Study on Detecting Emotion from Voice Based on a Bayesian Approach

2012

Jangsik Cho

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Humans communicate with people in their life, which has been recently expanded to include not only people of same cultural background, but also people who have different culture. Human communication involves psychological interactions such as comprehending mutual sentiments, sympathizing with the other person, and enjoying the conversation itself. For psychological interaction, human has to understand emotion from the communication partner exactly, and then express own emotion. Accordingly, understanding emotion is an important intelligence for human communication.

In the field of computer science, detecting emotion of human, for human-computer interaction, has been studied. Detecting emotion technique is applied to communication robot, virtual avatar, call center, car navigation, and so on. This study focuses on detecting emotion for application of the communication robot as described in the following. Firstly, the robots can communicate with human expressively as recognizing emotion, having emotion, and expressing emotion. Secondly, the robots are applied to support for smooth communication such as detecting emotion of people in different cultures. As a first step, this paper proposes detecting emotion method, and compares the sensibilities of emotion recognition between Japanese and Korean using detecting emotion method on each language.

Humans express their emotion by acoustic features, such as level, speed, volume in spoken voice, and thus, voice is relevant to detecting emotion. On the other hand, Bayesian method is a probabilistic reasoning technique for circumstances involving uncertainty, and it is becoming an increasingly important in research and applications of artificial intelligence. This study uses acoustic features of human voice for emotion detection and Bayesian network model as emotion detection method.

This chapter describes related research on detecting emotion from voice, emotion and communication, communication robot, Bayesian network that is proposed as detecting

emotion model, and composition of this paper.

1.1 Related Research on Detecting Emotion from voice

Several researchers have reported methods using acoustic features for detecting emotions from human voice. Shigenaga [1] proposed a method to classify voice samples into five emotions (happiness, disgust, anger, sadness, and neutral). The method covers eight sentences spoken by five participants and is based on the normalization of differences of acoustic features between neutral and each of the other emotions. Shirasawa [2] proposed a method to classify voice samples into six emotions (anger, sadness, happiness, surprise, disgust, and neutral) by using principal component analysis (PCA) on the acoustic features. The method covers eight sentences spoken by nine participants. Kinjou [3] proposed a PCA-based method to classify voice samples into four emotions (anger, sadness, happiness, and neutral) using acoustic features. The method covers twenty-four words spoken by six participants. Moriyama [4] proposed a fuzzy control based method to classify vocally expressed emotions into five emotions (anger, sadness, happiness, fear, and neutral). The voice samples consist of nine words spoken by eight participants. Ververidis [5] proposed a method to classify voice samples into five emotions (anger, happiness, neutral, sadness, and surprise) by using a Bayes classifier. This method detects emotion from specified four participants. These methods all restrict human voice to certain fixed phrases. This restriction would be fatal for human-robot communication. On the other hand, Lee [6] proposed a method to classify voice samples into two emotions (negative, non-negative) by using a linear discriminant classifier and k-nearest neighborhood classifier as combining acoustic, lexical, and discourse. This method had the accuracy rates over 80% from free sentences of unspecified participants in the two emotions. In this study, we devise a practical application for multi-emotion detection that covers free talking voice samples and is speaker independent.

1.2 Emotion and Communication

1.2.1 Emotion

Theories of emotion have been proposed by many researchers. The James-Lange theory [7] stated that emotion is feeling that arises out of physical changes. The two-

factor theory by Schachter [8] stated that the physiological arousal occurs first, and then the individual labels emotion in order to identify the reason behind this arousal. Psychologists distinguished between emotion and mood. They claimed that emotion sustains a mind in the short term, and mood is a mind in the long time. Physiologists argued that emotion is nonconscious mind from a somesthetic sense whereas feeling is conscious mind. From these theories, emotion is short state of mind that nonconsciously reacts to the outside stimulus. That is to say, human emotion arises nonconsciously out of stimuli in life with people.

Basic Emotions

Evolution theory argued that emotion has remained for human evolution because it needs for human subsistence. It means that emotion is innate, and cross-culturally universal. Basic emotions are based on evolution theory. Ekman [9] devised six basic emotions by examining identification of facial expression between isolated cultures. The six basic emotions are happiness, surprise, sadness, anger, fear, and disgust. Additionally, he argued that facial expression has in common between nations. His emotion classification is widely accepted in research of detecting and expressing emotion.

Expression and Extraction of Emotion in Voice

The main channels for emotion expression of human are voice, sentence meaning, face, and gesture. Visual information express emotion by changing eyes, eyebrows and lips from face, acting body and arms. Traditionally, researches of emotion expression have focused on visual information because of perceiving those by a visual sense easily. Sentence meaning also has an obvious emotion expression. Accordingly, it is seldom focused on emotion expression.

Voice expresses emotion from level, speed, and volume, called prosodic properties, which are influenced by a physiological process during pronouncing. For example, someone's voice will become higher, faster, and stronger when he feels angry. That is to say, voice has a connection with emotion expression. Recently, research of emotion expression from voice is becoming more concentration thanks to advance in signal processing technology.

Prosodic properties that express emotion are extracted by acoustic features. Many researches reported that acoustic features strongly influence emotion expression. Typical acoustic features on emotion extraction are fundamental frequency, energy, rate of speech, and formant, and these are widely used for detecting emotion from voice.

1.2.2 Emotion Communication

Animal communication only conveys nothing more than signals for protection of their species, such as hazard warning, feed, courtship, and so on. Conversely, humans create meaningful sentences by combining words of their language, and thus can communicate with people interactively. Interaction in human communication involves not only information share but also psychological sympathy. Needless to say, information share is a main part in human communication. Moreover, psychological sympathy is a unique activity of human communication. Psychological sympathy has the roles in the human communication as follows:

- Interest induction.
  Humans catch useful information from communication. However, sometimes they can not take an interest in communication. Psychological sympathy causes interest in the communication, which helps information share by staying in communication.

- Desire satisfaction.
  Humans live in a social environment that has a relationship with other people. They have a desire that meets empathetic people. Psychological sympathy in the communication satisfies their desire.

For psychological sympathy, humans have several emotional mechanisms such as understanding emotion, having emotion, and expressing emotion. Therefore, emotion is essential for human communication.

Approach of This Study

Recently, human centered robots, aimed at communicating expressively with human, have been developed. To live with people as a partner, robot in communication with human needs to psychological interaction such as detecting emotion, expressing emotion, and having emotion. As a first step, this study focuses on detecting emotion.

Speech, face, and gesture are communication channels of human. Human communication starts almost from speech, and face and gesture support communication. Moreover, consider as an example a telephone, communication is done by only speech without other channels. Speech has verbal and nonverbal properties. Verbal property is effective and unique channel to humans. In conjunction with verbal property, nonverbal properties called acoustic features are also expressed by speech. In Section 1.2.1,

we described availability of acoustic features from voice for detecting emotion. This study uses acoustic features as a evidence for detecting emotion.

1.2.3 Emotion Communication between People of Different Cultural Background

Emotion of human is different in cultures because of influence by cultural background, language, and so on. This section describes the difference of emotion between cultures. Human emotion arises in culture dependence. Mesquita [10] addresses cultural differences in four aspects about emotion generation as follows:

- Differences in antecedent events.
  The differences in prevalent antecedent events underlie difference in emotions. Cultures tend to promote and create events that elicit culturally desirable emotion.

- Differences in experience.
  Culture differences influence experience frequency of emotion. A much higher experience frequency of positive emotion tends to appraise emotion situations as more positive.

- Differences in appraisal.
  Agency attribution is made to the self, a particular other, fate, God, all circumstances together, or nobody in particular. Cultural differences in the frequencies of agency appraisal reflect different patterns of emotions.

- Differences in expression and behavior.
  Cultural differences in the frequencies of expressions and behaviors tend to reflect differences in emotions.

In Section 1.2.1, we described universality of emotion expression from face. However, speech expression of emotion is different in each culture. Take Japan and Korea for example. Japanese speaks Japanese sentence “si-a-wa-se” when he feels happy. Conversely, Korean speaks Korean sentence “hang-bok-ha-da” when he feels happy. Japanese and Korean are different in emotion expression from sentence meaning because they have different phonologies and vocabularies. Prosodic properties are also pronounced on each country peculiarly, which is caused by not only language but also cultural background, such as manner, values, history, and so on. It causes difficulty

of understanding emotion from people of different cultural background, and communication with people who have different culture leads to a problem with psychological sympathy.

Approach of This Study

Advance of internationalization has promoted more trade of media contents and immigration. This trend is becoming more and more increasing chance of psychological sympathy with people of different cultural background. However, understanding emotion from speech is difficult as described in Section 1.2.3. Understanding emotion from verbal information has nothing to do but understands sentence meaning. In contrast to, support robot that understands emotion from voice of different cultural people can help psychological sympathy with them. As a basic study, this study compares the sensibilities of emotion recognition from voice between Japanese and Korean using detecting emotion model on each language.

1.3 Communication Robot

For several past decades, robot technology has been done with distinct success in industrial and manufacturing. Recently, robotics research has been shifting from industrial to domestic application, aimed at communicating with human, have been developed. Its robots have not only communication skill, but also the other functions. This section introduces communication robots. Papero [11] can communicate with human by recognizing about 200 words from human speech, and respond to him. It also reacts to the touch by emotion expression, and has a lot of functions, such as recognizing face, conveying message, quiz, and games. Breazeal [12] developed Kismet that can express various emotions from facial expression, and communicate with human. It creates facial expression by movements of ears, eyebrows, eyelids, and lips. Marina-1 [13] is Internet correspondence robot that handles mobile home health care through communication system.

Our industry-university joint research project has developed a novel robot, Ifbot that can communicate with humans through conversations and facial expressions[14],[15]. The targets of this study are to enable Ifbot to communicate expressively with humans, and support smooth communication between people in different cultures. Section 1.3.1 describes Ifbot in detail.

1.3.1 Ifbot

Ifbot can communicate with humans through conversations and facial expressions[14, 15]. Figure 1.1 shows Ifbot's appearance. With two arms, wheels instead of legs, and an astronaut's helmet, Ifbot is 45 centimeters tall and weighs seven kilograms. It uses voice recognition and synthesis engines to converse with a person. Ifbot can also communicate by showing its emotions through facial expression mechanisms [16] and gestures. The mechanism for controlling Ifbot's emotional facial expressions has 10 motors and 101 LEDs. The motors actuate Ifbot's neck (2 DOFs), both sides of the eyes (2 DOFs for each), and both sides of the eyelids (2 DOFs for each). Ifbot's vision system using CCD cameras can recognize up to 10 persons [17], and the robot has a vocabulary of thousands of words and adapts its conversation to the habits and personalities of different people. Figure 1.2 and figure 1.3 show communication mechanism and facial expression of Ifbot.

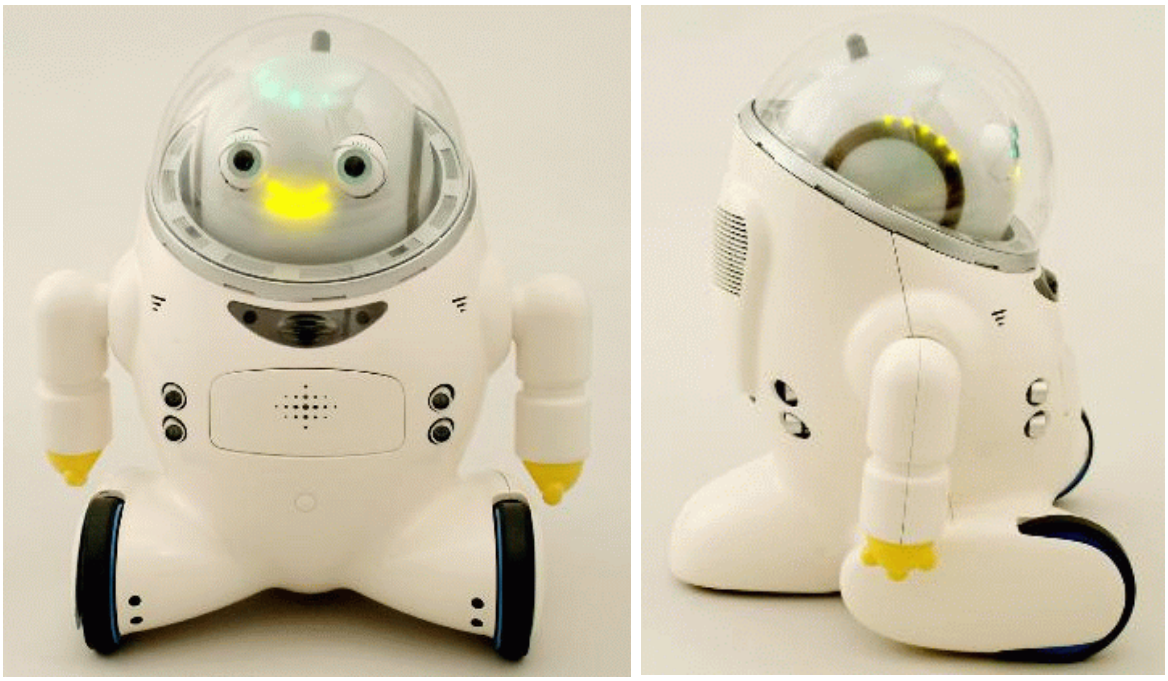


Figure 1.1: Appearance of Ifbot

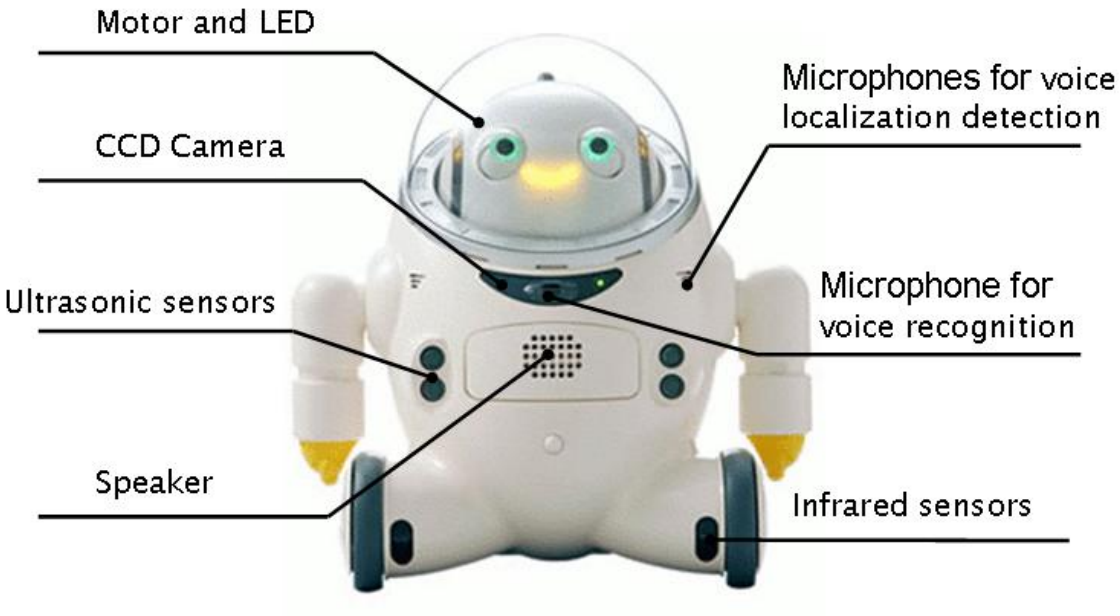
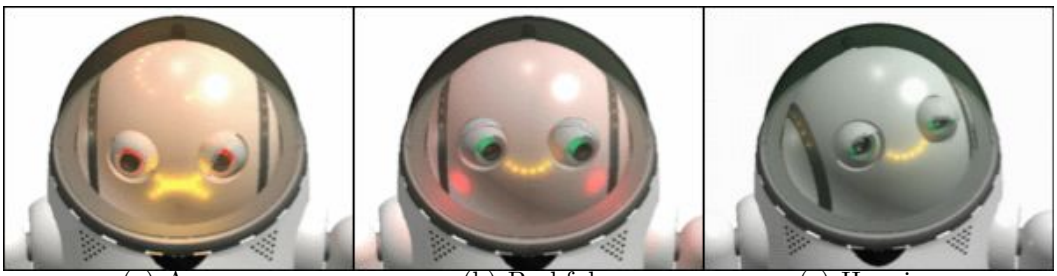


Figure 1.2: Communication mechanism of Ifbot



(a) Anger (b) Bashful (c) Happiness

Figure 1.3: Facial expression of Ifbot

1.4 Detecting Emotion Model: Bayesian Network

This study made a Bayesian network (BN) that is able to acquire the ability to detect emotional content in human voices. BN is one of the eminently practical probabilistic reasoning techniques for reasoning under uncertainty (e.g.,[18, 19]). This section describes BN.

1.4.1 What is Bayesian Network

A BN is a graphical structure that allows us to represent and reason about uncertain domain [20]. The graph structure is constrained to be a directed acyclic graph (or simply dag). A node in a BN represents a set of random variables from the domain. A set of directed arcs (or links) connects pairs of nodes, representing the direct dependencies between variables. Assuming discrete variables, the strength of the relationship between variables is quantified by conditional probability distributions associated with each node.

Most commonly, BNs are representations of joint probability distributions. Consider a BN containing n nodes, X_1 to X_n, taken in that order. A particular value in the joint distribution $P(X_1, ..., X_n)$ is calculated as follows:

$$P(X_1, ..., X_n) = \prod_{i=1}^{n} P(x_i|Pa(X_i)), \tag{1.1}$$

where $Pa(X_i) \subseteq \{X_1, ..., X_{n-1}\}$ is a set of parent nodes of X_i. This equation means that node X_i is dependent on only $Pa(X_i)$ and is conditionally independent of nodes except all nodes preceding X_i.

Table 1.1: The Conditional Probability Table for X_i

$p(X_i) = y_1 \mid Pa(X_i) = x_1$	...	$p(X_i) = y_1 \mid Pa(X_i) = x_m$
.....	...	.....
$p(X_i) = y_n \mid Pa(X_i) = x_1$	...	$p(X_i) = y_n \mid Pa(X_i) = x_m$

Once the topology of the BN is specified, the next step is to quantify the relationships between connected nodes. Assuming discrete variables, this is done by specifying a conditional probability table (CPT). Consider that node X_i has n possible values $y_1, ..., y_n$ and its parent nodes $Pa(X_i)$ have m possible combinations of values $x_1, ..., x_m$. The conditional probability table for X_i is as shown in Table 1.1.

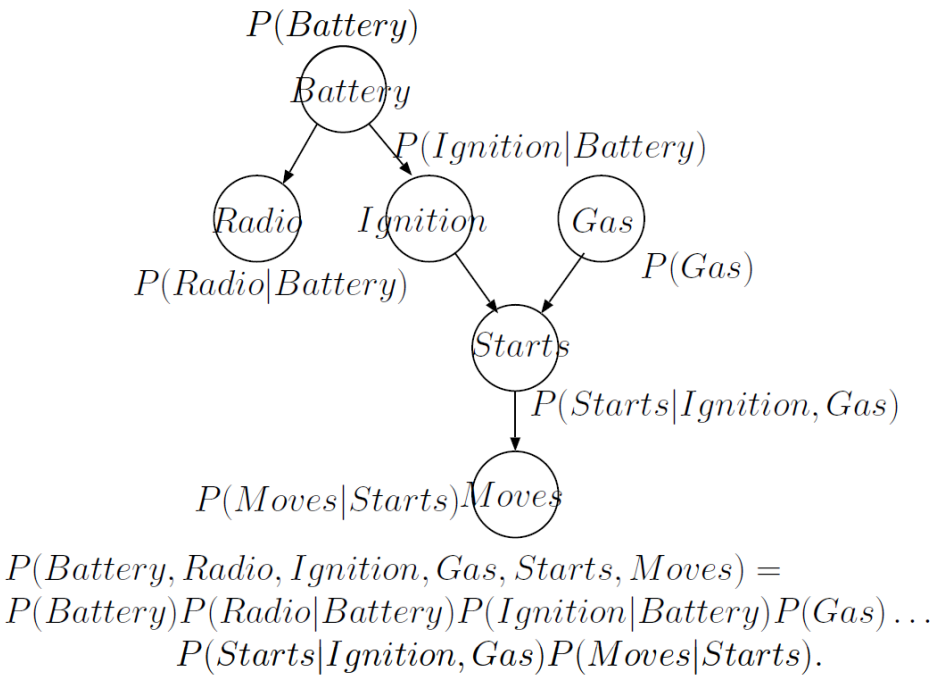


Figure 1.4: An example of BN

Once the topology of the BN and the CPT are given, we can do the probabilistic inference in the BN by computing the posterior probability for a set of query nodes, given values for some evidence nodes. Belief propagation (BP) [21] is a well-known inference algorithm for singly connected BNs, which have a simple network structure called a polytree. In the most general case, the BN structure is a multiply connected network, where at least two nodes are connected by more than one path in the underlying undirected graph. In such networks, the BP algorithm does not work; instead several enhanced algorithms, junction tree [22], logic sampling [23] and loopy BP [24] are used as exact or approximate inference methods. Figure 1.4 shows an example of BN.

1.4.2 Bayes' Theorem

Bayes' theorem is derived from the definition of conditional probability. The probability of event X given event Y is as follows:

$$P(X|Y) = \frac{P(X,Y)}{P(Y)} \tag{1.2}$$

Additionally, the probability of event Y given event X is as follows:

$$P(Y|X) = \frac{P(X,Y)}{P(X)} \tag{1.3}$$

Two equations are combined as follows:

$$P(X,Y) = P(X|Y)P(Y) = P(Y|X)P(X) \tag{1.4}$$

$$P(X|Y) = \frac{P(Y|X)P(X)}{P(Y)} \tag{1.5}$$

- $P(X)$ is prior probability of Y.
- $P(X|Y)$ is probability of X given Y.
- $P(Y|X)$ is probability of Y given X.
- $P(Y)$ is prior probability of Y.

This equation is Bayes' theorem. It is the foundation of the artificial intelligence system based on probability inference.

1.4.3 Conditional Independence

Conditional independence is important for understanding how to work BN. Conditional independence between A and C given B satisfies the following equation.

$$P(A, C|B) = P(A|B)P(C|B) \tag{1.6}$$

The joint distribution of A, B and C given B is as follows:

$$P(A, B, C) = P(A, C|B)P(B) \tag{1.7}$$

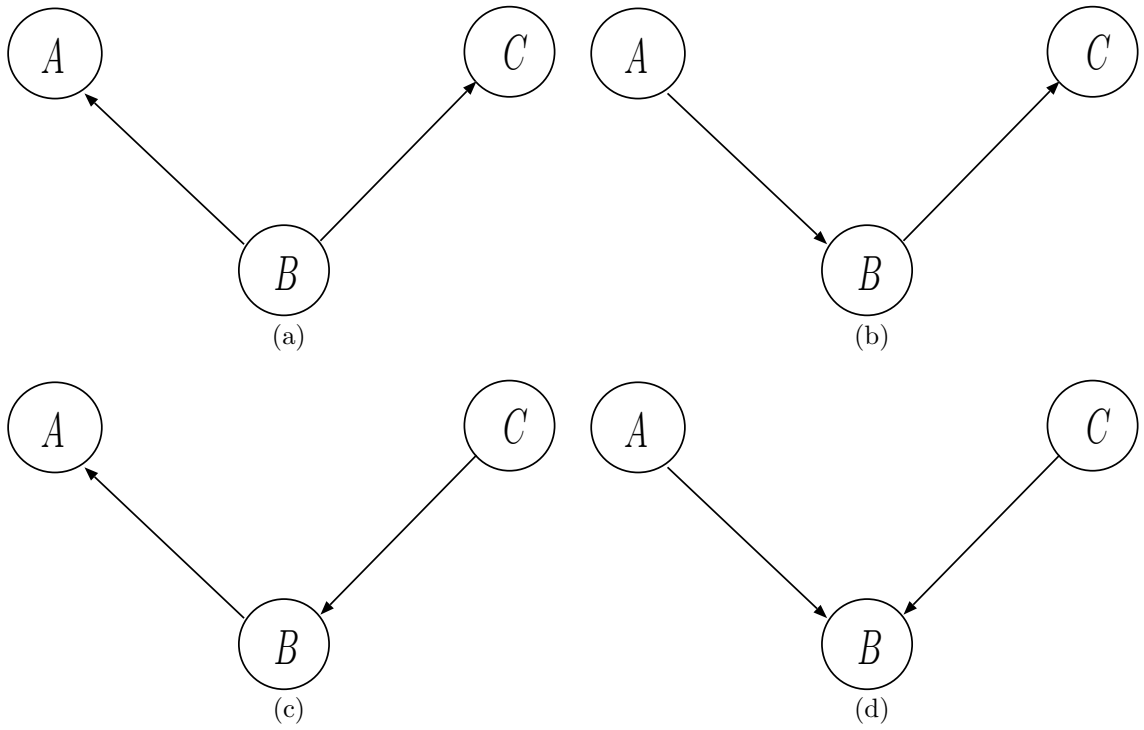


Figure 1.5: Examples of network

Figure 1.5 shows examples of network. The joint distribution of A and C given B in Figure 1.5 (a) is as follows:

$$P(A, B, C) = P(A|B)P(B)P(C|B)$$

$$P(A, C|B) = P(A|B)P(C|B) \tag{1.8}$$

And the joint distribution of A and C given B in Figure 1.5 (b) is as follows:

$$\begin{aligned} P(A, B, C) &= P(A)P(B|A)P(C|B) \\ &= P(A, B)P(C|B) \\ &= P(A|B)P(B)P(C|B) \end{aligned}$$

$$P(A, C|B) = P(A|B)P(C|B) \tag{1.9}$$

And the joint distribution of A and C given B in Figure 1.5 (c) is as follows:

$$\begin{aligned} P(A, B, C) &= P(A|B)P(B|C)P(C) \\ &= P(A|B)P(B, C) \\ &= P(A|B)P(C|B)P(B) \\ &= P(A|B)P(B)P(C|B) \end{aligned}$$

$$P(A, C|B) = P(A|B)P(C|B) \tag{1.10}$$

A and C given B in Figure 1.5 (a)-(c) are conditional independence. However, the joint distribution of A and C given B in Figure 1.5 (d) is as follows:

$$\begin{aligned} P(A, B, C) &= P(A)P(B|A, C)P(C) \\ &= P(A)P(B|A)P(C)P(B|C) \end{aligned}$$

$$P(A, C|B) \neq P(A|B)P(C|B) \tag{1.11}$$

A and C given B in Figure 1.5 (d) is not conditional independence. Node given parent nodes (Pa) is uninfluenced by parent node of that, and node is uninfluenced by unconnected nodes. It means that node given Pa is independence with other nodes except for child nodes of that. Section 1.4.4 shows an example of BN.

