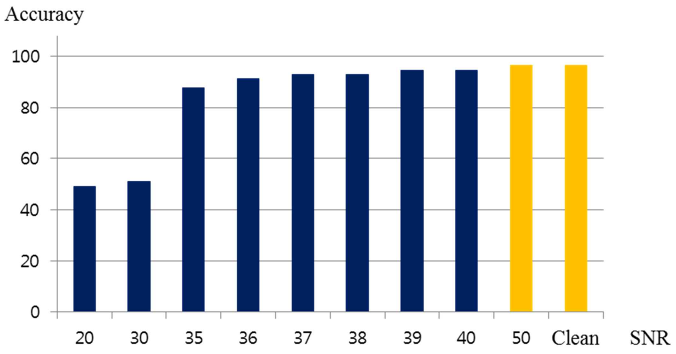


Figure 13. Accuracy with each SNR when detecting two sound events: a baby crying and a scream.

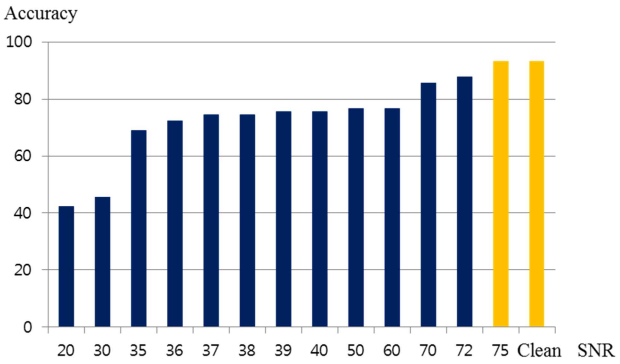


Figure 14. Accuracy with each SNR when detecting four sound events: a baby crying and a scream, breaking glass, and a gunshot.

Figure 13 shows the detection of two sound events with each SNR. When SNR is less than 35 (87.7%), accuracy is unacceptable. If SNR is close to 50, the result of classification is the same (96.5%) as when using the original sound samples. Figure 14 shows four unusual sound events with each SNR. When SNR is close to 75, accuracy is same as test of clean sound sample in contrast with figure 13. The reason of this difference is that sound of a gunshot is more sensitive than other sounds. After firing, an echo sound remains. If AWGN is added in the remaining of, the first layer SVM cannot classify it as a gunshot. So when SNR is set at 75, the accuracy of total result eventually is the same as the original test which used clean sound samples.

V. Conclusion

Sound includes enough information to distinguish an event, so sound is an excellent resource to recognize an event. As the population of people who live alone increases, the recognition of context became a significant issue; sound event detection can help to solve this problem. However, the variability of sound is one of the main challenges and causes incorrect sound event detection. To solve this variability, the DLC is suggested, which is composed of SVM and the Viterbi search. The DLC uses HMM and defines states, observations and HMM parameters to detect a sound event.

Our experiment detected four unusual sounds at home: a baby crying, a scream, breaking glass and a gunshot. The accuracy of detection of four unusual sounds was 94.4% with MFCCs, LPC, and MinMax sound features. Also, we tested the performance of DLC with in a noisy environment and we observed that the performance of DLC decreased. However, the DLC performance is outstanding when using clean sounds compared to SVM which is a popular machine learning algorithm used in related sound works.

Real-time event detection has been emphasized to prevent emergency situations for people who live alone. The ED mechanism helps to solve the quick decision making process. The ED found unusual sound events before the end of event.

The DLC classifier can increase accuracy in sound event detection compared to other classifiers such as SVM and consider real-time detection for emergency situations.

VI. Reference

[1] Administration of Aging, U.S. Department of Health and Human Services. “A Profile of Older Americans : 2011”. 2011.

[2] FBI(Federal Bureau of Investigation). “Crime in the United States, 2011”. 2011.

[3] M. Keally, G. Zhou, G. Xing, J. Wu, and A. Pyles. “PBN : Towards Practical Activity Recognition Using Smartphone-Based Body Sensor Networks”. *SenSys’11*, 2011. pp. 246-259

[4] E. Hoque and J. Stankovic, “AALO : Activity recognition n smart homes using Active Learning in the presence of Overlapped activities”. *Pervasive Computing Technologies for Healthcare, 2012*. pp. 139-146

[5] B. Settle, “Active learning literature survey”. University of Wisconsin-Madison, Computer Sciences Technical Report. 2009.

[6] R. Caroline, C. St-Arnaud, and J. Rousseau, “Video Surveillance for Fall Detection”. Video Surveillance, *InTech*. 2011. pp. 357-382.

[7] C. Chang and C. Lin, “LIBSVM : A library for support vector machines”. *ACM Transactions on Intelligent System and Technology (TIST)*. 2011. Vol. 2, Article No.27.

[8] H. Parmar and B. Sharma, “Control System with Speech Recognition Using MFCC and Euclidian Distance Algorithm”. *International Journal of Engineering Research & Technology (IJERT).* 2013. Vol. 2.

[9] T. Heittola, A. Mesaros, A. Eronen, and T. Virtanen, “Context-dependent sound event detection”. *EURASIP Journal on Audio, Speech, and Music Processing*. 2013.