1.4.4 Burglar Alarm

Figure 1.6 shows an example of BN called Burglar alarm. Burglar alarm can detect burglary (B) precisely whereas alarm sounds to weak earthquake (E). Two neighbors, John (J) and Mary (M), promised to call the police when anybody hears the alarm (A). John always calls the police when he hears the alarm, however, sometimes he confuses the alarm with the telephone ring. On the other hand, sometimes Mary doesn't hear the alarm because she listens to loud music. Table 1.4.4 shows CPT of Burglar alarm.

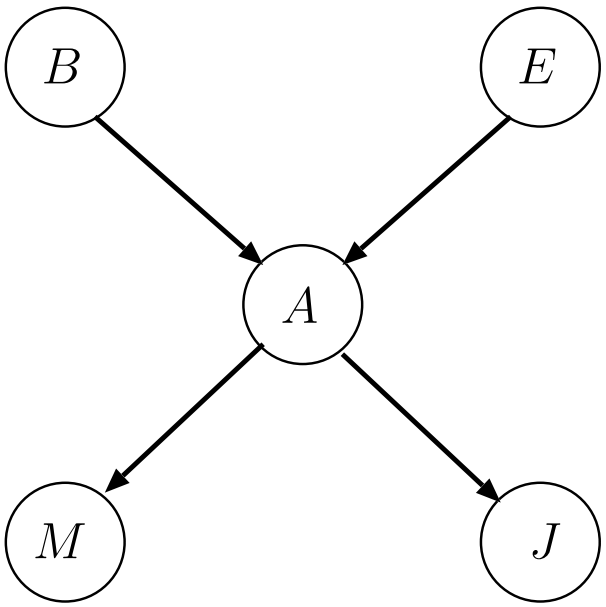


Figure 1.6: Burglar alarm

Table 1.2: Conditional Probability Table of Burglar Alarm

P(B=T)
.001

P(E=T)
.002

A	P(J=T)
T	.90
F	.05

B	E	P(A=T)
T	T	.95
T	F	.94
F	T	.29
F	F	.001

A	P(M=T)
T	.70
F	.01

The joint probability of Figure1.6 is represented as follows:

$$P(B, E, A, J, M) = P(B)P(A|B, E)P(E)P(J|A)P(M|A). \tag{1.12}$$

The probability, the burglar breaks in house when alarm sounds, is calculated as follows:

$$\begin{aligned} P(B = 1|A = 1) &= \frac{P(A = 1, B = 1)}{P(A = 1)} \\ &= \frac{\sum_E \sum_J \sum_M P(B = 1)P(A = 1|B = 1, E)P(E)P(J|A = 1)P(M|A = 1)}{\sum_E \sum_J \sum_M \sum_B P(B)P(A = 1|B, E)P(E)P(J|A = 1)P(M|A = 1)} \\ &= \frac{P(B = 1)\sum_E P(A = 1|B = 1, E)P(E)}{\sum_E \sum_B P(B)P(A = 1|B, E)P(E)} \\ &= 0.374 \end{aligned} \tag{1.13}$$

The probability, the burglar breaks in house when earthquake occurs and alarm sounds, is calculated as follows:

$$\begin{aligned} P(B = 1|A = 1, E = 1) &= \frac{P(A = 1, B = 1, E = 1)}{P(A = 1, E = 1)} \\ &= \frac{\sum_J \sum_M P(B = 1)P(A = 1|B = 1, E = 1)P(E = 1)P(J|A = 1)P(M|A = 1)}{\sum_J \sum_M \sum_B P(B)P(A = 1|B, E = 1)P(E = 1)P(J|A = 1)P(M|A = 1)} \\ &= \frac{P(B = 1)P(A = 1|B = 1, E = 1)}{\sum_B P(B)P(A = 1|B, E = 1)} \\ &= 0.003 \end{aligned} \tag{1.14}$$

Therefore, the probability that the burglar breaks in the house decreases extremely. The occurrence of an earthquake largely influences the inference that the burglar breaks in the house. It means that the earthquake and the burglar are not independence.

1.4.5 K2 Algorithm

K2 algorithm is a method of structure determination by score-based greedy search. It provides a score to fitness evaluation for the BN structure from the data. Initially, each node has no parent. Then, it constructs graphs by adding each parent, and selects a parent node group acquiring the highest score. This performance on one node stops when all of parent nodes are evaluated. This algorithm performs all of nodes, and constructs BN model. K2 algorithm is as follows [25]:

Algorithm 1.4.1 K2 algorithm

procedure K2
Input : a set of n node, an ordering on the nodes, an upper bound u on the number of parents a node may have, and a database D containing m cases.
Output : for each node, a printout of the parents of the node.
for $i := 1$ to n **do**
  $\pi := \phi$;
  $P_{old} := f(i, \pi_i)$
  OKToProceed := **true**;
  while OKToProceed and $|\pi_i| < u$ **do**
    let z be the node in $\mathbf{Pred}(x_i) - \pi_i$ that maximizes $f(i, \pi_i \cup \{z\})$;
    $P_{new} := f(i, \pi_i \cup \{z\})$;
    if $P_{new} > P_{old}$ **then**
      $P_{old} := P_{new}$;
      $\pi_i := \pi_i \cup \{z\}$;
    else
      OKToProceed := **false**;
    end if
  end while;
  write('$Node$: ',x_i,' Parents of x_i:',π_i);
end for;
end K2;

$f(i, \pi_i)$ is score function, π_i is set of parents node of x_i, $\mathbf{Pred}(x_i)$ is parent nodes of previous step.

This algorithm is widely accepted in BN construction. However, it requires reasonable node order. In Chapter 2 and Chapter 4, we use K2 algorithm as construction

method of BN.

1.4.6 Bayesian Network Classifiers

BN is a strong method for representation of uncertain circumstances. Additionally, BN has been also studied as classifiers. This section describes a simple Bayesian classifier called Naive Bayes firstly, and then describes Tree Augmented Naive Bayes that is one of Bayesian network classifier.

Naive Bayes

Naive Bayes (NB) is a simple Bayesian classifier with strong independence assumption of attributes. Class node connects to all of attribute nodes, and attribute nodes have no connection to the other attribute nodes. Figure 1.7 shows an example of BN. Classification from attributes $x = a_1 \cdots a_n$ is conducted as follows:

$$\begin{aligned} C &= \underset{c_i \in C}{\operatorname{argmax}} P(c_i|x) \\ &= \underset{c_i \in C}{\operatorname{argmax}} P(c_i)P(a_1 a_2 \cdots a_n|c_i) \\ &= \underset{c_i \in C}{\operatorname{argmax}} P(c_i)P(a_1|c_i)P(a_2|c_i)\cdots P(a_n|c_i) \\ &= \underset{c_i \in C}{\operatorname{argmax}} P(c_i)\prod_{j=1}^{n} P(a_j|c_i) \end{aligned} \tag{1.15}$$

It has a high classification performance. However, it is unrealistic for strong independence assumption of the attributes.

Tree Augmented Naive Bayes

Tree Augmented Naive Bayes (TAN) [26, 27, 28] is a BN classifier that constructs a tree structure among the attributes in BN. TAN reduces independence of the attributes for constructing the tree structure, and it keeps on classification efficiency. Conditional mutual information between the attributes is used for constructing a tree structure. Conditional mutual information is defined as follows [26]:

$$I_P(\mathbf{X};\mathbf{Y}|\mathbf{Z}) = \sum_{\mathbf{x},\mathbf{y},\mathbf{z}} P(\mathbf{x},\mathbf{y},\mathbf{z}) \log \frac{P(\mathbf{x},\mathbf{y}|\mathbf{z})}{P(\mathbf{x}|\mathbf{z})P(\mathbf{y}|\mathbf{z})} \tag{1.16}$$

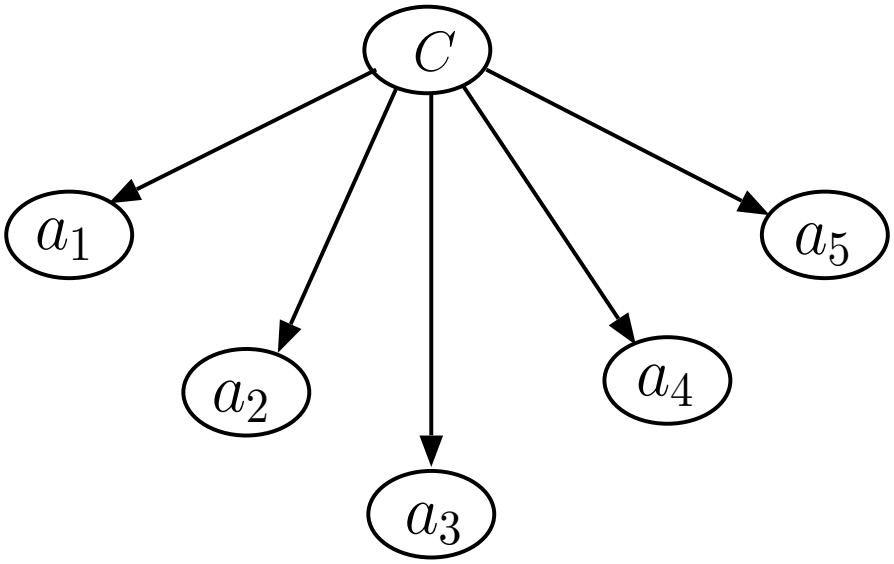


Figure 1.7: An example of NB

The tree structure is constructed as follows:

1. Compute $I_{\hat{P}_D}(a_i; a_j|c)$ between each pair of attributes, $i \neq j$.
2. Build a complete undirected graph in which nodes are attributes $a_1, \cdots, a_n$. Annotate the weight of an edge connecting a_i to a_j by $I_{\hat{P}_D}(a_i; a_j|c)$.
3. Build a maximum weighted spanning tree.
4. Transform the resulting undirected tree into a directed one by choosing a root variable and setting the direction of all edges to be outward from it.
5. Construct a TAN model by adding a vertex labeled by c and adding an arc from c to each a_i.

All attribute nodes except the root node of the tree can have only one parent from another attribute node. Class C is classified as follows:

$$C = \underset{c_i \in c}{\operatorname{argmax}} P(c_i) \prod_{j=1}^{n} P(a_j|c_i, Pa(a_j)) \tag{1.17}$$

$Pa(a_j)$ is a parent node of a_j but not of a class node. Figure 1.8 shows an example of TAN. In Chapter 3, we use TAN as detecting emotion model.

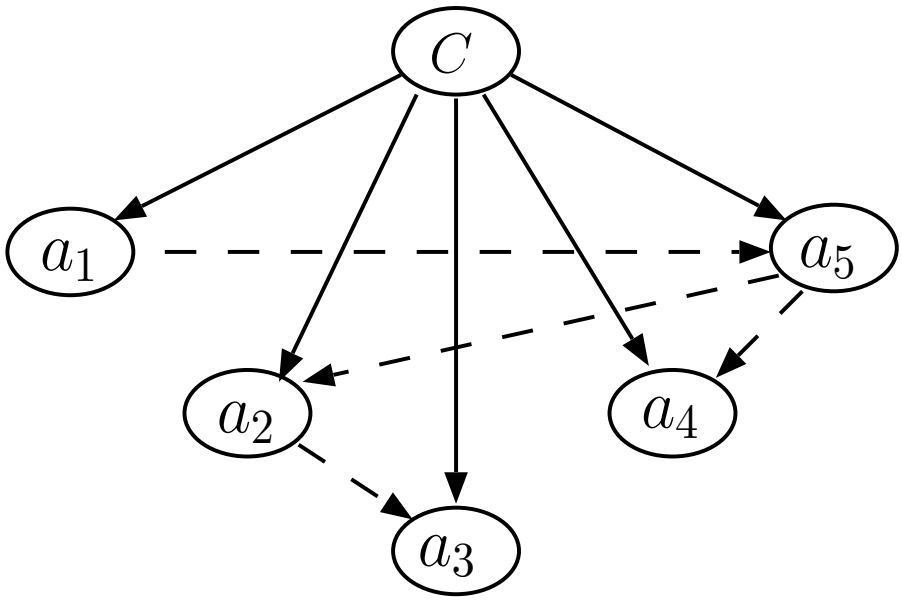


Figure 1.8: An example of TAN

1.5 Composition of This Paper

Detecting emotion mechanism of a robot is important for communication with human. This paper proposes Bayesian approach for detecting emotion from acoustic feature of voice, and detects emotion from native and foreign language.

In Chapter 2, we propose emotion detection methods from voice. The methods are singleton and biphase BN by using K2 algorithm.

In Chapter 3, we propose pairwise classification for emotion detection from voice. The method uses a series of binary TAN classifiers with feature selection on each pair of emotions.

In Chapter 4, we compare sensibilities of detecting emotion from voice between Japanese and Korean using each BN model.

In Chapter 5, we describe a conclusion.

Chapter 2

Detecting Emotion Method: BN using K2 Algorithm

2.1 Introduction

For several decades, robot technology has had distinct success in manufacturing. Robotics research has shifted from industrial to domestic applications, and several domestic robots, for communicating expressively with humans, have been developed (e.g.,[12, 29, 11, 13, 30, 31, 32]). This research field is called human centered robotics. Its robots must be not only safe, autonomous, and intelligent, but also able to help their human partners and to make him or her enjoy. We believe that these requirements need a smooth communication mechanism involving psychological interaction.

Communication involves not only conveying messages or instructions but also psychological interactions. Humans can communicate expressively as understanding emotion, having emotions, expressing emotion. To communicate in this way, a robot requires several mechanisms to make up for its lack of human intelligence.

The target of our study is to develop fundamental technology to recognize, control, and express the emotional content within robot-human interactions [33, 34, 35]. As a first step, this study focuses on detection of emotion. Humans can understand emotion by acquiring knowledge from the expressive features, such as, level, speed and volume of his voice, gesture and face of him. For emotional communication robots, we use the acoustic features that can precisely detect the emotions expressed in voice data.

In this chapter, we propose a method for emotional communication robots which detect emotion. The method uses BN in which the emotional content of voice are modeled by its acoustic features. We made a BN using K2 algorithm, which was able to acquire the ability to detect emotional content in human voices.

Here, we report on experiments on reasoning with emotion inference with complete and incomplete features, and discuss how the relationship between certain components of acoustic features and certain emotions affects reasoning performance.

2.2 Constructing Detecting Emotion Engine

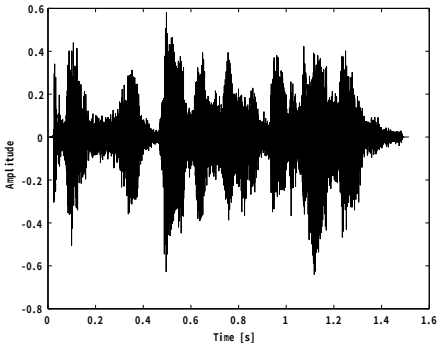
In this study, we focus on the acoustic features of the voice as a cue to what emotion expresses. This section describes a BN modeling for this problem.

2.2.1 Voice Data

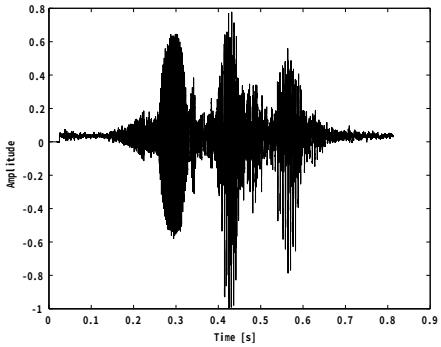
To construct the emotion detection engine, the voice data that have expressions of emotions are necessary as a learning data. A researcher collected segments of Japanese voice samples that were spoken by unspecified actors and actresses from free utterances in films, TV dramas, and so on. The segments we collected do not have loud noises. In this study, we supposed that the segments have a single emotion. All segments are labeled with one label from the six emotions (anger, sadness, disgust, fear, surprise, or happiness) by a researcher. The appropriateness of the emotional labels is confirmed through subjective evaluation experiments. Figure 2.1 shows examples of a voice waveforms on each emotion. These were spoken by different actors from free utterances.

2.2.2 Features Extraction

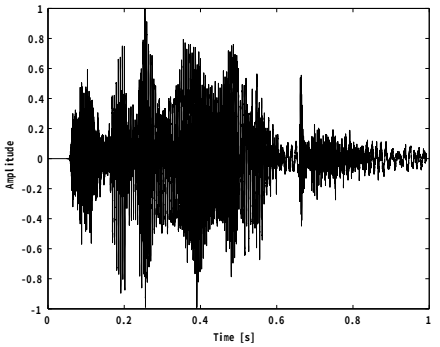
Voice has three components: prosody, tone, and phoneme. It became obvious from reviewing past research that the prosodic component is the most relevant to emotional expressions [36, 37]. As attributes of voice data, we chose three acoustic attributes: energy, fundamental frequency and duration as the acoustic parameters for BN modeling. Acoustic analysis was done on 11 ms frames passed through a Hamming window extracted from voice waveforms sampled at 22.05 kHz. The attributes of energy, maximum energy (PW_{MAX}), minimum energy (PW_{MIN}), mean energy (PW_{MEAN}) and its standard deviation (PW_S) are determined from the energy contours for the frames in the voice waveform. The attributes of fundamental frequency, maximum pitch ($F0_{MAX}$), minimum pitch ($F0_{MIN}$), mean pitch ($F0_{MEAN}$) and its standard deviation ($F0_S$) are determined from short time Fourier transforms for the frames in a voice waveform. As the attribute concerning duration, we measure the duration per a single mora (Tm). Then we added the attribute of the utterer's sexuality (SE).



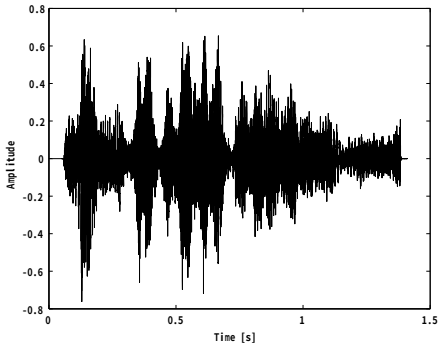
(a) Anger voice of Japanese sentence ze-tta-i-bo-ku-ni-wa-de-ki-ma-sen



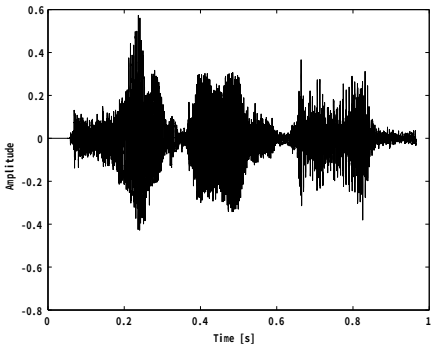
(b) Sadness voice of Japanese sentence u-so-da-ro



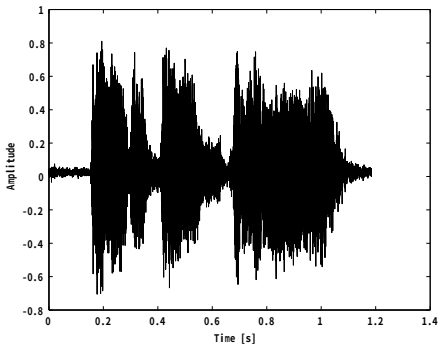
(c) Disgust voice of Japanese sentence so-re-de-i-i-no-ka



(d) Fear voice of Japanese sentence da-i-zyo-u-bu-de-syo-u-ka-sen-se-i



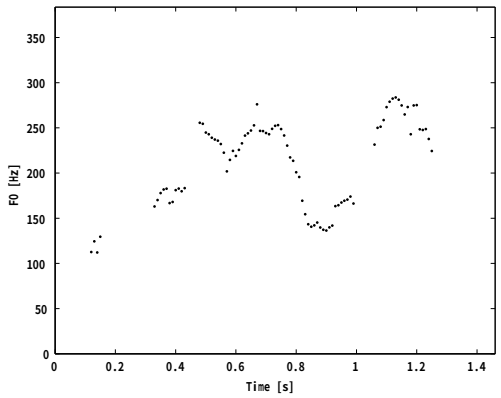
(e) Surprise voice of Japanese sentence hon-to-u-de-su-ka



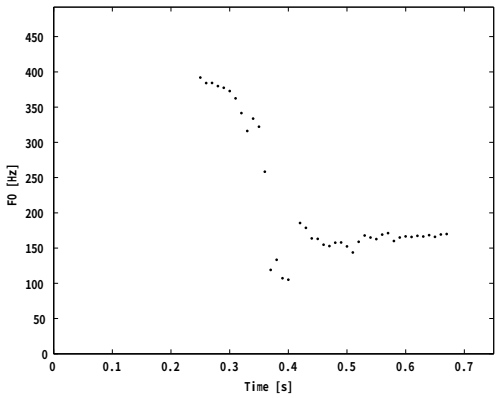
(f) Happiness voice of Japanese sentence a-ri-ma-si-ta

Figure 2.1: Examples of voice waveforms on each emotion

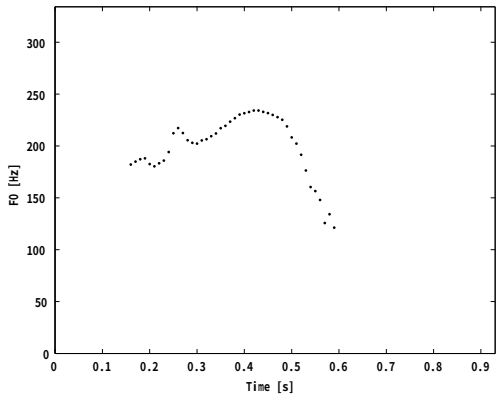
The goal attribute ($EMOT$) and the above nine prosodic feature values and utterer's sexuality (total eleven attributes) were assigned to the nodes of the BN model. Figure 2.2 shows plots of fundamental frequencies for voice samples shown in Figure 2.1, and Figure 2.3 shows plots of energies for those.



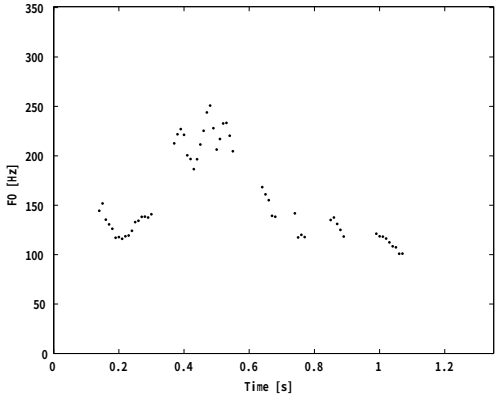
(a) Anger voice



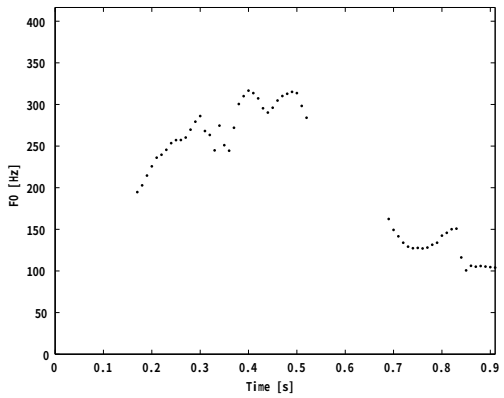
(b) Sadness voice



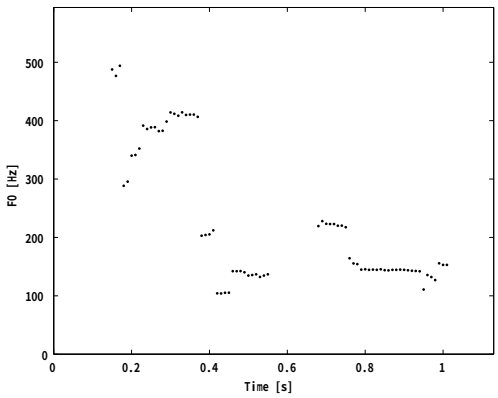
(c) Disgust voice



(d) Fear voice

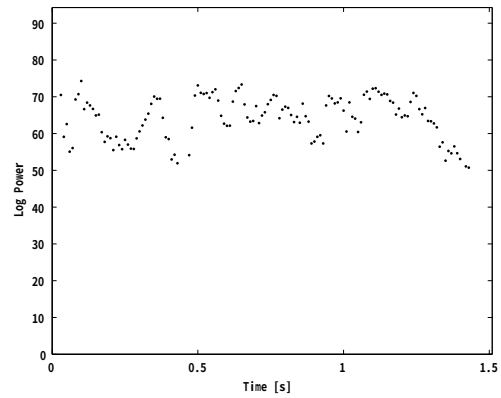


(e) Surprise voice

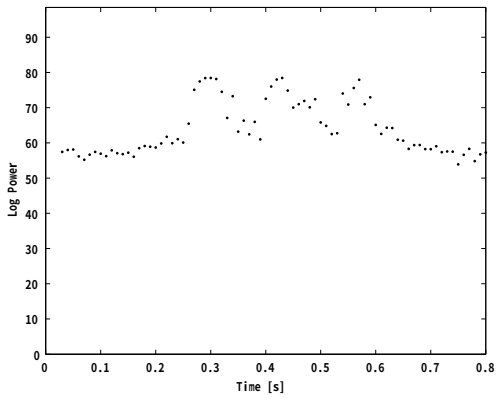


(f) Happiness voice

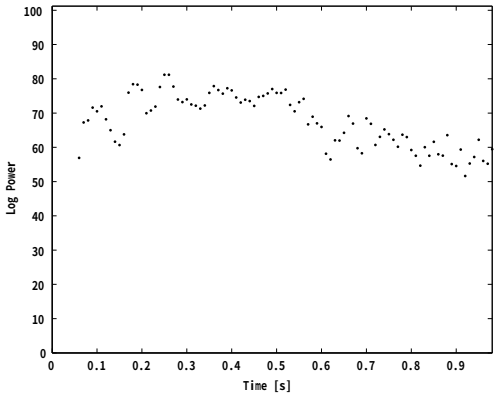
Figure 2.2: Plots of fundamental frequencies



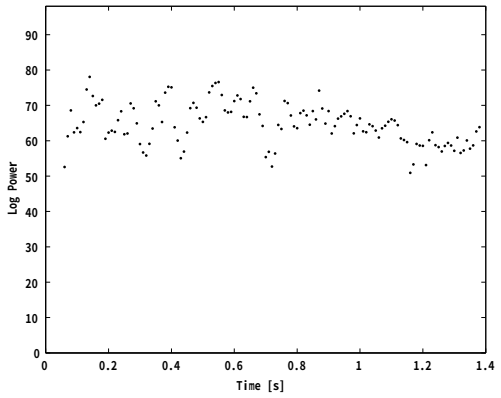
(a) Anger voice



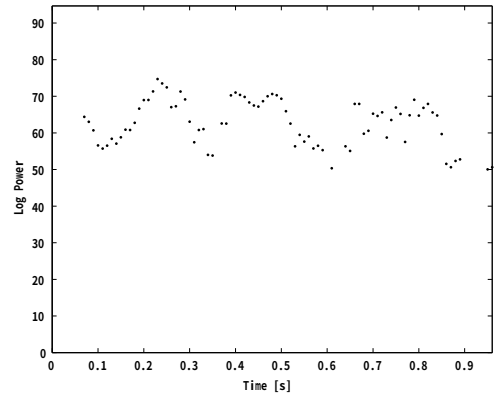
(b) Sadness voice



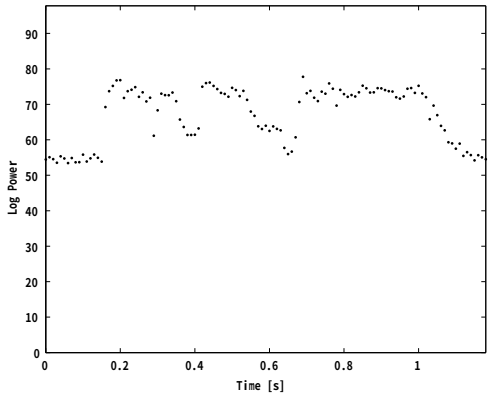
(c) Disgust voice



(d) Fear voice



(e) Surprise voice



(f) Happiness voice

Figure 2.3: Plots of energies

Table 2.1 - Table 2.3 show those feature values and feature averages of male's voices, and describe results that compare feature values with averages.

Table 2.1: Feature Values of Fundamental Frequency

Emotion of sample	$F0_S$	$F0_{MEAN}$	$F0_{MAX}$	$F0_{MIN}$
Anger	61.22814 high	197.5978 high	283.6697 low	67.42994 low
Sadness	91.99153 high	210.6101 high	391.9754 high	97.24525 high
Disgust	58.163 high	174.0668 low	234.2298 low	56.86691 low
Fear	46.54205 low	151.0857 low	250.8503 low	84.56413 low
Surprise	76.22251 high	215.8657 high	316.5243 high	100.6357 high
Happiness	116.9654 high	229.5691 high	493.8726 high	103.8682 high
Average	58.00001	193.9013	299.7783	89.92694

Table 2.2: Feature Values of Energy

Emotion of sample	PW_S	PW_{MEAN}	PW_{MAX}	PW_{MIN}
Anger	8.232096 high	62.6451 low	74.26471 low	21.87752 low
Sadness	7.443669 low	64.05325 low	78.45199 high	53.89079 high
Disgust	7.615725 high	67.53902 high	81.19223 high	51.64077 high
Fear	5.793229 low	64.69286 low	78.06022 low	50.9255 high
Surprise	7.288706 low	61.23113 low	74.69947 low	47.58789 high
Happiness	7.823879 high	67.08508 high	77.76905 low	53.45068 high
Average	7.607529	65.89302	78.19982	40.37269

Table 2.3: Feature Value of Duration Rates

Emotion of sample	Tm
Anger	0.0842 low
Sadness	0.162639 high
Disgust	0.141879 low
Fear	0.108547 low
Surprise	0.161172 high
Happiness	0.236943 high
Average	0.15251

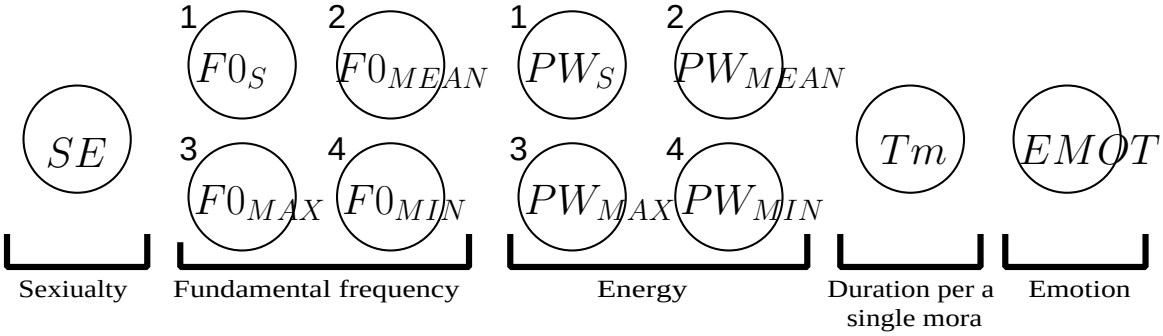
Take a happiness voice sample for example. Its features of fundamental frequency, energy, and duration rate have higher values than averages except for maximum of energy. From these values, we confirmed that happiness voice sample is spoken by fluctuant and high pitch and energy, slow speech.

2.2.3 Discretization of Features

The section describes the discretization of extracted acoustic features. We considered BN with discrete and multinomial variables only. In order to learn the discrete causal structure of the BN model, all acoustic features were converted into discrete values. The thresholds to discretize continuous values were determined from the distribution of the acoustic features extracted from the training voice samples. Discrete values of acoustic features are shown in Table 2.4.

Table 2.4: Samples of Training Data

SE	$F0_S$	$F0_{MEAN}$	$F0_{MAX}$	$F0_{MIN}$	PW_S	PW_{MEAN}	PW_{MAX}	PW_{MIN}	Tm	$EMOT$
F	4	4	4	2	5	5	4	5	2	Anger
M	1	1	1	1	2	3	5	4	2	Sadness
M	3	1	1	4	5	1	1	3	5	Disgust
F	5	5	4	5	2	2	2	4	4	Fear
F	2	4	5	4	2	5	5	4	3	Surprise
M	1	5	5	3	3	5	5	1	5	Happiness
M	3	4	3	1	3	3	2	2	1	Anger
F	2	1	3	2	5	4	5	5	2	Sadness
F	2	3	4	3	1	2	3	3	4	Disgust
M	1	2	1	2	3	1	3	3	4	Fear
M	1	1	1	3	3	1	3	4	5	Surprise
F	4	4	4	5	4	4	4	1	4	Happiness
...	...	...	...	...	...	...	...	...	...	...



1 $SE \prec F0group \prec PWgroup \prec Tm \prec EMOT$
2 $SE \prec F0group \prec Tm \prec PWgroup \prec EMOT$
3 $SE \prec PWgroup \prec F0group \prec Tm \prec EMOT$
4 $SE \prec PWgroup \prec Tm \prec F0group \prec EMOT$
5 $SE \prec Tm \prec F0group \prec PWgroup \prec EMOT$
6 $SE \prec Tm \prec PWgroup \prec F0group \prec EMOT$

Figure 2.4: Groups of nodes and the variable orders of node groups

2.2.4 Learning BN Structure

The section describes how to specify the topology of the BN model for emotion detection and to parameterize CPT for connected nodes. The emotion detection BN modeling is to determine the qualitative and quantitative relationships between the output node containing the goal attribute (emotions) and nodes containing acoustic features. We chose a model selection method based on the Bayesian information criterion (BIC) [38], which has information theoretical validity and is able to learn a high prediction accuracy model through avoidance of over-fitting to training data.

Let M be a BN model, θ_M be a parameter representing M, and d be the number of parameters. M is evaluated by the BIC of M, defined as

$$BIC(\hat{\theta_M}, d) = -2 \log P(D \,|\, \hat{\theta_M}) + d \log N. \tag{2.1}$$

where D is training samples, and $P(D \,|\, \theta_M)$ is the likelihood of D given θ_M; $\hat{\theta_M}$ is the parameter representing M giving the maximum likelihood (ML) estimate; and N is the number of the samples. If D is partially observed, expectation maximization (EM) algorithm [39] is utilized for estimating θ_M asymptotically with incomplete data in the training samples.

As the knowledge for emotion detection, we made a BN model that maximizes BIC

on voice data. We used K2 [25, 19] as the search algorithm. K2 needs a pre-selected variable order. We thus considered every possible permutation of three node groups: PW, $F0$ and Tm, such that node SE preceded all others shown in Figure 2.4.

2.3 Algorithm of Emotion Inference

Probabilistic reasoning in BN can make rational decisions even if some of the evidence lacks direct observation. Considering our purpose of developing a practical application of BNs for human-robot interaction, the inference algorithm should be able to handle evidence that has uncertainty associated with it, because robots often can not recognize the voice features. The topology of the BN is often multiply connected when there is a complicated relationship between variables. We chose junction tree [22] as the inference algorithm with BNs, it is an exact inference algorithm in multiply connected BNs. It is efficient clustering inference algorithms. Clustering inference algorithms transform the BN into a probabilistically equivalent polytree by merging nodes and removing multiple paths between two nodes along which evidence may travel.

2.4 Detecting Emotion using Singleton BN

Figure 2.5 is general of the emotion detection system using singleton BN. This section describes an experiment on emotion detection using singleton BN. First, we collected 1600 segments of voice waveforms and labeled them with six emotions, as described in Section 2.2.1. We then extracted nine acoustic features and the utterer's sexuality from each of the segments and assigned them to the attributes, as described in Section 2.2.2. The acoustic analysis used the Snack sound toolkit [40]. We then randomly selected 1400 samples as training data and discretized their attributes into five values. We determined the threshold for discretization on the basis of the idea of even-sized chunks; that is, each discrete value covers 20% of the training data. Table 2.4 shows brief excerpts of the training data. We then modeled the BNs with changing six variable orders by Bayes Net Toolbox [41]. Figure 2.6 shows results with the variable order $SE \prec F0 \prec PW \prec Tm \prec EMOT$. The parent nodes of $EMOT$ are $F0_{MEAN}, PW_S, T_m$ that strongly influence emotion inference. This model conducts emotion detection by probabilistic inference from conditional probability table (CPT) of $EMOT$ node shown in Table 2.5. To confirm feature values for detection of each emotion, we itemized the number of inference conditions of emotions on discretized

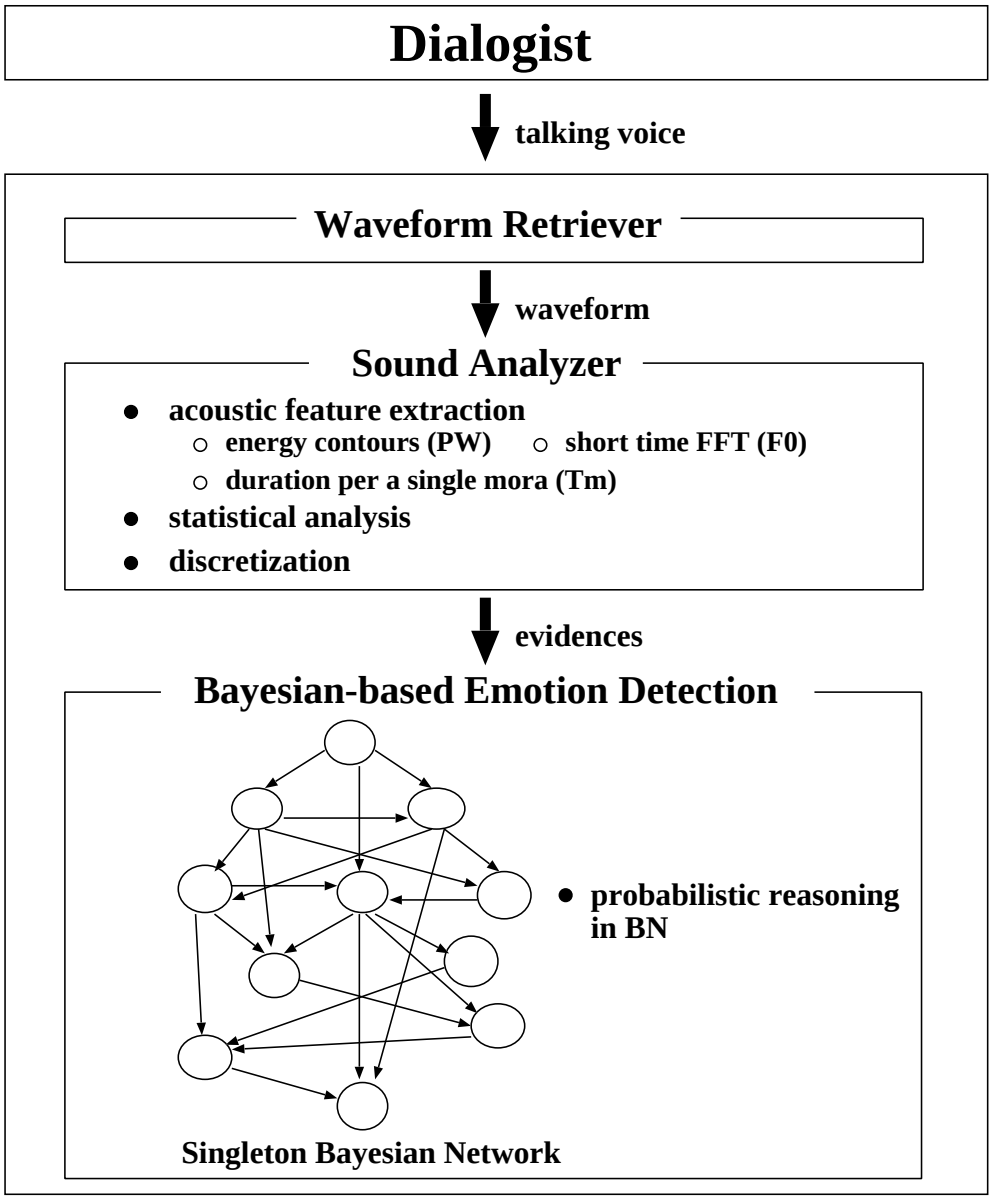


Figure 2.5: General of the emotion detection system using singleton BN

feature values for parent nodes of $EMOT$. Table 2.6 shows the results. Feature values that are included in condition 2 (discrete value is 3) are close to an average. BN model probably infers voice as sadness, disgust, or fear when $F0_{MEAN}$ is low. It probably distinguish these emotions by PW_S and Tm: voice is inferred as sadness when PW_S is high; low Tm causes inference as disgust; voice is inferred as fear when PW_S is low. It probably infers voice as surprise or happiness when PW_S is little high and Tm is high. It probably also infers voice as anger when PW_S is high and Tm is low. Therefore, these results correspond largely to comparison results with averages shown in Table 2.1 - Table 2.3.

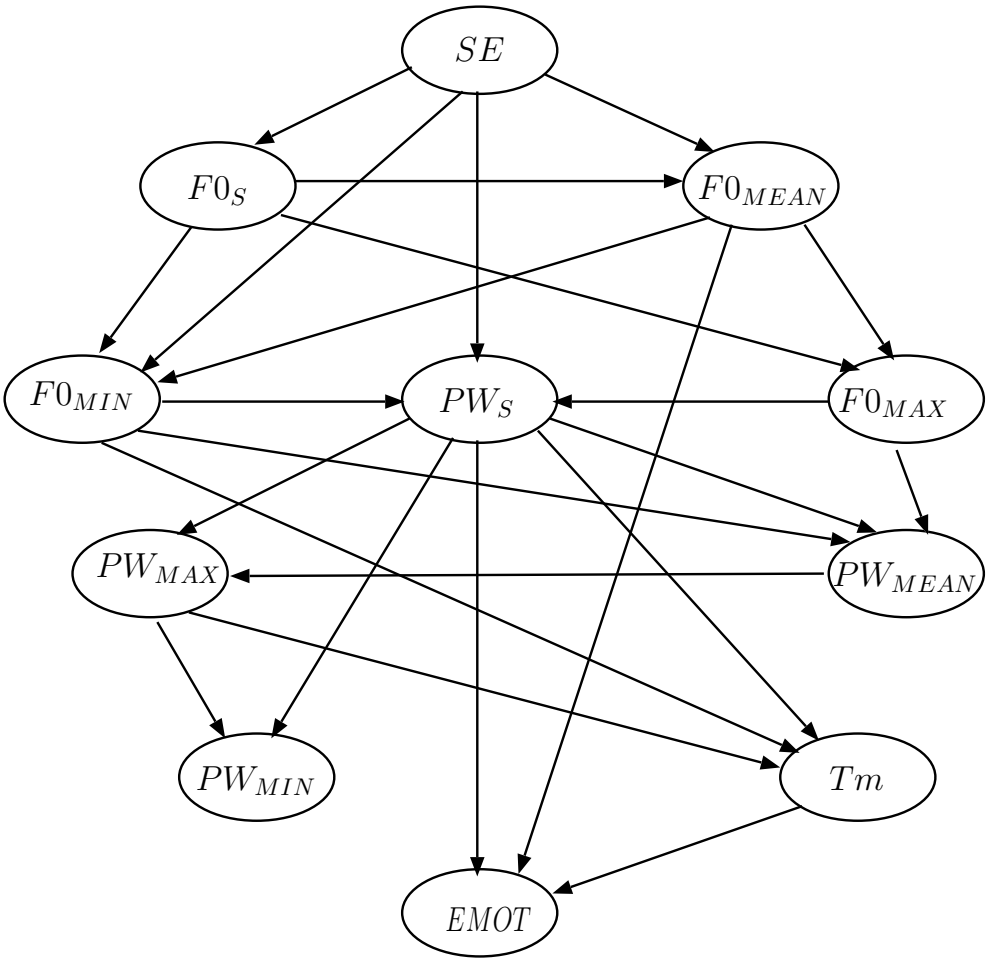


Figure 2.6: A BN structure learned from training data

Table 2.5: CPT of $EMOT$

$F0_{MEAN}$	PW_S	Tm	Anger	Sadness	Disgust	Fear	Surprise	Happiness
0	0	0	0.042	0.125	0.375	0.375	0.042	0.042
0	0	1	0.083	0.250	0.250	0.083	0.083	0.250
0	0	2	0.056	0.167	0.278	0.278	0.167	0.056
0	0	3	0.094	0.156	0.219	0.406	0.031	0.094
0	0	4	0.050	0.250	0.250	0.250	0.050	0.150
0	1	0	0.042	0.125	0.375	0.375	0.042	0.042
0	1	1	0.050	0.250	0.250	0.250	0.050	0.150
0	1	2	0.063	0.063	0.063	0.563	0.188	0.063
0	1	3	0.063	0.188	0.313	0.313	0.063	0.063
0	1	4	0.050	0.150	0.250	0.450	0.050	0.050
0	2	0	0.214	0.071	0.357	0.071	0.071	0.214
0	2	1	0.100	0.100	0.300	0.100	0.300	0.100
0	2	2	0.083	0.417	0.250	0.083	0.083	0.083
0	2	3	0.125	0.125	0.125	0.125	0.125	0.375
0	2	4	0.036	0.250	0.036	0.179	0.250	0.250
0	3	0	0.192	0.192	0.192	0.038	0.192	0.192
0	3	1	0.318	0.318	0.227	0.045	0.045	0.045
0	3	2	0.083	0.083	0.250	0.083	0.250	0.250
0	3	3	0.125	0.208	0.125	0.208	0.125	0.208
0	3	4	0.071	0.214	0.214	0.071	0.214	0.214
0	4	0	0.083	0.250	0.417	0.083	0.083	0.083
0	4	1	0.167	0.167	0.167	0.167	0.167	0.167
0	4	2	0.214	0.214	0.214	0.071	0.071	0.214
0	4	3	0.313	0.063	0.188	0.063	0.313	0.063
0	4	4	0.063	0.313	0.063	0.063	0.313	0.188
1	0	0	0.125	0.042	0.208	0.542	0.042	0.042
1	0	1	0.136	0.136	0.136	0.227	0.227	0.136
1	0	2	0.038	0.192	0.269	0.038	0.269	0.192
1	0	3	0.056	0.500	0.167	0.056	0.056	0.167
1	0	4	0.063	0.063	0.063	0.563	0.063	0.188
1	1	0	0.056	0.167	0.278	0.278	0.056	0.167
1	1	1	0.357	0.214	0.071	0.071	0.214	0.071
1	1	2	0.214	0.071	0.214	0.357	0.071	0.071
1	1	3	0.071	0.071	0.357	0.071	0.357	0.071
1	1	4	0.071	0.071	0.214	0.071	0.071	0.500
1	2	0	0.607	0.036	0.179	0.036	0.107	0.036
1	2	1	0.375	0.125	0.125	0.125	0.125	0.125
1	2	2	0.125	0.292	0.208	0.208	0.042	0.125
1	2	3	0.208	0.208	0.042	0.208	0.125	0.208
1	2	4	0.063	0.188	0.188	0.063	0.313	0.188
1	3	0	0.643	0.071	0.071	0.071	0.071	0.071
1	3	1	0.313	0.063	0.313	0.063	0.063	0.188
1	3	2	0.188	0.063	0.313	0.063	0.313	0.063
1	3	3	0.292	0.125	0.125	0.042	0.125	0.292
1	3	4	0.150	0.150	0.150	0.250	0.050	0.250
1	4	0	0.188	0.188	0.438	0.063	0.063	0.063
1	4	1	0.300	0.100	0.300	0.100	0.100	0.100
1	4	2	0.389	0.167	0.167	0.056	0.056	0.167
1	4	3	0.188	0.188	0.063	0.063	0.313	0.188
1	4	4	0.250	0.083	0.250	0.083	0.250	0.083