[10] J.E. Rougui, D. Istrate and W. Souidene, “Audio Sound Event Identification for Distress Situations and Context Awareness”. *IEEE Engineering in Medicine and Biology Society (EMBC)*. 2009. pp. 3501-3504

[11] L.A. Low, N.C. Maddage, M. Lech, L.B. Sheeber, and N.B. Allen, “Detection of Clinical Depression in Adolescent’s Speech During Family Interactions”. *IEEE Transactions on Biomedical Engineering*. 2011. Vol. 58. No. 3. pp. 574-586

[12] J.O. Cavenar, H. Keith, H. Brodie, and R.B. Weiner, *Signs and Symptoms in Psychiatry*, Philadelphia: Lippincott Williams & Wilkins. 1983.

[13] A. Tsanas, M. A. Little, P. E. McSharry, J. Spielmanm, and L. O. Ramig, “Novel Speech Signal Processing Algorithms for High-Accuracy Classification of Parkinson’s Disease”. *IEEE Transactions on Biomedical Engineering*. 2012. Vol. 59. No. 5. pp. 1264-1271

[14] C. Clavel, T. Ehretee, and G. Richard, “EVENT DETECTION FOR AN AUDIO-BASED SURVEILLANCE SYSTEM”. *IEEE ICME*. 2005. pp. 1306-1309.

[15] M. Guo, J. Wang, D. Li, and C. Liu, “Depression Detection Using the Derivative Features of Group Delay and Delta Phase Spectrum”*, IACC*, 2013, pp. 1275-1278

[16] N. Afza, M. Challa, and J. Mungara, “Speech Processing Algorithm for Detection of Parkinson’s Disease”, *International Journal of Engineering Research & Technology (IJERT)*, 2013, Vol. 2.

[17] K. Kipli, M.S. Muhammad, Sh.M.W. Masra, N. Zamhari, K. Lias, and D.A.A. Mat, “Performance of Levenberg-Marquardt Backpropagation for Full Reference Hybrid Image Quality Metrics”, *IMECS.* 2012. Vol. 1.

[18] S. Natlampiras, I. Potamitis, and N. Fakotakis, “On acoustic surveillance of hazardous situations”*, ICASSP,* 2009, pp. 165-168

[19] D.O. Shaughnessy, *Speech Communication : Human and Machine*, Universities Press, 2001.

[20] L.R. Rabiner and R.W. Schafer, *Digital Processing of Speech Signals*, Pearson Education, 1978.

[21] Viterbi AJ, “Error bounds for convolutional codes and an asymptotically optimum decoding algorithm”, *IEEE Transactions on Information Theory 13* ,1967, pp. 260-269

[22] Baum. L. E, Petrie, T. “Statistical Inference for Probabilistic Functions of Finite State Markov Chains”, *The Annals of Mathematical Statistics*, 1966, pp. 1554-1563

[23] Serguei A. Mokhov, “Introducing MARF : A Modular Audio Recognition Framework and its Applications for Scientific and Software Engineering Research”, *Advances in Computer and Information Sciences and engineering*, 2008, pp. 473-478

[24] Weimin Huang, Tuan Kiang Chiew, Haizhou Li, Tian Shing Kok, and Jit Biswas, “Scream Detection for Home Applications”, *IEEE conference on Industrial Electronics and Applicationsis*, 2010, pp. 2115-2120

[25] Man-Wai Mak and Sun-Yuan Kung, “LOW-POWER SVM CLASSFIERS FOR SOUND EVENT CLASSIFICATION ON MOBILE DEVICES”, *Acoustic, Speech and Signal Processing (ICASSP)*, 2012, pp. 1985-1988

요 약 문

소리를 이용하여 현재 상황을 감지하는 DLC 시스템

혼자 사는 사람들의 수가 점차 늘어나고 있으며 독거노인의 경우 미국에서만 2050 년에 현재보다 3 배 더 많아질 것으로 전망되고 있다. 혼자 사는 사람들은 각종 사고와 질병으로 인해 자신에게 급박한 일이 발생할 경우 이에 대한 대처가 늦어질 수 밖에 없으며 이를 방지하기 위해 각 종 센서들을 이용하여 현재 집안에 무슨 일이 발생하고 있는지를 실시간으로 감지해주는 시스템들이 많이 연구되어지고 있다. 그 중 소리를 이용하여 감지할 수 있는데, 소리는 무슨 상황이 일어났는지 알 수 있는 충분한 정보를 포함하고 있으며 마이크와 녹음할 수 있는 저장공간만 있다면 아주 쉽게 구할 수 있기 때문에 각광받고 있다. 소리를 이용하여 사건, 사고를 감지하기 위해서는 아날로그 형태의 소리를 디지털 형태로 가공하여야 하며 가공 후 각 종 알고리즘들을 이용하여 감지를 하는 것이 대표적인 방법이다. 현재 수많은 연구들이 진행되어 왔으며 대부분의 연구들은 아날로그 형태의 소리를 연구의 목적에 맞게 어떻게 가공할 것인가에 대해 초점을 두고 진행되어 왔다. 하지만 이 방법은 소리의 특성 중 하나인 다양함에 의해 한계가 있다. 본 연구에서는 소리의 가공에 초점을 두는 것이 아니라 어떠한 방법으로 감지를 할 것인지 분류하는 알고리즘에 포커스를 두어 진행하였다. HMM 로 모델링을 하고 SVM 과 Viterbi search 알고리즘을 사용함으로써 기존의 일차원적인 알고리즘에 비해 약 10%정도 더 높은 결과를 얻을 수 있었으며 아기 울음소리, 비명소리, 유리 깨지는 소리, 총소리 그리고 기타 집안에서 일어날 수 있는 웃음소리, 대화소리 등을 실험하여 94.4%의 정확도를 얻을 수 있었다.

이상혁