2	0	0	0.500	0.045	0.227	0.045	0.136	0.045
2	0	1	0.167	0.056	0.167	0.278	0.167	0.167
2	0	2	0.071	0.071	0.214	0.357	0.071	0.214
2	0	3	0.107	0.036	0.393	0.179	0.179	0.107
2	0	4	0.083	0.250	0.083	0.083	0.417	0.083
2	1	0	0.094	0.156	0.344	0.281	0.094	0.031
2	1	1	0.214	0.357	0.071	0.071	0.214	0.071
2	1	2	0.250	0.083	0.250	0.083	0.250	0.083
2	1	3	0.063	0.188	0.063	0.313	0.188	0.188
2	1	4	0.063	0.063	0.063	0.313	0.063	0.438
2	2	0	0.611	0.056	0.167	0.056	0.056	0.056
2	2	1	0.045	0.045	0.409	0.045	0.227	0.227
2	2	2	0.150	0.350	0.050	0.050	0.150	0.250
2	2	3	0.250	0.083	0.250	0.083	0.250	0.083
2	2	4	0.063	0.188	0.063	0.313	0.313	0.063
2	3	0	0.500	0.136	0.045	0.227	0.045	0.045
2	3	1	0.357	0.214	0.214	0.071	0.071	0.071
2	3	2	0.250	0.250	0.083	0.083	0.250	0.083
2	3	3	0.250	0.083	0.417	0.083	0.083	0.083
2	3	4	0.083	0.083	0.083	0.083	0.250	0.417
2	4	0	0.300	0.300	0.100	0.100	0.100	0.100
2	4	1	0.389	0.167	0.278	0.056	0.056	0.056
2	4	2	0.208	0.292	0.125	0.042	0.125	0.208
2	4	3	0.188	0.188	0.313	0.063	0.063	0.188
2	4	4	0.150	0.250	0.050	0.050	0.450	0.050
3	0	0	0.167	0.167	0.056	0.500	0.056	0.056
3	0	1	0.643	0.071	0.071	0.071	0.071	0.071
3	0	2	0.227	0.318	0.045	0.045	0.318	0.045
3	0	3	0.150	0.150	0.050	0.250	0.150	0.250
3	0	4	0.083	0.083	0.250	0.417	0.083	0.083
3	1	0	0.500	0.167	0.167	0.056	0.056	0.056
3	1	1	0.357	0.071	0.071	0.071	0.214	0.214
3	1	2	0.083	0.083	0.083	0.417	0.250	0.083
3	1	3	0.036	0.250	0.107	0.321	0.179	0.107
3	1	4	0.045	0.227	0.045	0.227	0.409	0.045
3	2	0	0.563	0.063	0.188	0.063	0.063	0.063
3	2	1	0.250	0.083	0.083	0.083	0.083	0.417
3	2	2	0.278	0.167	0.056	0.056	0.278	0.167
3	2	3	0.208	0.125	0.208	0.208	0.208	0.042
3	2	4	0.125	0.125	0.125	0.125	0.375	0.125
3	3	0	0.300	0.100	0.100	0.100	0.300	0.100
3	3	1	0.214	0.071	0.071	0.071	0.214	0.357
3	3	2	0.071	0.214	0.214	0.071	0.357	0.071
3	3	3	0.063	0.188	0.063	0.063	0.313	0.313
3	3	4	0.313	0.063	0.188	0.063	0.188	0.188
3	4	0	0.682	0.136	0.045	0.045	0.045	0.045
3	4	1	0.214	0.357	0.214	0.071	0.071	0.071
3	4	2	0.389	0.278	0.056	0.056	0.056	0.167
3	4	3	0.278	0.167	0.167	0.056	0.167	0.167
3	4	4	0.107	0.321	0.036	0.036	0.179	0.321

4	0	0	0.357	0.071	0.071	0.071	0.071	0.357
4	0	1	0.357	0.071	0.071	0.071	0.214	0.214
4	0	2	0.100	0.100	0.100	0.100	0.500	0.100
4	0	3	0.045	0.136	0.045	0.409	0.227	0.136
4	0	4	0.063	0.063	0.188	0.063	0.063	0.563
4	1	0	0.125	0.125	0.125	0.125	0.375	0.125
4	1	1	0.417	0.083	0.083	0.083	0.250	0.083
4	1	2	0.438	0.313	0.063	0.063	0.063	0.063
4	1	3	0.278	0.167	0.056	0.056	0.056	0.389
4	1	4	0.036	0.107	0.036	0.036	0.321	0.464
4	2	0	0.423	0.038	0.115	0.192	0.192	0.038
4	2	1	0.063	0.313	0.063	0.063	0.188	0.313
4	2	2	0.250	0.083	0.083	0.083	0.250	0.250
4	2	3	0.071	0.214	0.071	0.071	0.071	0.500
4	2	4	0.083	0.417	0.083	0.083	0.083	0.250
4	3	0	0.350	0.050	0.150	0.050	0.150	0.250
4	3	1	0.500	0.167	0.056	0.056	0.056	0.167
4	3	2	0.591	0.136	0.045	0.045	0.045	0.136
4	3	3	0.150	0.150	0.150	0.050	0.350	0.150
4	3	4	0.125	0.125	0.125	0.125	0.125	0.375
4	4	0	0.643	0.071	0.071	0.071	0.071	0.071
4	4	1	0.357	0.214	0.071	0.071	0.214	0.071
4	4	2	0.150	0.550	0.150	0.050	0.050	0.050
4	4	3	0.318	0.227	0.136	0.045	0.136	0.136
4	4	4	0.063	0.438	0.188	0.063	0.063	0.188

Table 2.6: Itemization of Inference Conditions of Emotions on the Discretized Feature Values

		The number of data														
Feature		$F0_{MEAN}$					PW_S					Tm				
Value		0	1	2	3	4	0	1	2	3	4	0	1	2	3	4
Emotion	Anger	5	10	9	11	12	4	6	10	13	**14**	**15**	14	9	7	2
	Sadness	**12**	3	5	3	4	4	2	7	5	**9**	2	7	8	3	7
	Disgust	**15**	8	7	1	0	6	7	5	6	7	**8**	7	6	7	3
	Fear	**11**	7	4	6	1	**13**	10	3	2	1	5	4	5	9	6
	Surprise	8	7	6	8	4	5	4	**9**	**9**	6	3	3	**10**	7	**10**
	Happiness	9	4	2	5	8	4	4	7	**10**	3	2	5	3	8	**10**

Table 2.7: The Accuracy Rates by BN and PCA

Emotion	Accuracy Rates (%)	
	BN	PCA
Anger	**72.1**	27.9
Sadness	**64.9**	18.9
Disgust	**64.1**	59.0
Fear	**55.6**	18.5
Surprise	**59.3**	33.3
Happiness	**63.0**	22.2

Table 2.8: The Accuracy Rates

Emotion	Accuracy Rates (%)		
	Whole evidence(10)	Six evidence	Four evidence
Anger	72.1	54.1	58.1
Sadness	64.9	24.3	27.0
Disgust	64.1	38.5	35.9
Fear	55.6	37.0	37.0
Surprise	59.3	48.1	44.4
Happiness	63.0	40.7	44.4

2.4.1 Detecting Emotion Performance

We converted acoustic features of the remaining 200 samples into discrete values by using the same thresholds for the training data and then examined the detecting emotion performance of the BN model shown in Figure 2.6 on the 200 samples [42, 43, 44]. Table 2.8 shows the results. The singleton BN had accuracy rates of detecting emotion higher than 50% for all emotions. The accuracy rate for anger was higher than the other emotions, whereas those for fear and surprise were lower than the other emotions.

For comparison, we used principal component analysis (PCA) and a classification based on the Mahalanobis distance in a four PC space in a four dimensional hyperplane using four PCs because the accumulated contribution relevance is more than 90%. Table 2.8 shows the results. The accuracy rate of detecting emotion by the singleton BN was higher than that of PCA. These results indicate that the BN had acceptable accuracy rates for the whole range of emotions.

Table 2.9: The Accuracy Rates under Complete Evidence and Incomplete Evidence

		Accuracy Rates (%)			
		Complete evidence (10 evidences)	Incomplete evidence		
			No F0 (6)	No PW (6)	No Tm (9)
Emotion	Anger	72.1	55.8	65.1	60.5
	Sadness	64.9	37.8	**27.0**	40.5
	Disgust	64.1	**17.9**	46.2	**23.1**
	Fear	55.6	25.9	**22.2**	59.3
	Surprise	59.3	40.7	**22.2**	**14.8**
	Happiness	63.0	33.3	48.1	**14.8**
Total		64.0	36.0	40.5	37.0

2.4.2 Reasoning with Incomplete

We examined the inference performance that a part of evidence is given [42, 44]. Table 2.8 shows the results. when four ($F0_{MAX}$, PW_S, PW_{MEAN}, Tm) and six (SE, $F0_S$, $F0_{MAX}$, PW_S, PW_{MEAN}, Tm) evidences are given, accuracy rates went down than whole evidences are given. However, accuracy rates were higher than probability(16.7%) that we randomly answered.

To confirm the influence of detecting emotion when acoustic features are mis-analyzed, we examined the detecting emotion of the BN when only incomplete evidence is available [43]. We firstly examined the situation where the BN lacked evidence on one acoustic feature[1] of $F0$, PW, or Tm and attempted to find which of the three acoustic features is essential for detecting emotion.

The right left half of Table 2.9 shows the results. The results indicate that accuracy rates except for anger decrease dramatically appreciably in comparison with inference using complete evidence. Sadness cannot be detected without the PW feature; disgust cannot be detected without both $F0$ and Tm features; fear cannot be detected without the PW feature; surprise cannot be detected without both PW and Tm features; and happiness cannot be detected without the Tm feature.

2.4.3 Runtime Practicality

We evaluated the runtime performance in an attempt to confirm that our Bayesian approach is applicable to the emotion detection engine of a robot's communication

[1]It should be noted that knowledge of acoustic feature for emotion detection is represented by four nodes for $F0$, four nodes for PW, and one node for Tm in the BN.

system [42, 43]. The time required for emotion detection from a voice waveform can be estimated as

$$\frac{1}{n}\sum_{i=1}^{n} T_{prosodic}(v_i) + T_{mora}(v_i) + T_{inference}(v_i), \tag{2.2}$$

where $T_{prosodic}(v_i)$ is the time required for extracting the acoustic features from a voice waveform v_i, $T_{mora}(v_i)$ is the time to syllabify v_i for measuring the duration per mora, and $T_{inference}(v_i)$ is the time required for detecting emotion. Regarding the above 1600 samples, the execution time except $T_{mora}(v_i)$ took 143.5 [msec] for a single voice waveform on an Athlon 64 3500+ (2.2GHz) / 2GB memory PC. The time required for emotion detection, therefore, is less than 300 [msec] on the assumption that $T_{mora} \leq T_{prosodic} + T_{inference}$. This runtime performance is acceptable in consideration of the latency of human-robot communication.

2.5 Detecting Emotion using Biphase BN

Section 2.4 provided a singleton Bayesian modeling of acoustic features of voice for emotion detection by constructing BN from numbers of segments of emotive, expressive voice samples. And the practical usefulness of probabilistic reasoning in the BN was reported. However, the reasoning method was quite simple: to output a certain emotion that has the highest probability in the BN. There is no consideration of the probability distribution among emotions. In this section, we propose a biphase inference method using the BNs [45].

Considering practical application for human-robot communication, a sophisticated inference mechanism using BNs should be invent for emotion detection in the case of subtle difference in probability values. BN can reduce the choice, even though it does not infer a certain emotion with height probability. In this case, our method can re-infer using a BN modelling with the reduced emotions at the second phase. The biphase Bayesian inference method is the following procedure.

Preparation: In advance of emotion detection inference, one BN and a set of BNs are constructed from training voice samples for the first and the second phase, respectively. The BN at the first phase distinguishes N kinds of emotions, and a BN at the second phase distinguishes $M(< N)$ kinds of emotions. The set of BNs for the second phase is a M-combination from a set with N emotions. The BN at the first phase is denoted by BN^N and a BN at the second phase is denoted by BN_i^M $(i = 1, \cdots, {}_N C_M)$.

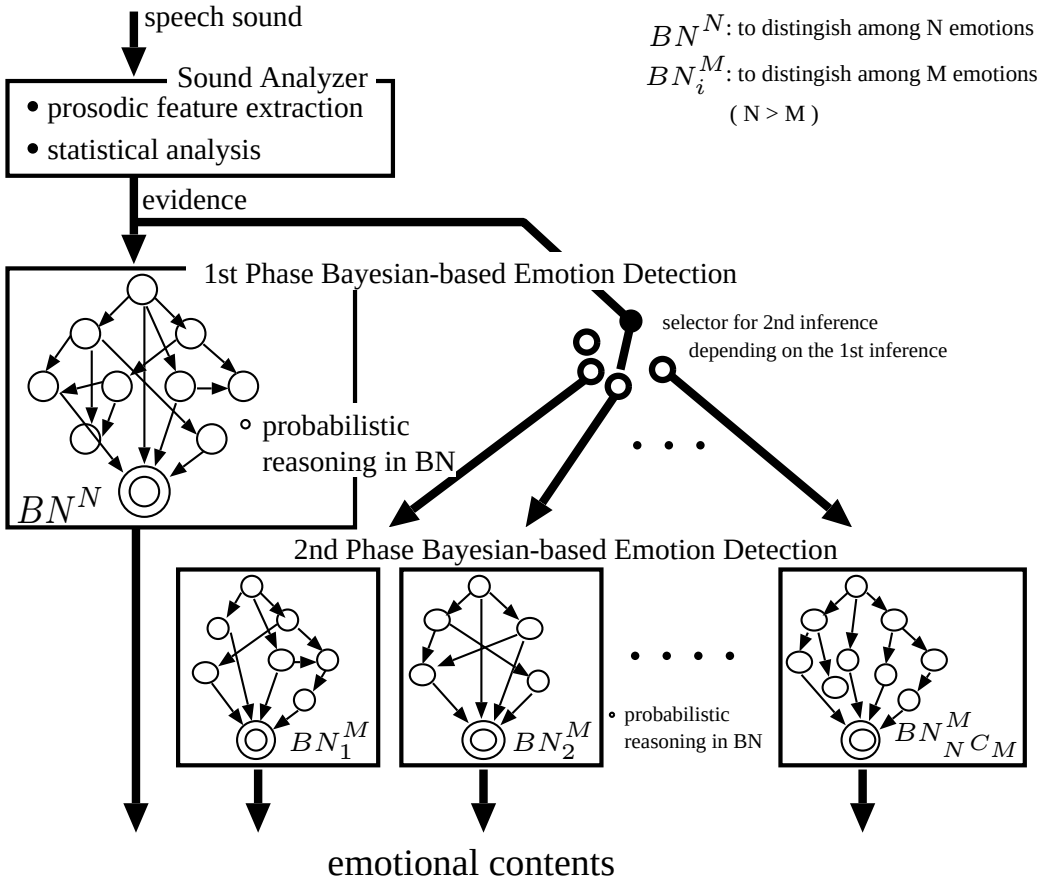


Figure 2.7: General of biphase Bayesian-based emotion detection system

First phase: Probabilistic reasoning in BN^N with some evidences of acoustic features of speech sound from sound analyzer makes conditional probability of emotion, given the evidences. The method outputs a certain emotional label and terminates if the probability satisfies the following condition:

$$\max_{e \in Emotions}(P(e)) > \alpha \quad \text{and} \quad \underset{e \in Emotions}{\text{argmax}}\,(P(e)) \text{ is unique,}$$

where α means the threshold of emotion determination at the first phase BN^N; otherwise the method selects BN_i^M according to the emotional labels within the M-th highest probability and then shifts to the second phase.

Second phase: Probabilistic reasoning in BN_i^M makes conditional probability of emotion, given the evidences. The method determines an emotional label with the highest probability.

Figure 2.7 show general of biphase Bayesian-based emotion detection system.

Table 2.10: Parent Nodes of $EMOT$s on Each Second Phase BN

Pair of emotions	Parent nodes
Anger, sadness	$SE, F0_S, F0_{MIN}, PW_S, PW_{MEAN}, T_m$
Anger, disgust	$SE, F0_{MEAN}, F0_{MAX}, PW_{MAX}, T_m$
Anger, fear	$SE, F0_{MEAN}, PW_S, PW_{MAX}, T_m$
Anger, surprise	$F0_{MIN}, PW_S, PW_{MAX}, PW_{MIN}, T_m$
Anger, happiness	$F0_S, PW_{MEAN}, PW_{MAX}, PW_{MIN}, T_m$
Sadness, disgust	$SE, F0_{MAX}, F0_{MIN}, PW_{MEAN}, PW_{MIN}$
Sadness, fear	$SE, F0_{MAX}, PW_S, PW_{MEAN}, PW_{MIN}$
Sadness, surprise	$SE, F0_{MAX}, PW_{MEAN}, PW_{MAX}, PW_{MIN}$
Sadness, happiness	$SE, F0_{MIN}, PW_{MEAN}, PW_{MAX}, PW_{MIN}, T_m$
Disgust, fear	$SE, F0_S, F0_{MEAN}, PW_S, PW_{MIN}$
Disgust, surprise	$F0_{MAX}, F0_{MIN}, PW_S, PW_{MIN}, T_m$
Disgust, happiness	$SE, F0_S, F0_{MAX}, T_m$
Fear, surprise	$SE, F0_S, F0_{MIN}, PW_S, PW_{MEAN}$
Fear, happiness	$SE, F0_{MAX}, F0_{MIN}, PW_S, PW_{MIN}$
Surprise, happiness	$F0_{MAX}, PW_S, PW_{MEAN}, PW_{MAX}, PW_{MIN}$

This section describes an experimentation of detecting emotion of our biphase-based Bayesian approach. We randomly selected 1400 samples as training data and discretized their attribute values with five values. In this experiment, we determined the threshold for the discretization on the basis of the idea of even-sized chunk, that is, each label of the discrete values covers 20% of the training data. Then we modeled one BN BN^6 which distinguishes six kinds of emotions and fifteen BNs $BN_i^2\,(i = 1, \cdots, 15)$ which distinguish two kinds of emotions for the first and the second phase, respectively (i.e., $N = 6$ and $M = 2$ for the biphase-based method). BNs were built from the above mentioned training voice samples with changing six variable orders by Bayes Net Toolbox [41].

Figure 2.6 shows the results with the variable order $SE \prec F0 \prec PW \prec Tm \prec EMOT$ for the first phase BN^6. In addition, appendix A shows BN models for the second phase BN_2^6, and Table 2.10 shows parent nodes of $EMOT$s on each second phase BNs. $EMOT$s on each second phase BNs connect to attribute nodes that strongly influence inference of their pair of emotions. Therefore, we expect that biphase method improves emotion inference performance.

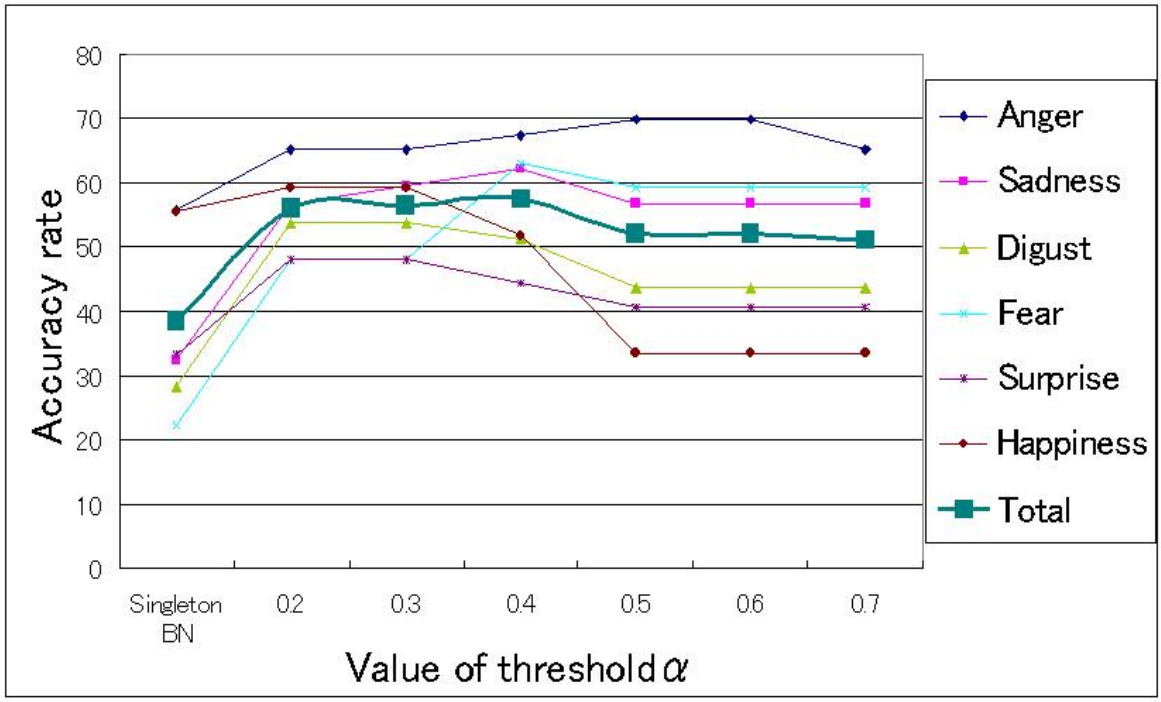


Figure 2.8: The accuracy rates of biphase detecting emotion

2.5.1 Detecting Emotion Performance

We converted acoustic features of the rest 200 samples into discrete value by the same thresholds for the training data, and then we examined the inference performance of the biphase BN model by the inference test. In reasoning with the BN, we used junction tree [22] as the inference algorithm with BNs. The examination was done by changing the threshold α.

Figure 2.8 shows the accuracy rates of detecting emotion. We examined the inference performance using singleton BN for comparison. Reasoner using first phase BN cannot determine a certain emotional label if there is more than one emotional label with the highest probability. This case is a frequent occurrence, and it makes a lower accuracy rate. On the other hand, the results indicate that biphase-based method has largely acceptable accuracy rates for all emotions in comparison with singleton BN (except for happiness on $\alpha = 0.4 - 0.7$). In this particular examples, the total accuracy rates were the best average using biphase-based method with $\alpha = 0.4$; In the method with $\alpha \geq 0.7$ BN^6, BN at the first phase, could not determinate and any of BN_i^2s, a BN at the second phase, was used for all of 200 samples.

2.6 Conclusion

In this chapter, we proposed detecting emotion method from acoustic feature of human voice. The method is singleton and biphase BN using K2 algorithm. Our methods proposed in this chapter gives much benefit to emotion detection by probabilistic inference from complete whole and partial evidence and reasoning under uncertainty. In addition, we confirmed that specific acoustic features strongly influence detection of each emotion respectively.

In this study, we used voices spoken by unspecified actors and actresses from free utterances. It causes inaccurate emotion detection performance. In Chapter 3, we propose pairwise classification using BN for detecting emotion.

Chapter 3

Detecting Emotion Method: Pairwise Classification using TAN

3.1 Introduction

In the research field of data mining, various classification methods have been studied. Many researches for classification focus on feature extraction, feature selection, and classification techniques. Feature extraction and feature selection are a data preprocessing step for acquiring efficient features. Feature extraction transforms input data into reduced representation features. As feature extraction techniques, PCA (Principal Component Analysis), ICA (Independent Component Analysis), and LDA (Linear Discriminant Analysis) are commonly used. Feature selection is done by selecting appropriate subset features for classifying. Generally, stepwise selection and genetic algorithm are used for feature selection. For efficient classification from the acquiring features, classification techniques have been researched (e.g., Support Vector Machine, Neural Network, Bayesian Network Classifier, Linear Discrimination classifier, and K-nearest neighbor).

A multi-class is more complex than a binary class in classification problem. pairwise classification (one against one) and one against all, for classifying a multi-class using a series of binary classifiers, have been researched. One against all transforms C classes problem into C binary classification for one class against other classes. Pairwise classification [46, 47, 48] converts a multi-class problem into the series of binary class problems ($C(C-2)/2$).

In Chapter 2, we studied Bayesian method using K2 algorithm for detecting emotion. In this chapter, we propose pairwise classification by weighted probability for detecting emotion [49, 50]: using TAN in which the emotional contents of the voice

are modeled by their selected acoustic features on every pair of emotions. Here, we report on experiment results of detecting emotion from voice and classification from open databases, and compare accuracy rates of our method with other methods.

3.2 Pairwise Classification

Generally, a binary class has higher classification ability than a multi-class. Accordingly, a multi-class sometimes resolves a series of binary class classifiers. Pairwise classification [47, 46, 48] converts a C class problem into $C(C-1)/2$ binary problems for a pair of classes. Otherwise, one against all classifies the C class problem into C binary problems for one class and other classes. Pairwise classification is more common between multi binary classifiers, because pairwise classification has relatively high ability. In pairwise classification, each binary classifier C_{ij}(classifiers for class i, j) is learned on the subset of training examples that belong to the classes C_i and C_j, all other examples are ignored for the training of C_{ij}. Classification from each binary classifier adopts voting system that classifies the most selected class by using each binary classifiers. However, it can not classify class exactly if the most win is equal over two class. In this chapter, we use weighted probability pairwise classification by posterior probability of BN.

3.3 Constructing Emotion Detection Engine

This section describes constructing binary classifiers for emotion detection.

3.3.1 Voice Data

We collected voice samples for five emotions (anger, sadness, disgust, surprise, or happiness) as mentioned in 2.2.1. These voice samples also are spoken by unspecified actors and actresses from free sentences. In this chapter, we divided them into male and female voice because we examined emotion detection for distinguishing between male and female.

3.3.2 Features Extraction

Acoustic features (i.e. fundamental frequency, energy, duration, formant, mel-frequency cepstral coefficients, duration, speech rate) are used for emotion detection

[36, 51]. In this chapter, as attributes of voice data, we chose acoustic components as: duration, fundamental frequency, energy, from first to fourth formant frequency and bandwidth. Acoustic analysis was done on 11 ms frames passed through a Hamming window extracted from voice waveforms sampled at 22.05 kHz. Then, we extracted 93 acoustic features from fundamental frequency (12 features), energy (16 features), formant (64 features), and duration rate (1 feature). All of acoustic features are as follows:

- Fundamental frequency features ($F0$).
  - $F0_{1-7}$. Standard deviation, mean, maximum, minimum, median, range between maximum and minimum, and timezone of maximum.
  - $F0_{8}$. Gradient of the linear regression line of the F0 contour.
  - $F0_{9-12}$. Amplitude of F0 contour during t seconds after the beginning of the phrase ($t = 0.05, 0.1, 0.15, 0.2$).
- Energy features (PW).
  - PW_{1-7}. Standard deviation, mean, maximum, minimum, median, range between maximum and minimum, and timezone of maximum.
  - PW_{8}. Gradient of the linear regression line of the power envelope.
  - PW_{9-12}. Median value of the first derivative of the power envelope during the t seconds after the beginning of the phrase ($t = 0.05, 0.1, 0.15, 0.2$).
  - PW_{13-16}. Ratio of the power at t seconds after the beginning of the phrase to the maximum power ($t = 0.05, 0.1, 0.15, 0.2$).
- First formant frequency features ($F1$).
  - $F1_{1-7}$. Standard deviation, mean, maximum, minimum, median, range between maximum and minimum, and timezone of maximum.
  - $F1_{8}$. Gradient of the linear regression line of first formant frequency.
- Second formant frequency features ($F2$).
  - $F2_{1-7}$. Standard deviation, mean, maximum, minimum, median, range between maximum and minimum, and timezone of maximum.
  - $F2_{8}$. Gradient of the linear regression line of second formant frequency.
- Third formant frequency features ($F3$).

  - $F3_{1-7}$. Standard deviation, mean, maximum, minimum, median, range between maximum and minimum, and timezone of maximum.
  - $F3_8$. Gradient of the linear regression line of third formant frequency.

- Fourth formant frequency features($F4$).
  - $F4_{1-7}$. Standard deviation, mean, maximum, minimum, median, range between maximum and minimum, and timezone of maximum.
  - $F4_8$. Gradient of the linear regression line of fourth formant frequency.

- First formant bandwidth features ($BW1$).
  - $BW1_{1-7}$. Standard deviation, mean, maximum, minimum, median, range between maximum and minimum, and timezone of maximum.
  - $BW1_8$. Gradient of the linear regression line of first formant bandwidth.

- Second formant bandwidth features ($BW2$).
  - $BW2_{1-7}$. Standard deviation, mean, maximum, minimum, median, range between maximum and minimum, and timezone of maximum.
  - $BW2_8$. Gradient of the linear regression line of second formant bandwidth.

- Third formant bandwidth features ($BW3$).
  - $BW3_{1-7}$. Standard deviation, mean, maximum, minimum, median, range between maximum and minimum, and timezone of maximum.
  - $BW3_8$. Gradient of the linear regression line of third formant bandwidth.

- Fourth formant bandwidth features ($BW4$).
  - $BW4_{1-7}$. Standard deviation, mean, maximum, minimum, median, range between maximum and minimum, and timezone of maximum.
  - $BW4_8$. Gradient of the linear regression line of fourth formant bandwidth.

- Tm. The average duration rate per a single mora.

3.3.3 Feature Selection

Feature selection [52, 53] is a data pre-processing that selects appropriate subset features for classification. Generally, feature selection is done by selecting efficient features that determine the class label. This can also reduce the dimensionality, which can otherwise worsen the performance of the pattern classifiers. Feature selection algorithms for selecting appropriate subset features, such as stepwise, genetic algorithm, are proposed. Stepwise feature selection computes scores of features by adding or removing each feature. Then, one feature acquiring maximum score is selected for adding or removing the feature. In this study, we used the forward-backward stepwise selection method. The score uses accuracy rate of classification using NB. This method consists of two parts: forward stepwise and backward stepwise. Feature selection is conducted as follows:

- forward stepwise.
    1. Compute scores by adding each feature to selected features (initial selected feature is none, and initial maximum score is zero).
    2. Select a feature acquiring maximum score.
    3. If the score is higher than the previous maximum score, add the feature, otherwise, stop feature selection.
    4. If selected features are less than three, go to forward stepwise 1), otherwise, go to backward stepwise 1).
- backward stepwise.
    1. Compute scores by removing each selected features.
    2. Select a feature acquiring maximum score.
    3. If the score is higher than the maximum score of forward stepwise, remove the feature, otherwise, go to forward stepwise 1).
    4. If selected features are less than three, go to forward stepwise 1), otherwise, go to backward stepwise 1).

In this approach, variables once entered may be dropped if those are no longer significant as other variables are added. Feature selection conducts on every pair of emotions for pairwise classification, which obtains the important subset features for classification on every pair of emotions. We expect that feature selection done on each

pair of emotions improves emotion detecting performance for pairwise classification and reduces a computational effort of emotion detection. Subset features criterion use the accuracy rate of classification by NB, and the accuracy rate is estimated by the leave one out method.

3.3.4 Learning Emotion Detection Engine

We constructed the binary classifiers on every pair of emotions. We used TAN as binary classifiers. For detection of the five emotions, we constructed ten ($C(C-1)/2$) binary classifiers. Each classifier used different features that were selected on every pair of emotions from all features.

3.4 Emotion Detection Algorithm

The voting system is a common pairwise classification. However, if the most amount of vote is tied over two classes, the voting system randomly classifies class among them. In this chapter, we used weighted voting for posterior probability by TAN binary classifiers. Emotion detection algorithm is as follows:

Algorithm 3.4.1 Emotion detection algorithm : weighted probability

for $i = 1$ to k-1 **do**
  for $j = i + 1$ to k **do**
    Compute probability P_i, P_j using classifier $C_{i,j}$
    if $P_i > P_j$ **then**
      $w = P_i$
      $sum_i = sum_i + w$ //weighted probability
    else if $P_i < P_j$ **then**
      $w = P_j$
      $sum_j = sum_j + w$ //weighted probability
    else if $P_i = P_j$ **then**
      Continue
    end if
  end for
end for
Detect emotion as $\underset{c=1\cdots k}{\operatorname{argmax}} sum_c$

In algorithms, C_{ij} is classifier of i, j class, P_i is posterior probability of i as classifier C_{ij}, and k is class number.

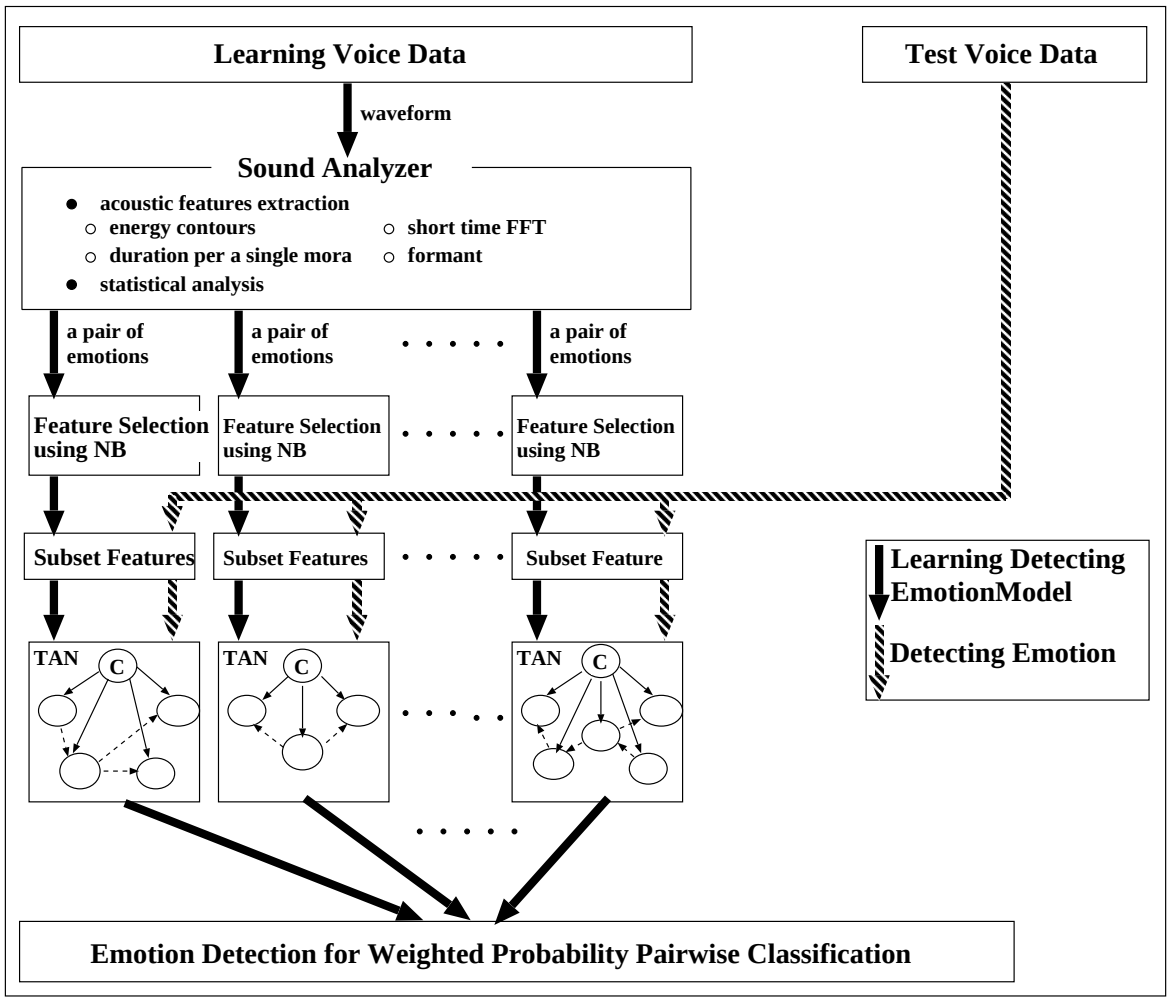


Figure 3.1: General of detecting emotion by proposed method

3.5 Experimental Evaluation of Emotion Detection

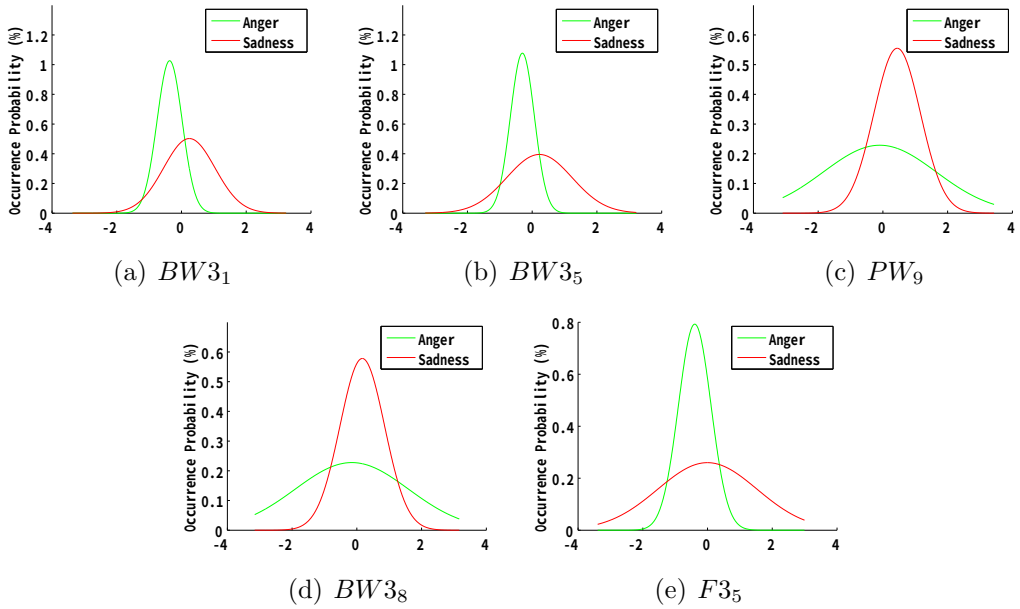
Figure 3.1 shows general of detecting emotion by proposed method. The section describes an experimentation of emotion detection. First, we collected 200 segments of male and female voice waveforms (400 segments in total) and labeled them with the five emotions. Voice samples are same number segments for each emotion (40 segment per emotion). Then, we extracted 93 acoustic features in each of the segments and assigned them to the attributes, as described in Section 3.3.2. The acoustic analysis used the Snack sound toolkit [40]. Then, we conducted feature selection on every pair of emotions, as described in Section 3.3.3. Then we constructed binary classifiers from subset features, as described in Section 3.3.4, with Bayes Net Toolbox [41]. Then, we detected emotion from voice samples, as described in Section 3.4. To evaluate emotion detection performance, we adopted ten fold cross-validation. For ten fold cross-validation, we partitioned voice samples into ten subsamples in which each subsample had the same number of emotion segments. One subsample assigns the test data, and the remaining nine subsamples are used as training data. Then we repeated this ten times as changing test data.

3.5.1 Results of Feature Selection

TAN classifiers calculate posterior probability of emotion from Gaussian distributions of attributes. Table 3.1 shows selected features on pairs of emotions for male, and Figure 3.2 - Figure 3.11 show those Gaussian distributions. Table 3.2 also shows selected features on pairs of emotions for female, and Figure 3.12 - Figure 3.21 show those Gaussian distributions. Table 3.1 (for male) and Table 3.2 (for female) also show feature values between pairs of emotions from Gaussian distributions. Take emotion detection between anger and disgust of female for example. The selected features $F0_3$ (maximum of fundamental frequency), PW_6 (range between maximum and minimum of energy), and $BW2_8$ (gradient of the linear regression line of second formant bandwidth) influence emotion detection between anger and disgust for female: $F0_3$ value of disgust is lower than that of anger; anger has higher PW_6 value than disgust; $BW2_8$ value of anger is broader than that of disgust. Therefore, we can confirm relative features and those feature value for detection on each pair of emotions.

Table 3.1: Selected Features and Feature Values between Pairs of Emotions (male)

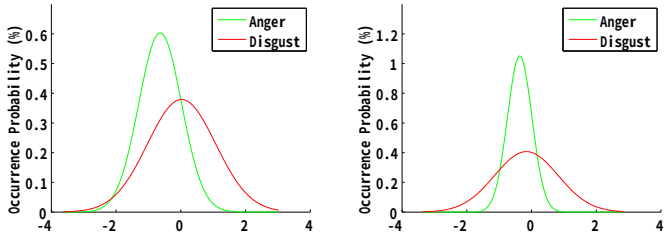
Pair of emotions	Selected features	Feature values on pair of emotions
Anger, sadness	$BW3_1$	Anger - low, sadness - high
	$BW3_5$	Anger - low, sadness - high
	PW_9	Anger - broad, sadness - narrow
	$BW3_8$	Anger - broad, sadness - narrow
	$F3_5$	Anger - narrow, sadness - broad
Anger, disgust	$F4_2$	Anger - low, disgust - high
	$BW3_2$	Anger - narrow, disgust - broad
Anger, surprise	$BW3_5$	Anger - low, surprise - high
	PW_4	Anger - low, surprise - high
	$F3_4$	Anger - low, surprise - high
Anger, happiness	$F4_2$	Anger - low, happiness - high
	$BW3_2$	Anger - narrow, happiness - broad
Sadness, disgust	PW_9	Sadness - high, disgust - low
Sadness, surprise	PW_2	Sadness - low, surprise - high
	$BW1_1$	Sadness - narrow, surprise - broad
Sadness, happiness	PW_2	Sadness - low, happiness - high
	PW_7	Sadness - low, happiness - high
	PW_{12}	Sadness - high, happiness - low
Disgust, surprise	$F0_4$	Disgust - narrow, surprise - broad
	$F0_1$	Disgust - low, surprise - high
Disgust, happiness	$F0_2$	Disgust - low, happiness - high
	PW_8	Disgust - low, happiness - high
Surprise, happiness	$BW3_8$	Surprise - narrow, happiness - broad
	$BW2_2$	Surprise - high, happiness - low
	$F0_8$	Surprise - narrow, happiness - broad



(a) $BW3_1$ (b) $BW3_5$ (c) PW_9

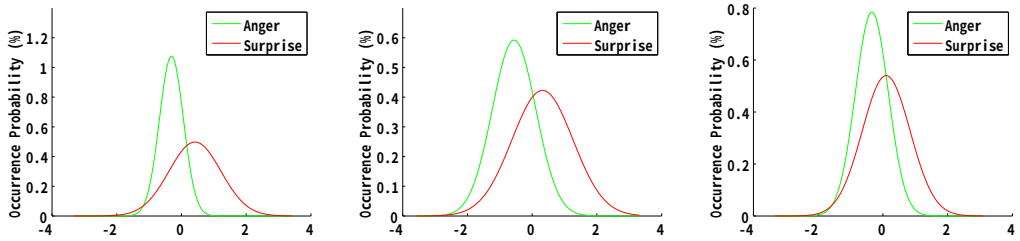
(d) $BW3_8$ (e) $F3_5$

Figure 3.2: Gaussian distributions of selected features on anger and sadness (male)



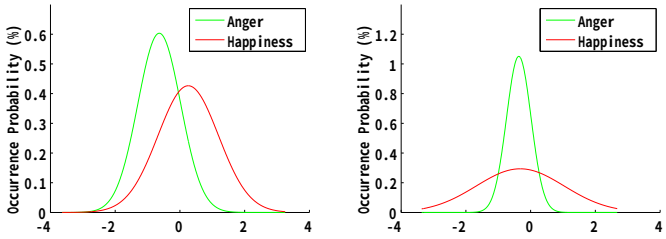
(a) $F4_2$ (b) $BW3_2$

Figure 3.3: Gaussian distributions of selected features on anger and disgust (male)



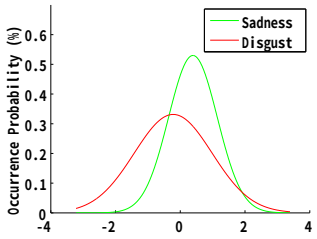
(a) $BW3_5$ (b) PW_4 (c) $F3_4$

Figure 3.4: Gaussian distributions of selected features on anger and surprise (male)



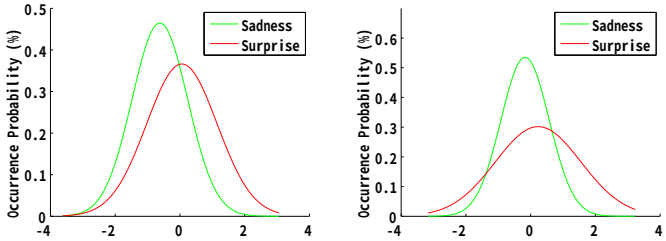
(a) $F4_2$ (b) $BW3_2$

Figure 3.5: Gaussian distributions of selected features on anger and happiness (male)



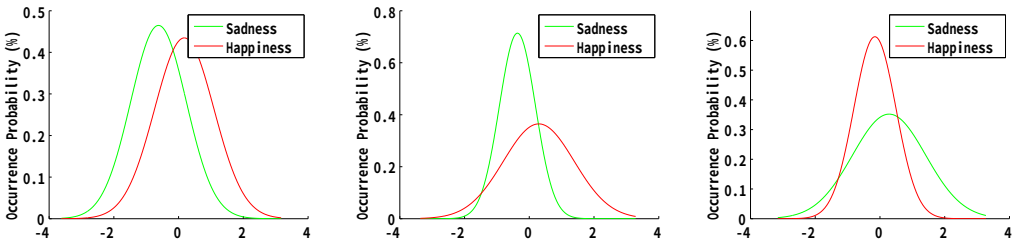
(a) PW_9

Figure 3.6: Gaussian distribution of selected feature on sadness and disgust (male)



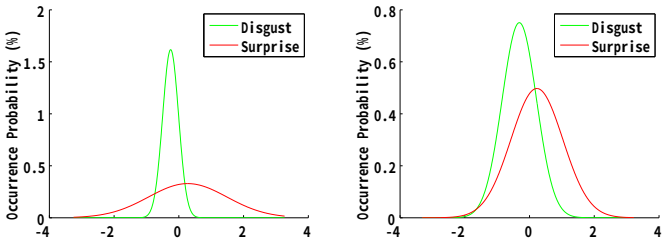
(a) PW_2 (b) $BW1_1$

Figure 3.7: Gaussian distributions of selected features on sadness and surprise (male)



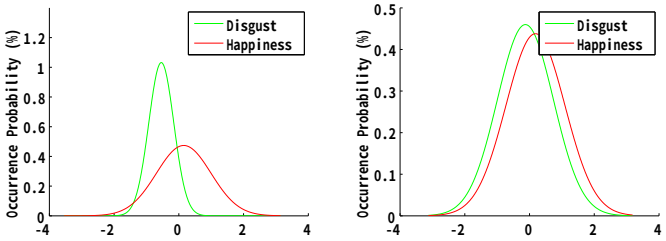
(a) PW_2 (b) PW_7 (c) PW_{12}

Figure 3.8: Gaussian distributions of selected features on sadness and happiness (male)



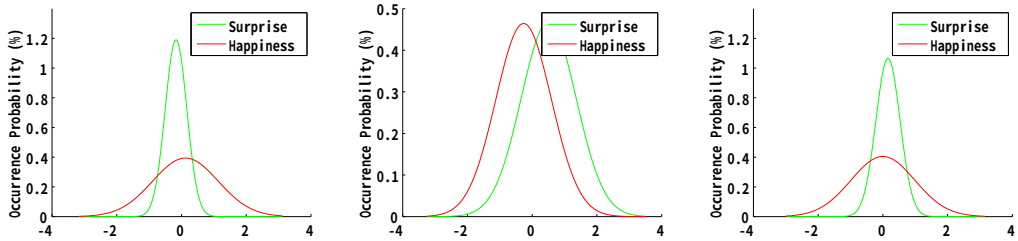
(a) $F0_4$ (b) $F0_1$

Figure 3.9: Gaussian distributions of selected features on disgust and surprise (male)



(a) $F0_2$ (b) PW_8

Figure 3.10: Gaussian distributions of selected features on disgust and happiness (male)

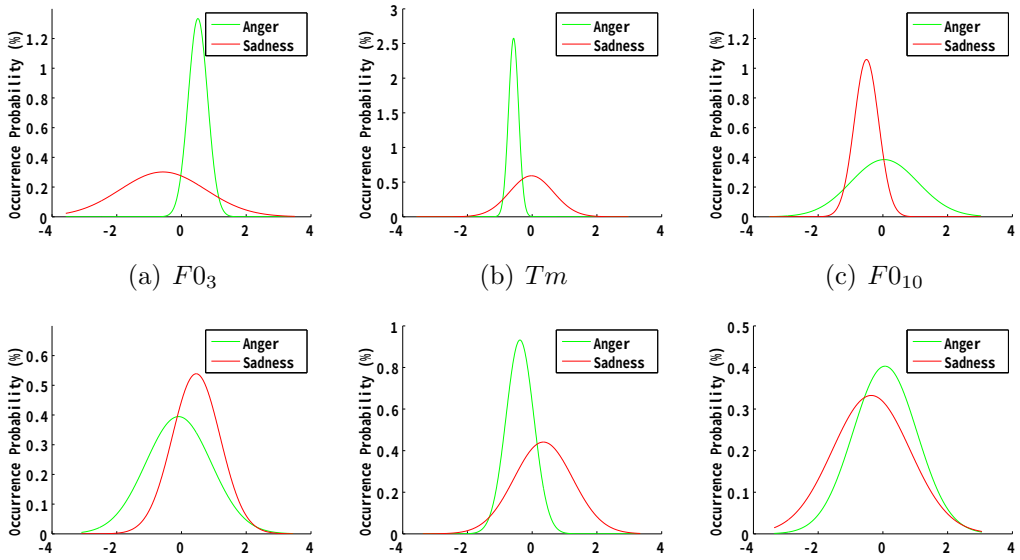


(a) $BW3_8$ (b) $BW2_2$ (c) $F0_8$

Figure 3.11: Gaussian distributions of selected features on surprise and happiness (male)

Table 3.2: Selected Features and Feature Values between Pairs of Emotions (female)

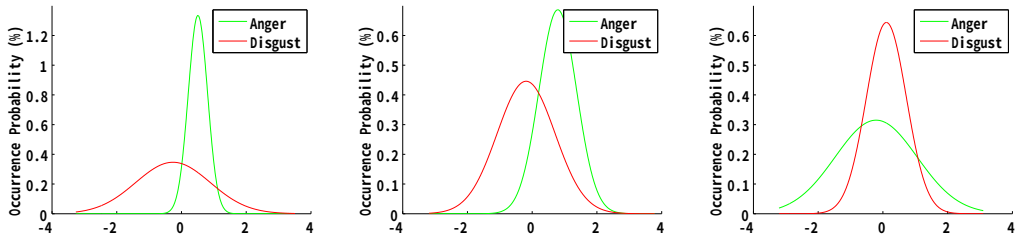
Pair of emotions	Selected features	Feature values on pair of emotions
Anger, sadness	$F0_3$	Anger - high, sadness - low
	Tm	Anger - low, sadness - high
	$F0_{10}$	Anger - high, sadness - low
	$F1_6$	Anger - low, sadness - high
	$BW1_2$	Anger - low, sadness - high
	PW_2	Anger - high, sadness - low
Anger, disgust	$F0_3$	Anger - high, disgust - low
	PW_6	Anger - high, disgust - low
	$BW2_8$	Anger - broad, disgust - narrow
Anger, surprise	$F2_5$	Anger - narrow, surprise - broad
	PW_5	Anger - high, surprise - low
	Tm	Anger - narrow, surprise - broad
	PW_1	Anger - high, surprise - low
	$BW4_8$	Anger - low, surprise - high
	$F1_6$	Similar
Anger, happiness	PW_1	Anger - high, happiness - low
	Tm	Anger - narrow, happiness - broad
	$F2_3$	Similar
	$BW1_5$	Anger - narrow, happiness - broad
Sadness, disgust	$F0_{10}$	Sadness - low, disgust - high
	$F1_3$	Sadness - high, disgust - low
	$F3_2$	Sadness - low, disgust - high
Sadness, surprise	PW_5	Sadness - high, surprise - low
	PW_2	Sadness - low, surprise - high
	$F0_1$	Sadness - low, surprise - high
	$BW4_7$	Sadness - low, surprise - high
	PW_1	Sadness - high, surprise - low
	$BW1_2$	Sadness - high, surprise - low
	PW_8	Sadness - low, surprise - high
	$BW2_2$	Sadness - high, surprise - low
	$F1_7$	Sadness - narrow, surprise - broad
Sadness, Happiness	PW_9	Sadness - broad, happiness - narrow
	$BW1_5$	Sadness - high, happiness - low
	$F0_4$	Sadness - narrow, happiness - broad
Disgust, surprise	$F1_7$	Disgust - low, surprise - high
	$F0_5$	Disgust - low, surprise - high
Disgust, happiness	PW_1	Disgust - broad, happiness - narrow
	$BW4_6$	Disgust - narrow, happiness - broad
	$BW1_1$	Disgust - high, happiness - low
	$F1_4$	Disgust - low, happiness - high
	$F3_3$	Disgust - high, happiness - low
	PW_{16}	Disgust - broad, happiness - narrow
	$BW2_8$	Disgust - narrow, happiness - broad
	PW_7	Disgust - narrow, happiness - broad
	PW_{13}	Disgust - high, happiness - low
Surprise, happiness	PW_7	Surprise - narrow, happiness - broad
	$BW3_5$	Surprise - high, happiness - low



(a) $F0_3$ (b) Tm (c) $F0_{10}$

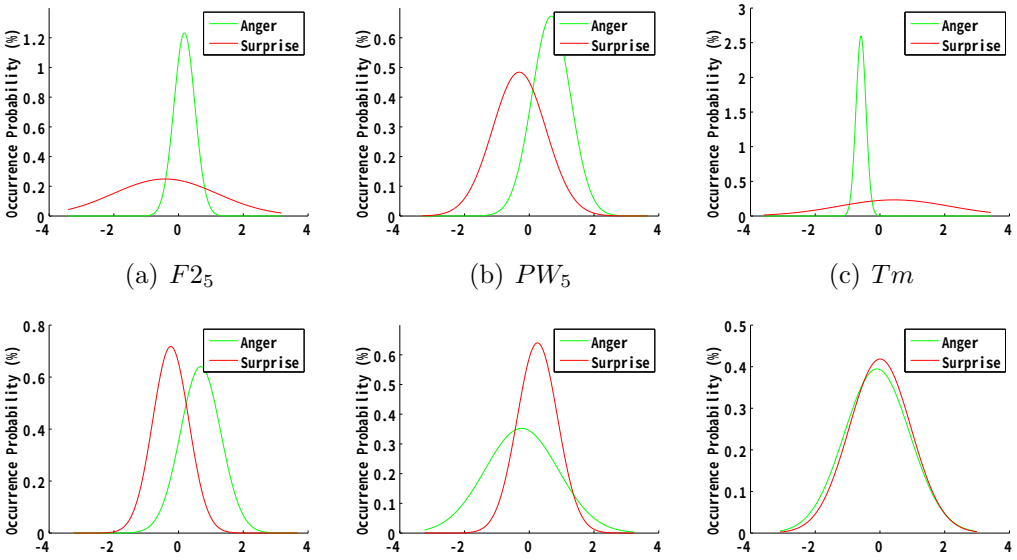
(d) $F1_6$ (e) $BW1_2$ (f) PW_2

Figure 3.12: Gaussian distributions of selected features on anger and sadness (female)



(a) $F0_3$ (b) PW_6 (c) $BW2_8$

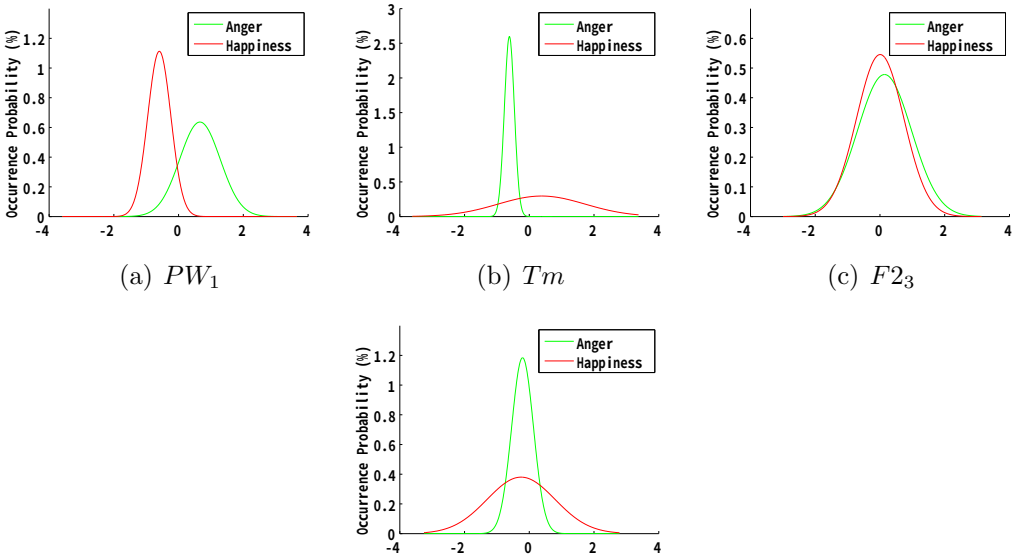
Figure 3.13: Gaussian distributions of selected features on anger and disgust (female)



(a) $F2_5$ (b) PW_5 (c) Tm

(d) PW_1 (e) $BW4_8$ (f) $F1_6$

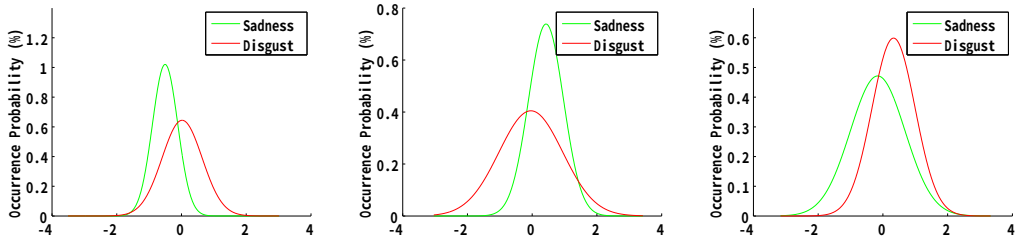
Figure 3.14: Gaussian distributions of selected features on anger and surprise (female)



(a) PW_1 (b) Tm (c) $F2_3$

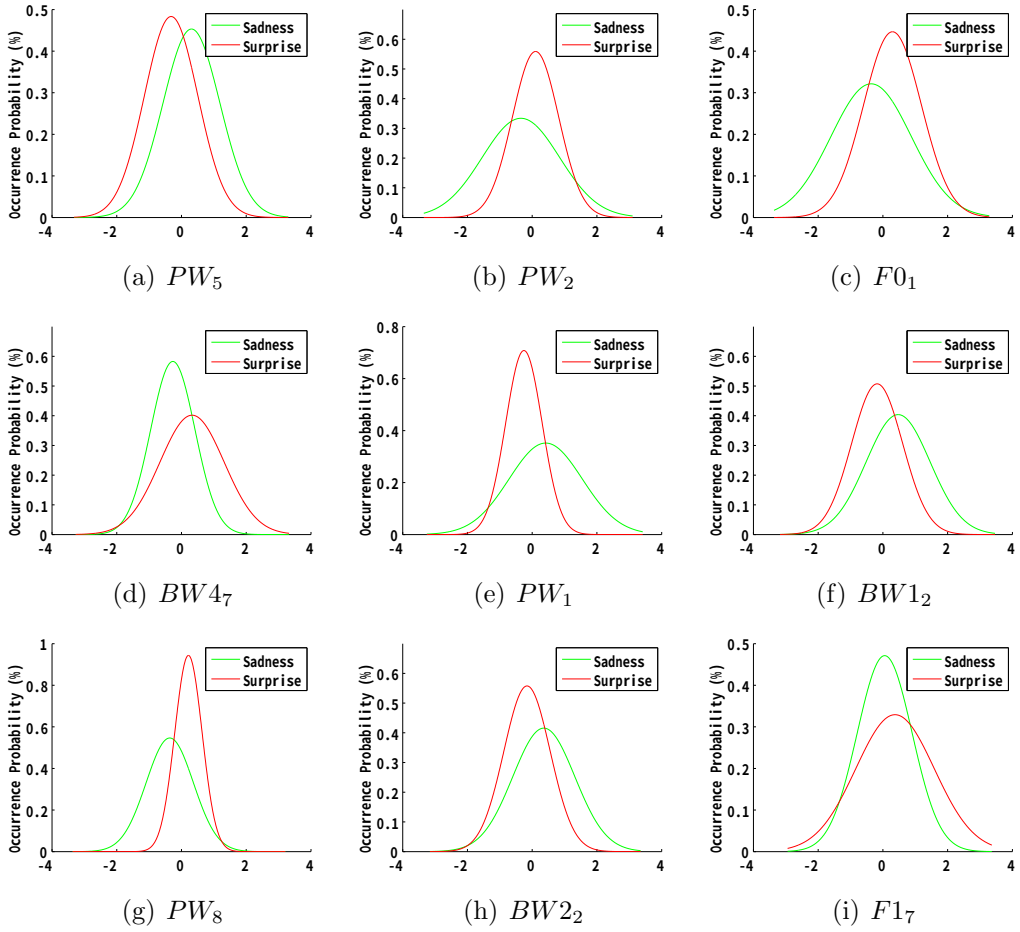
(d) $BW1_5$

Figure 3.15: Gaussian distributions of selected features on anger and happiness (female)



(a) $F0_{10}$ (b) $F1_3$ (c) $F3_2$

Figure 3.16: Gaussian distributions of selected features on sadness and disgust (female)



(a) PW_5 (b) PW_2 (c) $F0_1$

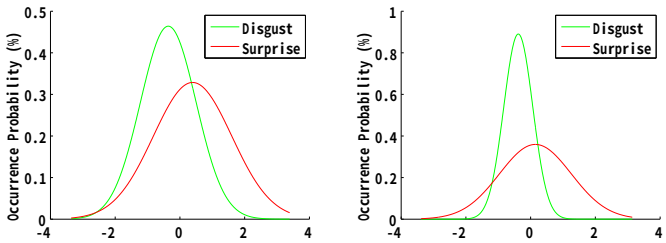
(d) $BW4_7$ (e) PW_1 (f) $BW1_2$

(g) PW_8 (h) $BW2_2$ (i) $F1_7$

Figure 3.17: Gaussian distributions of selected features on sadness and surprise (female)

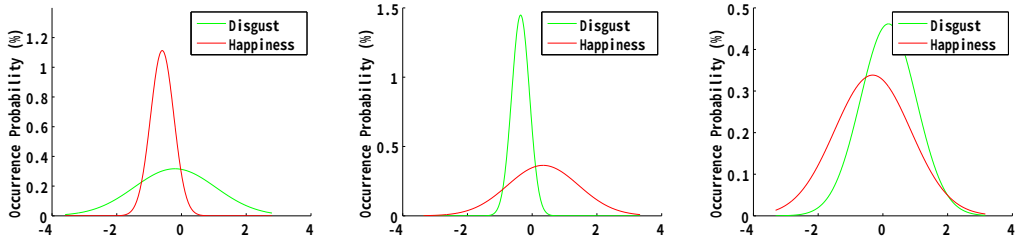
(a) PW_9 (b) $BW1_5$ (c) $F0_4$

Figure 3.18: Gaussian distributions of selected features on sadness and happiness (female)

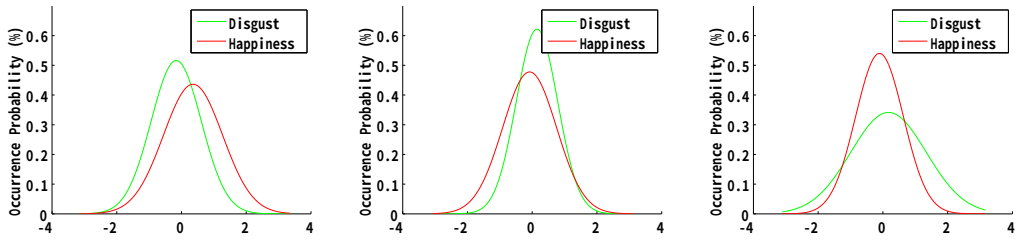


(a) $F1_7$ (b) $F0_5$

Figure 3.19: Gaussian distributions of selected features on disgust and surprise (female)



(a) PW_1 (b) $BW4_6$ (c) $BW1_1$



(d) $F1_4$ (e) $F3_3$ (f) PW_{16}

(g) $BW2_8$ (h) PW_7 (i) PW_{13}

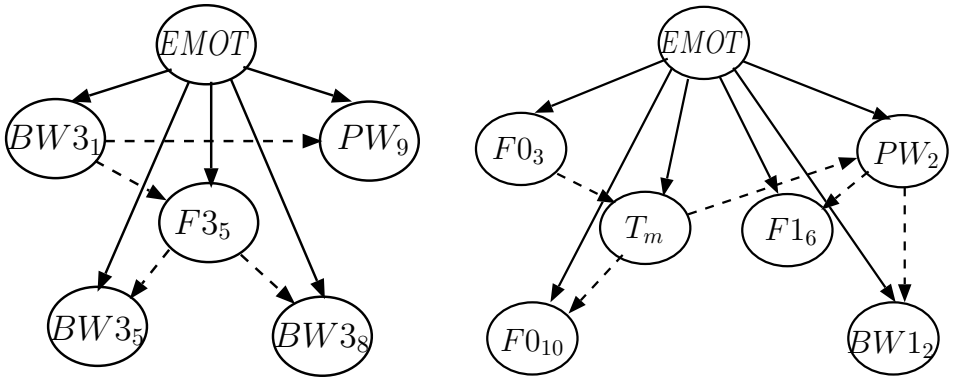
Figure 3.20: Gaussian distributions of selected features on disgust and happiness (female)

(a) PW_7 (b) $BW3_5$

Figure 3.21: Gaussian distributions of selected features on surprise and happiness (female)

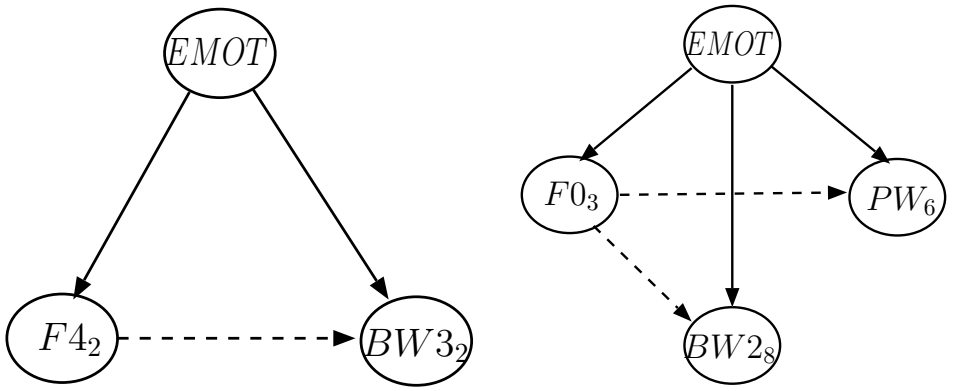
3.5.2 Binary Classifiers

Figure 3.22 - Figure 3.31 show TAN binary classifiers on each pair of emotion. Take binary classifiers of anger and sadness for example. The male model for anger and sadness uses 5 features ($BW3_1$, $BW3_5$, PW_9, $BW3_8$, $F3_5$) (Table 3.1). It constructs tree among eight attributes as $BW3_1$ is root node. Also the female model uses 6 features ($F0_3$, Tm, $F0_{10}$, $F1_6$, $BW1_2$, PW_2) (Table 3.2). It also constructs tree among seven attributes as $F0_3$ is root node.



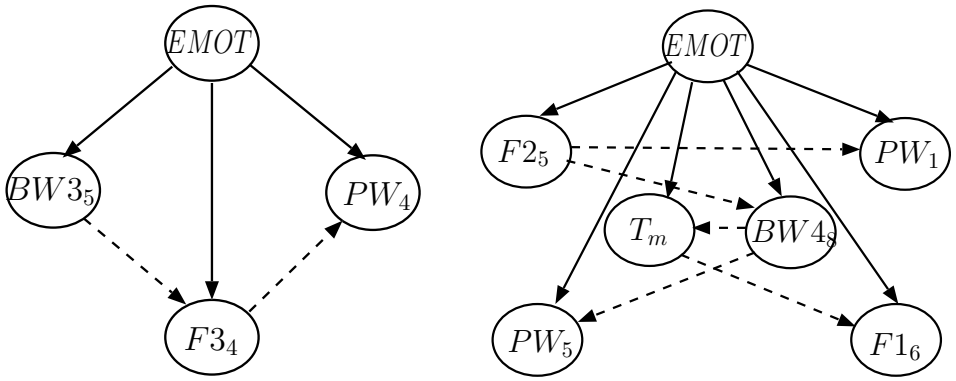
(a) Male (b) Female

Figure 3.22: TAN classifiers of anger and sadness



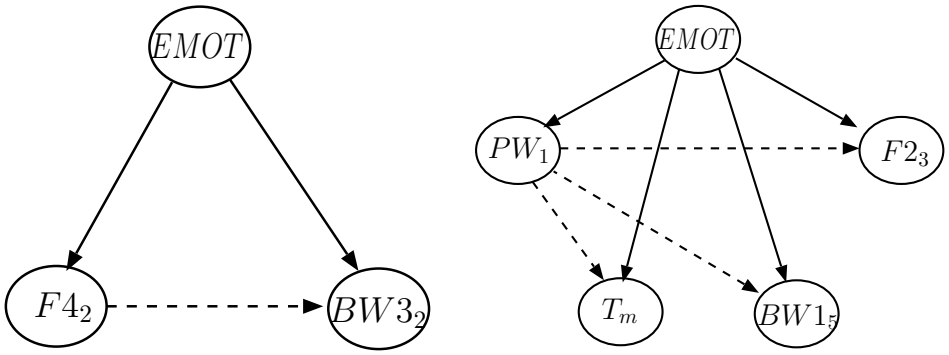
(a) Male (b) Female

Figure 3.23: TAN classifiers of anger and disgust



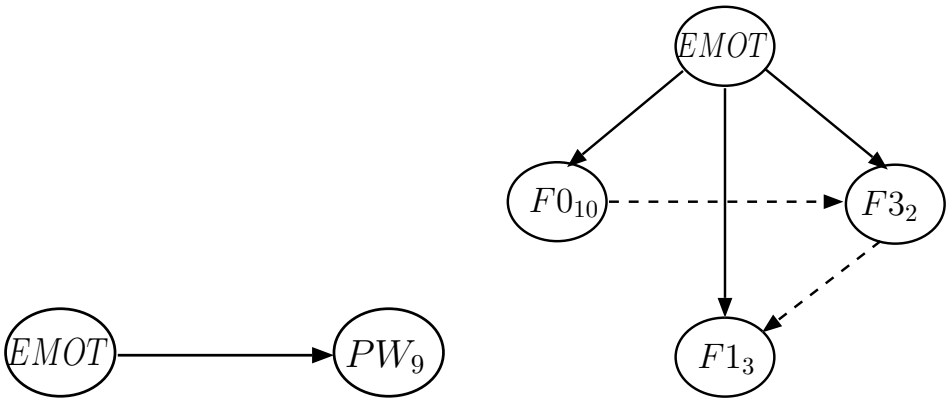
(a) Male (b) Female

Figure 3.24: TAN classifiers of anger and surprise



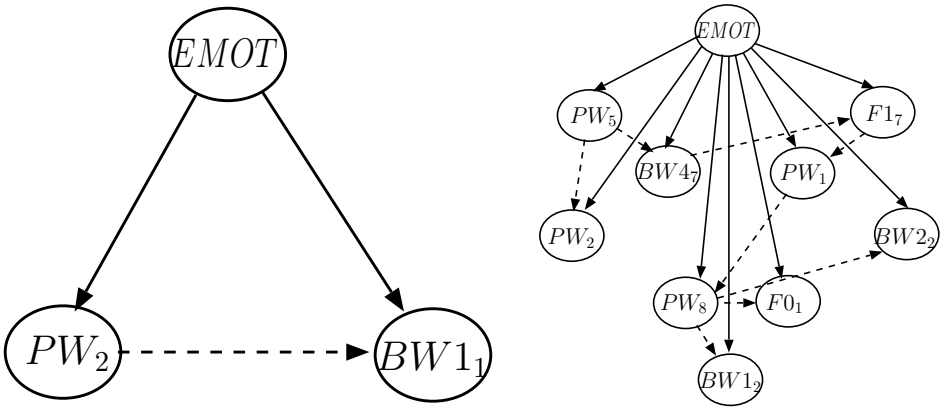
(a) Male (b) Female

Figure 3.25: TAN classifiers of anger and happiness



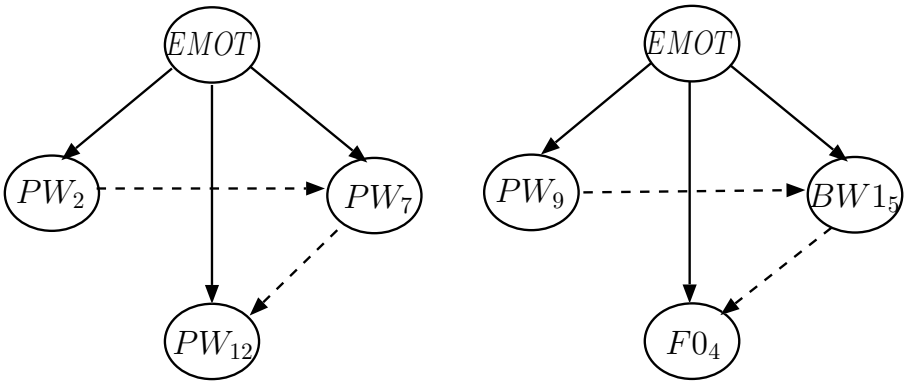
(a) Male (b) Female

Figure 3.26: TAN classifiers of sadness and disgust



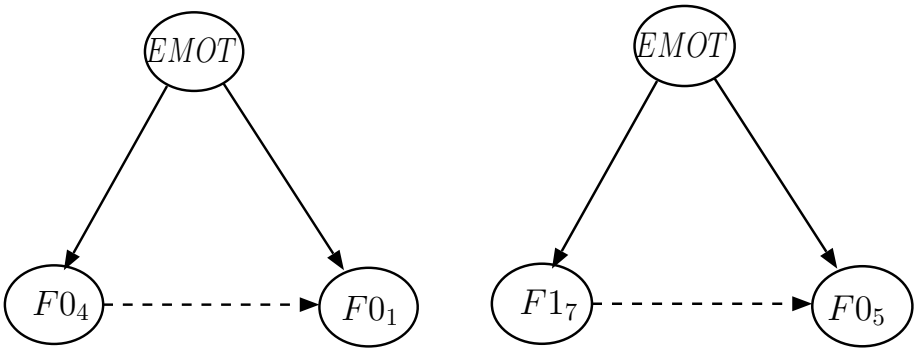
(a) Male (b) Female

Figure 3.27: TAN classifiers of sadness and surprise



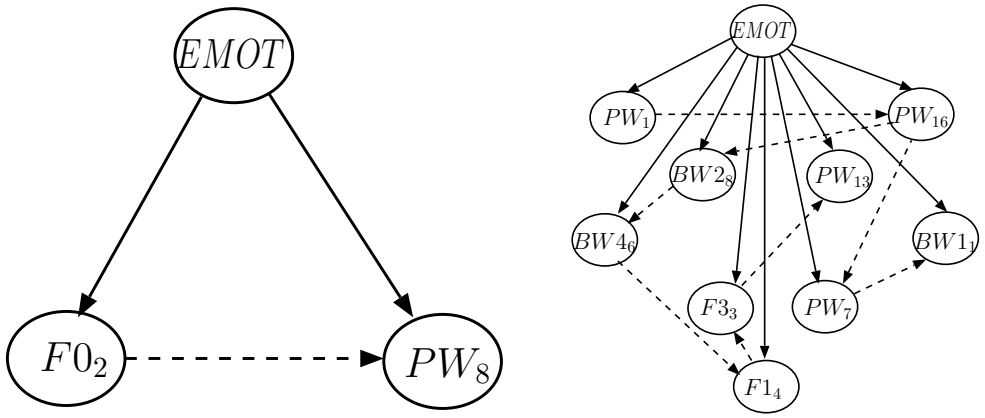
(a) Male (b) Female

Figure 3.28: TAN classifiers of sadness and happiness



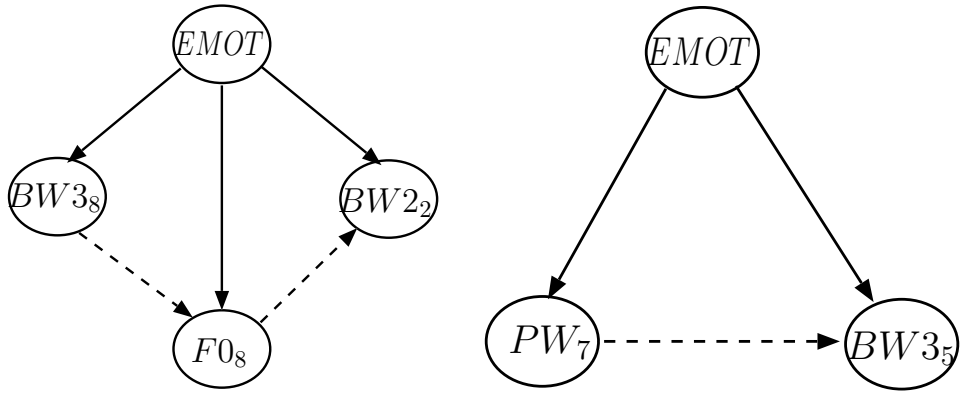
(a) Male (b) Female

Figure 3.29: TAN classifiers of disgust and surprise



(a) Male (b) Female

Figure 3.30: TAN classifiers of disgust and happiness



(a) Male (b) Female

Figure 3.31: TAN classifiers of surprise and happiness

3.5.3 Detecting Emotion Performance

We examined the emotion detection for our method. Table 3.3 shows the accuracy rates for emotion detection. The total accuracy rate of emotion detection was higher than 50%. The accuracy rates for male and female anger was higher than the other emotions, whereas those for male and female happiness, and male sadness were lower than the other emotions.

Table 3.4 shows the confusion matrix for detecting emotion. Many male sadness voice samples were mis-detected as disgust. Table 3.5 shows posterior probability of those sadness voice samples by sadness and disgust classifier for male. Sample 1 was detected as disgust because weighted probabilities of disgust were extremely high.

Table 3.3: Accuracy Rates of Emotion Detection

Emotion	Accuracy rates(%)		
	Male	Female	Total
Anger	**70.0**	**82.5**	**76.3**
Sadness	30.0	60.0	45.0
Disgust	47.5	47.5	47.5
Surprise	55.0	42.5	48.8
Happiness	37.5	40.0	38.8
Total	48.0	54.5	51.3

Table 3.4: Confusion Matrix

Gender	Actual Emotion	The number of classified samples					
		Anger	Sadness	Disgust	Surprise	Happiness	Total
Male	Anger	**28**	0	4	2	6	40
	Sadness	8	**12**	**10**	5	5	40
	Disgust	9	3	**19**	3	6	40
	Surprise	6	3	5	**22**	4	40
	Happiness	7	5	5	8	**15**	40
Female	Anger	**33**	1	1	3	2	40
	Sadness	8	**24**	3	4	1	40
	Disgust	6	9	**19**	3	3	40
	Surprise	3	7	7	**17**	6	40
	Happiness	5	6	3	**10**	**16**	40

Misdetection as anger by anger and sadness classifier caused inaccurate emotion detection on sample 2. However, many samples (sample 3 - sample 10) were mis-detected as disgust because sadness and disgust classifier detected those samples as disgust. FS for male's sadness and disgust classifier selected one feature (PW_9), which causes inaccurate emotion detection performance. This result indicates that we extracted few features for classification between sadness and disgust. Many female happiness voice samples also were classified as surprise. Table 3.6 shows posterior probability of those happiness voice samples by surprise and happiness classifier. Sample 1 was mis-detected as surprise because classifier of sadness and happiness detected that sample as sadness; sample 2 was mis-detected as surprise because classifier of disgust and happiness detected that sample as disgust; sample 3 was mis-detected as surprise because classifier of disgust and happiness detected that sample as disgust. However, many samples (sample 4 - sample 10) are mis-detected as surprise by surprise and happiness classifier for female. Gaussian distributions from selected 2 features shown in Fig-

Table 3.5: Posterior Probability by Sadness and Disgust Classifier

	Probability of sadness (%)	Probability of disgust (%)
Sample 1	**51.8**	48.2
Sample 2	**63.1**	36.9
Sample 3	48.6	**51.4**
Sample 4	40.1	**59.9**
Sample 5	35.6	**64.4**
Sample 6	37.6	**62.4**
Sample 7	35.3	**64.7**
Sample 8	46.7	**53.3**
Sample 9	31.8	**68.2**
Sample 10	27.0	**73.0**

Table 3.6: Posterior Probability by Surprise and Happiness Classifier

	Probability of happiness (%)	Probability of surprise (%)
Sample 1	**50.9**	49.2
Sample 2	**60.4**	39.6
Sample 3	**51.8**	48.2
Sample 4	47.5	**52.5**
Sample 5	44.5	**55.5**
Sample 6	39.6	**60.4**
Sample 7	48.6	**51.4**
Sample 8	46.4	**53.6**
Sample 9	38.7	**61.3**
Sample 10	34.6	**65.4**

ure 3.21 rarely distinguish between surprise and happiness. This result indicates that extracted features are not appropriate for classification between surprise and happiness. Therefore, we need to extract more acceptable features for classification between those pairs of emotions.

3.5.4 Comparing Results for Emotion Detections

To confirm the influence of feature selection and pairwise classification on emotion detection, we examined emotion detection by pairwise classification and multi-class classification for weak FS and without FS. Weak FS detects emotion with feature selection on all emotions, and selects subset features ($BW3_1$, $BW2_4$, PW_4, $F0_3$, PW_{11}, $F1_2$, $F4_2$, Tm for male, and PW_1, $F2_5$, $F2_8$, $F0_3$, $F2_4$, $F3_8$ for female). And without FS conducts emotion detection by all features. Table 3.7 shows the accuracy rates.

Table 3.7: Accuracy Rates of Emotion Detection Methods

Emotion	Accuracy rates(%)				
	Pairwise classification			Multi-class classification	
	Strong FS(Our method)	Weak FS	without FS	Weak FS	without FS
Anger	**76.3**	62.5	58.8	63.8	58.8
Sadness	**45.0**	30.0	26.3	32.5	26.3
Disgust	**47.5**	35.0	35.0	23.8	30.0
Surprise	**48.8**	35.0	31.3	42.5	33.8
Happiness	**38.8**	37.5	30.0	32.5	31.3
Total	**51.3**	40.0	36.3	39.0	36.0

The accuracy rates of weak FS were higher than their without FS on both pairwise classification and multi-class classification. This result indicates that feature selection has benefit for emotion detection performance. Total accuracy rate of strong FS (our method) were higher than the other methods. This result indicates that our method improves emotion detection performance from voice.

For comparison of pairwise classification methods for emotion detection, we detected emotion by simple voting system. The algorithm is as follows:

Algorithm 3.5.1 Simple voting

for $i = 1$ to k-1 **do**
  for $j = i + 1$ to k **do**
    Compute probability P_i, P_j using classifier C_{ij}
    if $P_i > P_j$ **then**
      $sum_i = sum_i + 1$ //simple voting system
    else if $P_i < P_j$ **then**
      $sum_j = sum_j + 1$ //simple voting system
    else if $P_i = P_j$ **then**
      Continue
    end if
  end for
end for
if One emotion is max win **then**
  Detect emotion as $\underset{c=1\cdots k}{\operatorname{argmax}}\, sum_c$
else
  Emotion detection is failure
end if

This method mis-detects emotion to equal amount of wins on more than two emo-

Table 3.8: Accuracy Rates of Pairwise Classifications

Emotion	Accuracy rates(%)		
	Our method	Simple voting	Sum of probability
Anger	**76.3**	72.5	70.0
Sadness	45.0	42.5	**47.5**
Disgust	**47.5**	41.3	41.3
Surprise	**48.8**	32.5	46.3
Happiness	**38.8**	31.3	37.5
Total	**51.3**	44.0	48.5

tions. Table 3.8 (middle) shows the accuracy rates. The accuracy rates of our method (Table 3.8 (left)) were quite higher than those of simple voting system. This result indicates that our method improves emotion detection performance.

We also detected emotion by sum of probability. The algorithm is as follows:

Algorithm 3.5.2 Sum of probability

for $i = 1$ to k-1 **do**
  for $j = i + 1$ to k **do**
    Compute probability P_i, P_j using classifier C_{ij}
    $w_i = P_i$
    $sum_i = sum_i + w_i$ //sum of probability
    $w_j = P_j$
    $sum_j = sum_j + w_j$ //sum of probability
  end for
end for
Detect emotion as $\underset{c=1\cdots k}{\operatorname{argmax}} sum_c$

This method adds all posterior probability from TAN classifiers. Table 3.8 (right) shows the accuracy rates. Total accuracy rate of our method was higher than sum of probability. The result indicates that our method has more acceptable accuracy rates.

3.6 Classification Results from Open Datasets

We confirmed classification performance of our method using open database from the University of California at Irvine (UCI) [54] machine learning repository [55]. These datasets are shown in Table 3.9. The datasets consist of multi-class for pairwise classification. We selected appropriate subset features on every pair of classes from all features. Then, we used selected subset features for learning each binary classifiers

Table 3.9: Description of Datasets

Dataset	# of training data	# of test data	# of classes	# of features
Iris	150	CV-5	3	4
Wine	178	CV-5	3	13
Wall-Follow	5456	CV-5	4	24
Breast Tissue	106	CV-5	6	9
Glass	214	CV-5	6	9
Segment	2310	CV-5	7	19
Vowel	528	462	11	10

Table 3.10: Accuracy Rates of Classification with FS

Dataset	Pairwise (Our method)	Multi-class
Iris	94.7	94.7
Wine	90.4	**94.9**
Wall-Follow	**62.7**	61.2
Breast Tissue	**69.8**	58.5
Glass	**55.1**	48.6
Segment	**90.5**	88.3
Vowel	**50.6**	50.0

using TAN. Then, we conducted classification using testing data. In regard to six datasets (except for Vowel), we conducted classification by five-fold cross validation from all training data.

Table 3.10 shows the accuracy rates of classification with FS. FS of multi-class is conducted on all classes. In regard to five datasets (Wall-Follow, Breast Tissue, Glass, Segment, and Vowel), our method has the higher accuracy rates than multi-class classification. However, our method from Wine had the lower accuracy rate than other, and Iris had the same accuracy rate for two methods. Table 3.11 also shows the accuracy rates of pairwise classification. Our method from four datasets (Wall-Follow, Breast Tissue, Glass, and Segment) had the higher accuracy rates than without FS, whereas our method from other datasets (Iris, Wine, and Vowel) had the lower accuracy rates than without FS. Therefore, our method is appropriate to the four datasets having four to seven classes (Breast Tissue, Glass, Segment and Wall-Follow). However, our method shows poor performance to the datasets having three classes (Iris and Wine) and many classes (Vowel).

Table 3.11: Accuracy Rates of Pairwise Classification

Dataset	FS (Our method)	without FS
Iris	94.7	**96.7**
Wine	90.4	**96.1**
Wall-Follow	**62.7**	51.7
Breast Tissue	**69.8**	58.5
Glass	**55.1**	45.3
Segment	**90.5**	79.3
Vowel	50.6	**58.0**

3.7 Conclusion

This chapter proposed a method for detecting emotion from human voice. The method provided pairwise classification using selective TAN of acoustic features. Our method improves emotion detection performance for freely uttered voice samples from many and unspecified actors and actresses. We also reported on classification performance from open database.

We proposed detecting emotion method in Chapter 2 and Chapter 3. Unfortunately, considering the practical constraints of human robot interaction, voice analysis can not be fully guaranteed. In the future work, we will propose more precise detecting emotion method. Finally, we will aim to apply detecting engine system to a robot communication system.

Chapter 4

Comparison of Sensibilities of Japanese and Korean

4.1 Introduction

Recently, the world's communities are growing more and more inter-dependent. This trend has promoted economic cooperation, foreign trade, immigration, etc. In light of this trend, there has been a large increase in opportunities for cross-cultural exchange, and thus, mutual understanding of emotions in order to comprehend or sympathize with speakers of different languages is becoming even more important.

Besides vocal expression, facial expression is an important element of emotional expression. According to Paul Ekman's study [9] of human faces, people in many different cultures innately share facial expressions conveying six basic emotions (anger, sadness, disgust, fear, surprise, and happiness). With respect to emotional expression of the human voice, however, there has been almost no research comparing vocal emotional expressions of people who speak different languages and have different cultural backgrounds. Although there are several reports on detecting emotions in human voice (e.g., [56, 57]), they deal with a specific language. We chose Japan and Korea as the two different cultures for our study. The grammars of Japanese and Korean show some similarity: they have the same word order, and nominatives are indicated by particles. On the other hand, speakers of these languages can't understand the other language because they have different phonologies, vocabularies, and writings.

In Chapter 2 and Chapter 3, we focused on a method for detecting emotion. We have previously presented a Bayesian network-based method for detecting emotions in human voices. The method focuses on acoustic features in emotionally expressive human voices and models the causal relationship between emotions and the features

by using a BN. The BN built by our method can detect emotions from human voices expressing non-verbal information.

In this chapter, we model the sensibilities that Japanese and Korean have for emotional voices by learning BNs that can detect emotions in emotional voices of native Japanese and Koreans [58, 59]. We then compare the sensibilities of emotion recognition from speech between Japanese and Korean by examining the cross-inference through two BNs with speech in the respective foreign language.

4.2 Constructing Detecting Emotion Engine

We focus on the acoustic features of the speaker's voice as a cue to what emotion he or she expresses. This section describes a BN modeling for this problem.

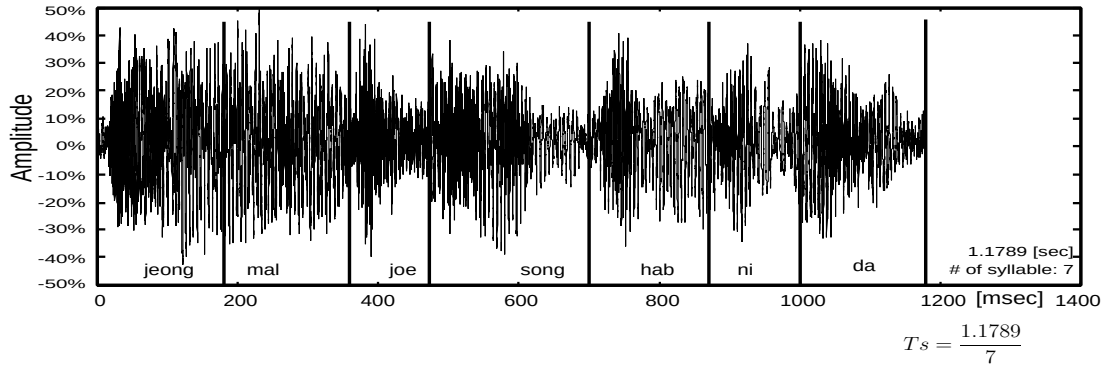
4.2.1 Voice Data

We collected segments of Japanese and Korean voice samples that were spoken emotionally by actors and actresses in films, TV dramas, and so on. We labeled all segments with five emotional labels (anger, sadness, disgust, surprise or happiness). We extracted voice samples from many and unspecified actors and actresses, and Japanese and Korean collected sound data in their native language.

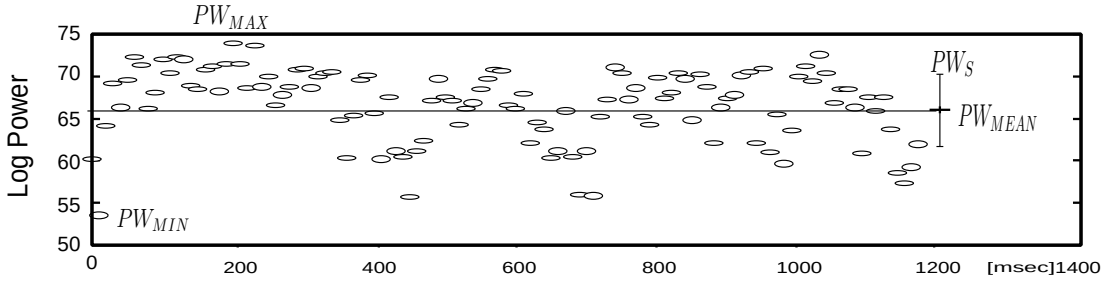
4.2.2 Features Extraction

In this chapter, we chose three acoustic attributes: energy, fundamental frequency and duration as the acoustic parameters for BN modeling. Acoustic analysis was done on 11 ms frames passed through a Hamming window extracted from voice waveforms sampled at 22.05 kHz.

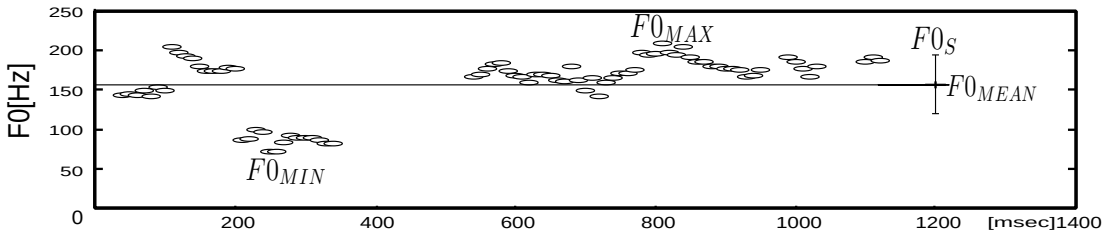
Figure 4.1 shows an example of feature extraction from voice data of a Korean male speaking the Korean sentence "jeong-mal-joe-song-hab-ni-da" (that means "I am so sorry"). The attributes of energy, maximum energy (PW_{MAX}), minimum energy (PW_{MIN}), mean energy (PW_{MEAN}) and its standard deviation (PW_S) are determined from the energy contours for the frames in a voice waveform (see Figure 4.1 (b)). The attributes of fundamental frequency, maximum pitch ($F0_{MAX}$), minimum pitch ($F0_{MIN}$), mean pitch ($F0_{MEAN}$) and its standard deviation ($F0_S$) are determined from short time Fourier transforms for the frames in a voice waveform (see Figure 4.1 (c)). As the attribute concerning duration, we measure the average duration per a single



(a) Amplitude



(b) Energy



(c) Fundamental frequency

Figure 4.1: An example of attribute extracted from a Korean male's voice "jeong-mal-joe-song-hab-ni-da": (a) the speech waveform of "jeong-mal-joe-song-hab-ni-da" and number of word's syllable where continuous lines is syllable boundaries derived on a phonological basis, (b) a plot of energy extracted from the energy contours for the frames in the waveform, (c) a plot of fundamental frequency extracted by short time Fourier transforms for the frames in the waveform.

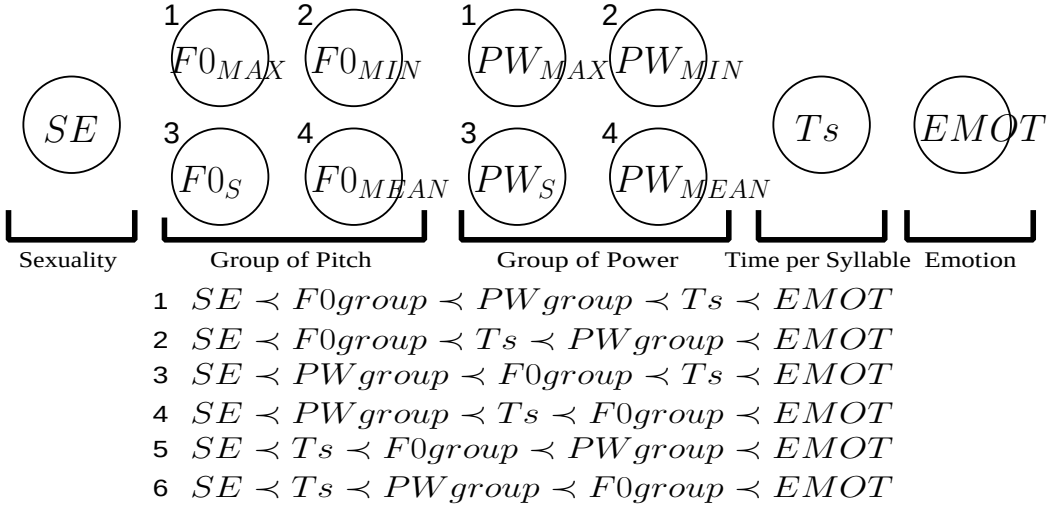


Figure 4.2: Possible variable orders of node groups

syllable (Ts) (see Figure 4.1 (a)). Then we added the attribute of the speaker's sexuality (SE). The goal attribute ($EMOT$) and the above nine acoustic feature values and speaker's sexuality (total eleven attributes) were assigned to the nodes of the BN model.

4.2.3 Discretization of Feature

The section describes the discretization of extracted acoustic features. We considered BNs with discrete and multinominal variables only. In order to learn the discrete causal structure of the BN model, all acoustic features were converted into discrete values. The thresholds to discretize continuous values were determined from the distribution of the acoustic features extracted from the training voice samples.

4.2.4 Learning BN Structure

The section describes how to specify the topology of the BN model for emotion detection and to parametrize CPT for connected nodes. The emotion detection BN modeling is to determine the qualitative and quantitative relationships between the output node containing the goal attribute (emotions) and nodes containing acoustic features. We chose a model selection method based on the Bayesian information criterion (BIC) [38], and we used K2 as the search algorithm. We described construction of BN model using K2 algorithm based on BIC scoring in Section 2.2.4. We considered every possible permutation of three node groups: PW, $F0$ and Ts, such that node SE

preceded all others shown in Figure 4.2.

The BN has no verbal information: it only has acoustic features and speaker's sexuality. With respect to the native language, verbal information is often dominant in the emotion recognition from speech, but it is of no use for the foreign language. We used BNs composed of non-verbal information to enable a pure comparison of native and foreign languages.

4.3 Inference Algorithm

The topology of the BN is often multiply connected when there is a complicated relationship between variables. We chose junction tree [22] as the inference algorithm with BNs, it is an exact inference algorithm in multiply connected BNs. It is efficient clustering inference algorithms. Clustering inference algorithms transform the BN into a probabilistically equivalent polytree by merging nodes and removing multiple paths between two nodes along which evidence may travel.

4.4 Comparison of Sensibilities through Emotion Inference

The section describes an experiment comparing the sensibilities of emotion recognition of Japanese and Koreans by using the Bayesian approach. Figure 4.3 shows an overview of the experiment. First, we collected 500 segments of voice waveforms in Japanese and Korean (1000 segments in total) and labeled them with five emotions, as described in Section 4.2.1. Then, we extracted nine acoustic features and the speaker's sexuality in each of the segments and assigned them to the attributes, as described in Section 4.2.2. The acoustic analysis used the Snack sound toolkit [40]. After that we randomly selected 400 samples from Japanese and Korean (800 in total) as training data and discretized their attributes into four values. We determined the threshold for discretization on the basis of the idea of even-sized chunks; that is, each discrete value covers 25% of the training data. We then modeled the BNs for Japanese and Korean by changing six variable orders (see Figure 4.2) with Bayes Net Toolbox [41].

Figure 2.6 shows the results of learning: BN for Japanese (BN_{JP}) and BN for Korean (BN_{KR}) with the variable order $SE \prec F0_{MAX} \prec F0_{MIN} \prec F0_S \prec F0_{MEAN} \prec PW_{MAX} \prec PW_{MIN} \prec PW_S \prec PW_{MEAN} \prec Ts \prec EMOT$, where both the BNs have the highest accuracy rates for their native voice samples. First, we compared the

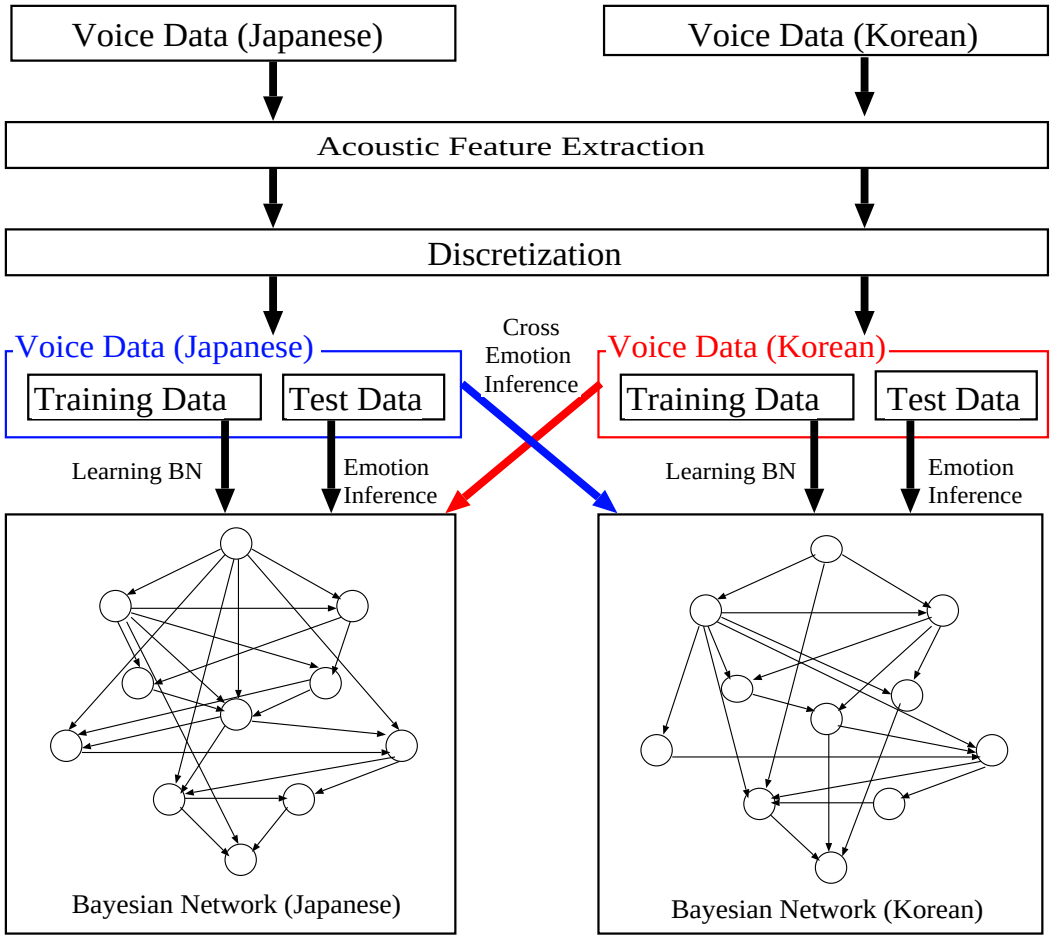
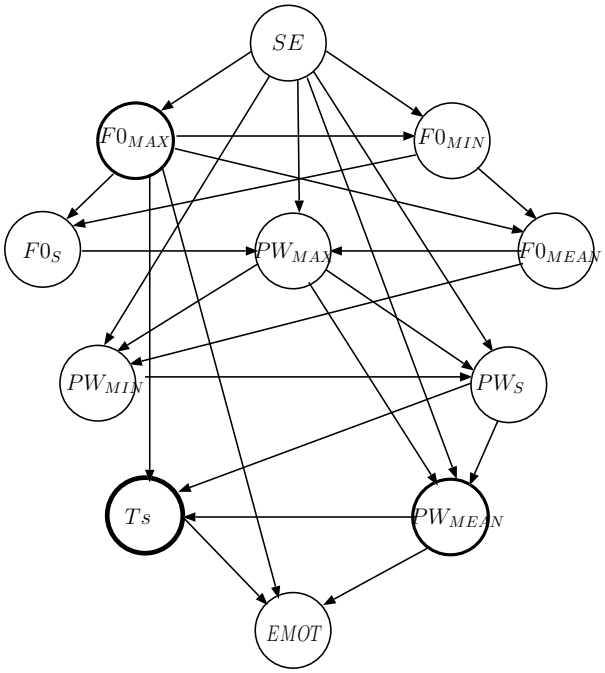


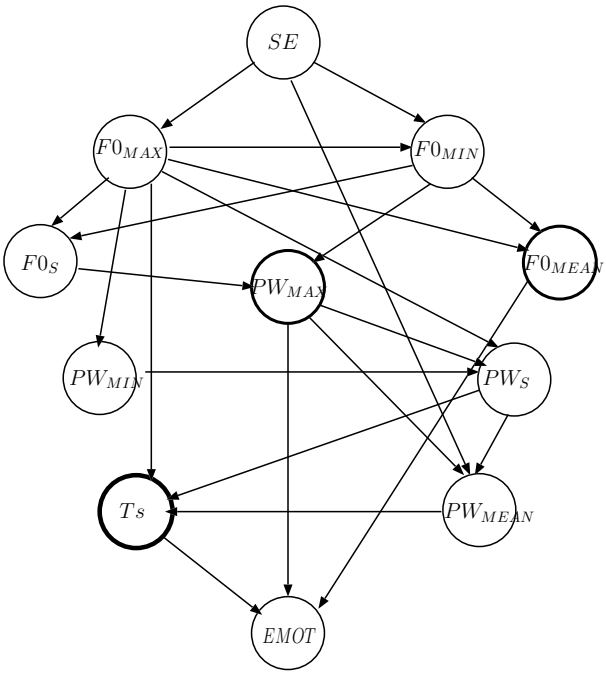
Figure 4.3: Cross-inference for emotion detection from Japanese and Korean speech data

parent nodes of $EMOT$ between the BNs because these nodes strongly influence emotion inference. The parent nodes of $EMOT$ are $F0_{MAX}$, PW_{MEAN}, and Ts in BN_{JP} and are $F0_{MEAN}$, PW_{MAX}, and Ts in BN_{KR}. The results indicate roughly that the three acoustic features ($F0$, PW, Ts) are largely related with emotion inference in Japanese and Korean. Concerning the fundamental frequency ($F0$), Japanese sensibilities depend on the maximum value and Korean sensibilities depend on the average value. Concerning energy (PW), Japanese sensibilities depend on the average value and Korean sensibilities depend on the maximum value.

We then conducted two experiments on emotion inference: detecting emotions from native speech and from foreign speech. The first experiment attempted to confirm the effectiveness of the two BNs as sensibility models of Japanese and Korean. The second experiment was to enable a comparative discussion of sensibilities between Japanese and Koreans.



(a) A BN_{JP} learned from Japanese voices



(b) A BN_{KR} learned from Korean voices

Figure 4.4: BN structure learned from training data

Table 4.1: Accuracy Rates of Emotion Inference in the Native Language

		Accuracy Rates [%]			
		BN		PCA	
Language		Japanese	Korean	Japanese	Korean
Emotion	Anger	70.0	65.0	90.0	30.0
	Sadness	55.0	75.0	10.0	35.0
	Disgust	60.0	60.0	50.0	50.0
	Surprise	40.0	50.0	50.0	10.0
	Happiness	50.0	50.0	5.0	30.0
Total		55.0	60.0	41.0	31.0

Table 4.2: Accuracy Rates of Emotion Inference in the Foreign Language

		Accuracy Rates [%]	
BN		BN_{JP}	BN_{KR}
Testdata		Korean	Japanese
Emotion	Anger	55.0	22.0
	Sadness	9.0	20.0
	Disgust	29.0	33.0
	Surprise	18.0	10.0
	Happiness	31.0	29.0
Total		28.4	22.8

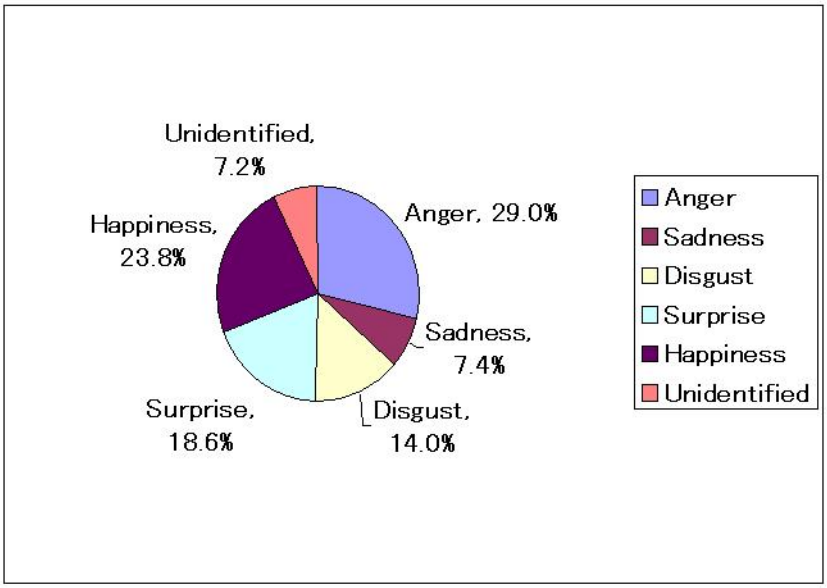
4.4.1 Emotion Inference from Voices in the Native Language

We converted acoustic features of the remaining 100 samples in Japanese and Korean (200 in total) into discrete values by using the same thresholds for the training data and then examined the inference performance of each of the BN models shown in Figure 2.6. The left side of Table 4.1 shows the results. The BNs had accuracy rates of inference higher than 50%, except for Japanese surprise, and the total accuracy rates were higher than 55% in both Japanese and Korean.

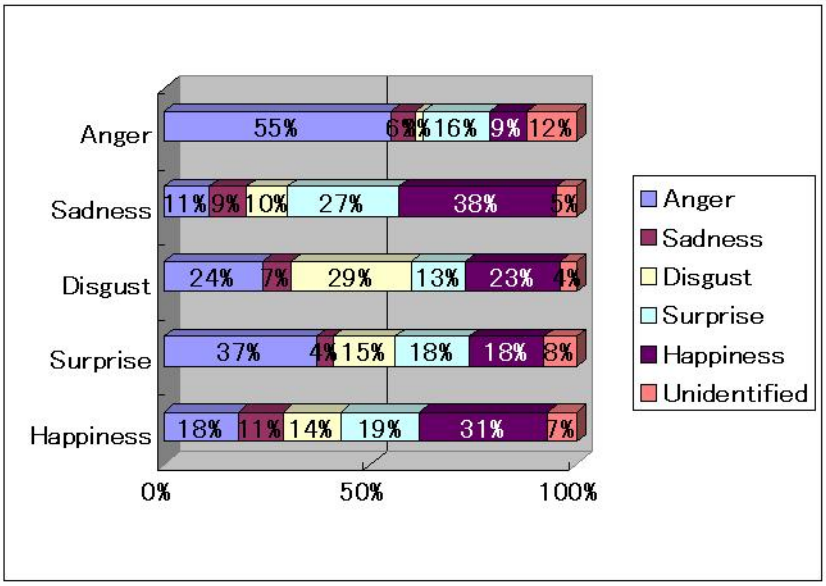
For comparison, we used principal component analysis (PCA) using ten acoustic features of training data and a classification based on the linear discriminant in a four-dimensional hyper-plane using four PCs because the accumulated contribution relevance was more than 70%. The right side of Table 4.1 shows the results. The BN for Korean had accuracy rates higher than the PCA for all emotions. The BN for Japanese had accuracy rates higher than the PCA, except for anger and surprise. The

PCA for Japanese had very high accuracy rate for anger. To infer a specific emotion accurately is, however, totally insignificant, unless the BN can adequately infer all other emotions. The emotion inference has to get high average accuracy rates for all emotions. Note that the BNs for Japanese and Korean have total accuracy rates higher than the PCAs.

From now on, our discussion will proceed under the assumption that the BNs for Japanese and Korean reflect their sensibilities of emotion recognition from voices. With respect to each emotion, both BNs had high accuracy rates for anger. The BN for Korean had high accuracy rate for sadness as well. Both BNs had lower accuracy rates for surprise and happiness. These results suggest that Japanese and Koreans easily recognize anger and Koreans easily recognize sadness in their own native speech, and that it is slightly difficult for Japanese and Korean to recognize surprise and happiness in their own native speech.

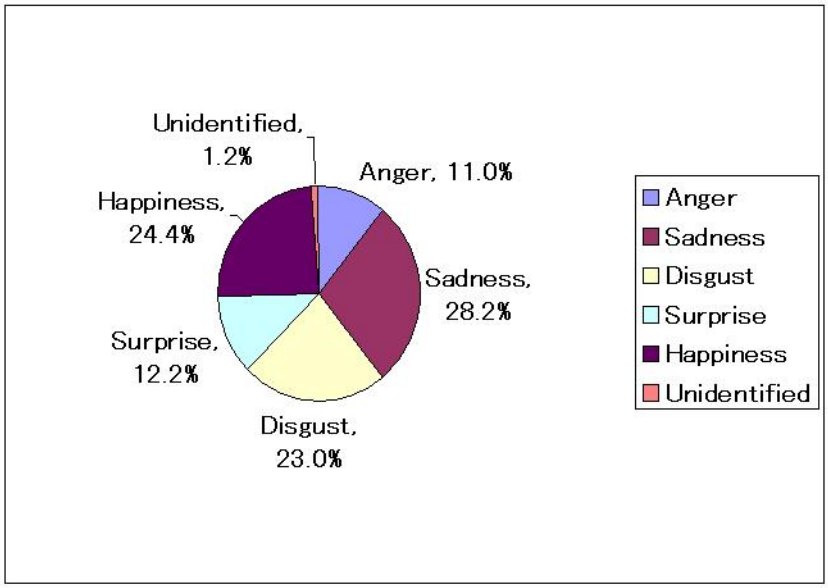


(a) Inference rates

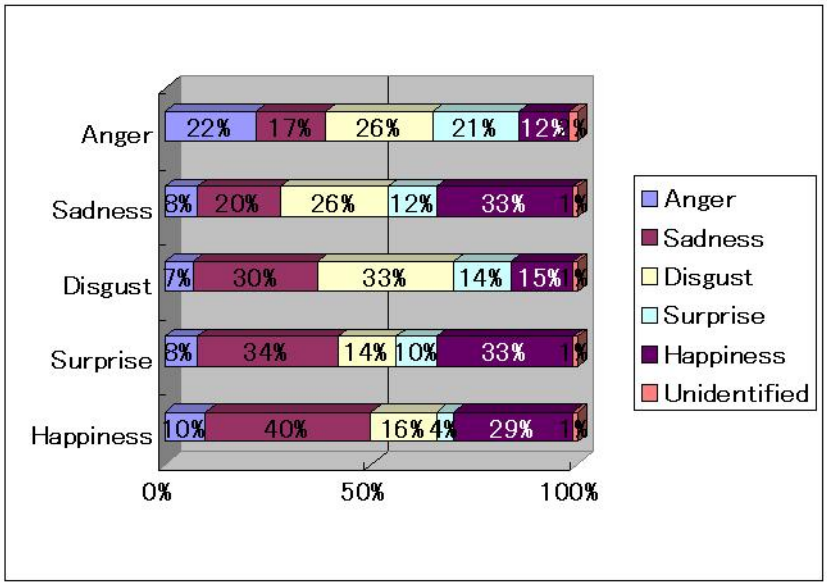


(b) Detailed inference rates of each emotion

Figure 4.5: Detailed emotion inference rates for Korean voice using BN_{JP}



(a) Inference rates



(b) Detailed inference rates of each emotion

Figure 4.6: Detailed emotion inference rates for Japanese voice using BN_{KR}

4.4.2 Emotion Inference from Voices in a Foreign Language

To compare the sensibilities of Japanese and Koreans in recognizing emotions from voices, we examined the cross-emotion inference through two BNs with speech in the respective foreign language. The cross-emotion inference was done by giving 500 Korean voice samples to BN_{JP} and giving 500 Japanese voice samples to BN_{KR}. Table 4.2 shows the resulting accuracy rates. The accuracy rates for the respective foreign languages (see Table 4.2) are lower than those for the native languages (see Table 4.1). BN_{JP}'s accuracy rate for Korean anger was higher than 50%. This result suggests that Japanese sensibilities can easily recognize anger from Korean speech. BN_{KR}'s accuracy rate for Japanese disgust was higher than those of other emotions. The result suggests that Korean sensibilities can fairly easily recognize disgust from Japanese speech. With respect to sadness and surprise, both BN_{JP} and BN_{KR} had accuracy rates not more than 20%. The results suggest that it is difficult for Japanese and Koreans to recognize sadness and surprise from the other's speech.

Figure 4.5 (a) shows the inference rates of emotions using BN_{JP} with Korean voice samples. For example, 145 Korean voice samples (29%) are recognized as anger. The figure indicates that BN_{JP} recognizes most Korean voices as expressing anger or happiness and a few Korean voices as expressing sadness or disgust. Figure 4.5 (b) shows the detailed inference rates of each emotion. For example, 37 surprised voice samples (37%) are recognized as anger. The figure indicates that most Korean angry voices are recognized correctly but that many Korean surprised voices are mis-recognized as angry. The figure also indicates that many Korean sad voices are mis-recognized as happy. These results suggest that Japanese sensibilities often recognize Korean voices as angry and they often mis-recognize Korean surprise and sadness as anger and happiness, respectively.

Figure 4.6 (a) shows the inference rates for BN_{KR} on Japanese voice samples. For example, 141 samples (28.2%) from Japanese voice samples are recognized as sad. The figure indicates that most Japanese voices are recognized as sad or happy and a few Japanese voices are recognized as angry or surprised. Figure 4.6 (b) shows the detailed inference rates of each emotion. For example, 40 happy voice samples (40%) are recognized as sad. The figure indicates that lots of Japanese happy voices and surprised voices are mis-recognized as sad and lots of Japanese sad and surprised voices are mis-recognized as happy. These results suggest that Korean sensibilities often recognize Japanese voices as expressing sadness or happiness and they often mis-

Table 4.3: Itemization of the Discretized Feature Values of Korean Voice Samples

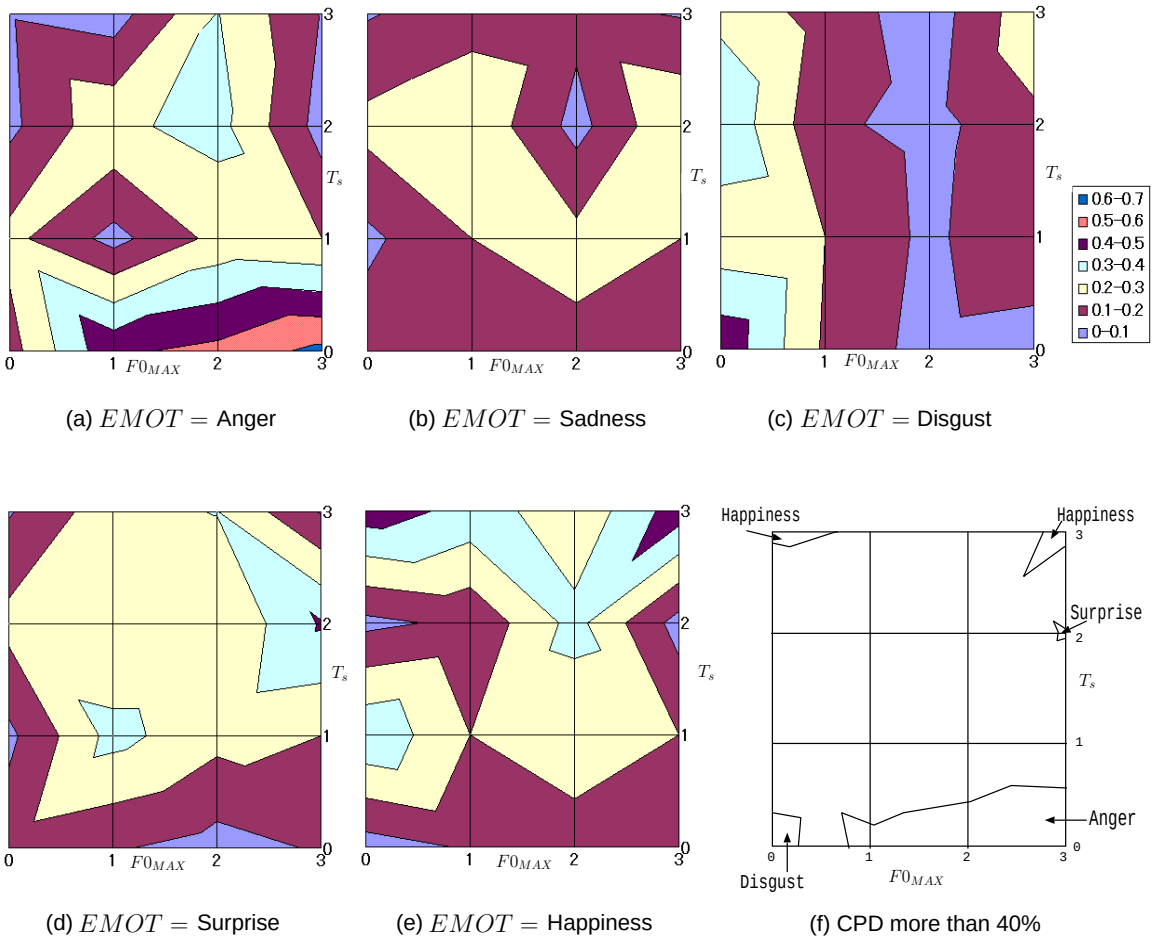
		The number of data											
Feature		$F0_{MAX}$				PW_{MEAN}				Ts			
Value		0	1	2	3	0	1	2	**3**	0	1	2	3
Emotion	Anger	0	19	32	49	3	4	19	**74**	48	31	16	5
	Sadness	13	27	18	42	12	13	21	**54**	6	25	27	42
	Disgust	55	33	9	3	8	20	26	**46**	33	45	14	8
	Surprise	19	24	22	35	7	10	18	**65**	38	37	15	10
	Happiness	23	33	22	22	5	8	30	**57**	20	33	26	21
Subtotal		110	136	103	151	35	55	114	**296**	145	171	98	86
Total		500				500				500			

Table 4.4: Itemization of the Discretized Feature Values of Japanese Voice Samples

		The number of data											
Feature		$F0_{MEAN}$				PW_{MAX}				Ts			
Value		0	1	2	3	**0**	1	2	3	0	1	2	3
Emotion	Anger	13	31	29	27	**33**	30	30	7	40	28	16	16
	Sadness	38	24	27	11	**57**	15	18	10	15	22	22	41
	Disgust	46	39	11	4	**63**	25	10	2	38	13	30	19
	Surprise	20	31	28	21	**57**	19	15	9	8	15	23	54
	Happiness	25	31	18	26	**58**	18	18	6	10	14	22	54
Subtotal		142	156	113	89	**268**	107	91	34	111	92	113	184
Total		500				500				500			

recognize Japanese happiness and surprise as sadness. These results also suggest that Korean sensibilities often mis-recognize Japanese sadness and happiness.

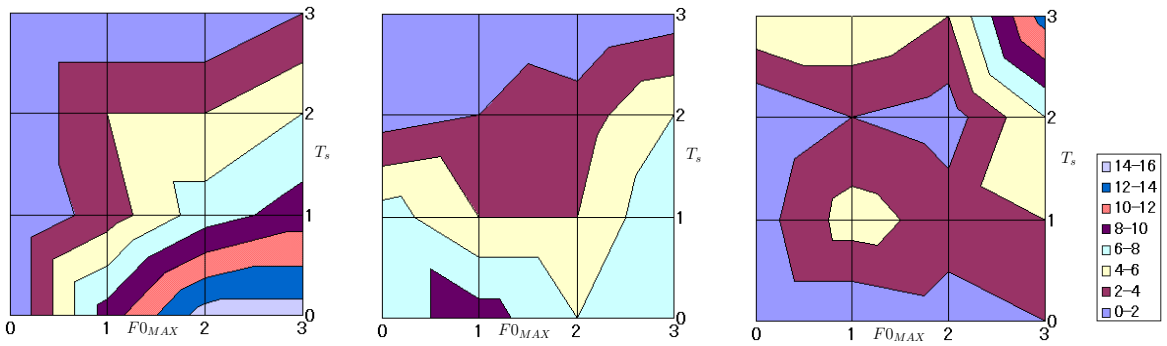
We investigated the causal relation in BN_{JP} with Korean voice. We focus on the relationship between $EMOT$ and its parent nodes $F0_{MAX}$, PW_{MEAN}, and Ts. Table 4.3 shows the itemization of the discretized feature values of Korean voice samples. For visualization in two dimensions, we selected Korean voice samples such that $PW_{MEAN} = 3$ because they had the most of voice samples. Figure 4.7 shows the conditional probability distribution (CPD) of $EMOT$ each emotion on $F0_{MAX}$, Ts, and $PW_{MEAN} = 3$. According to the CPD in BN_{JP}, Japanese sensibilities probably recognize a speech voice as anger when $F0_{MAX}$ is not low and Ts is fast; they rarely recognize it as sadness; they probably recognize it as disgust when $F0_{MAX}$ is low and Ts is fast; they probably recognize it as surprise when $F0_{MAX}$ is high and Ts is a little slow; they probably recognize it as happiness when $F0_{MAX}$ is low or high and Ts is slow. Figure 4.7 (f) shows the CPD where $\mathrm{P}(EMOT \mid F0_{MAX}, Ts, PW_{MEAN} = 3) \geq 0.4$. Figure 4.8 shows the frequency distribution of Korean voice samples of anger, surprise,



(a) $EMOT =$ Anger (b) $EMOT =$ Sadness (c) $EMOT =$ Disgust

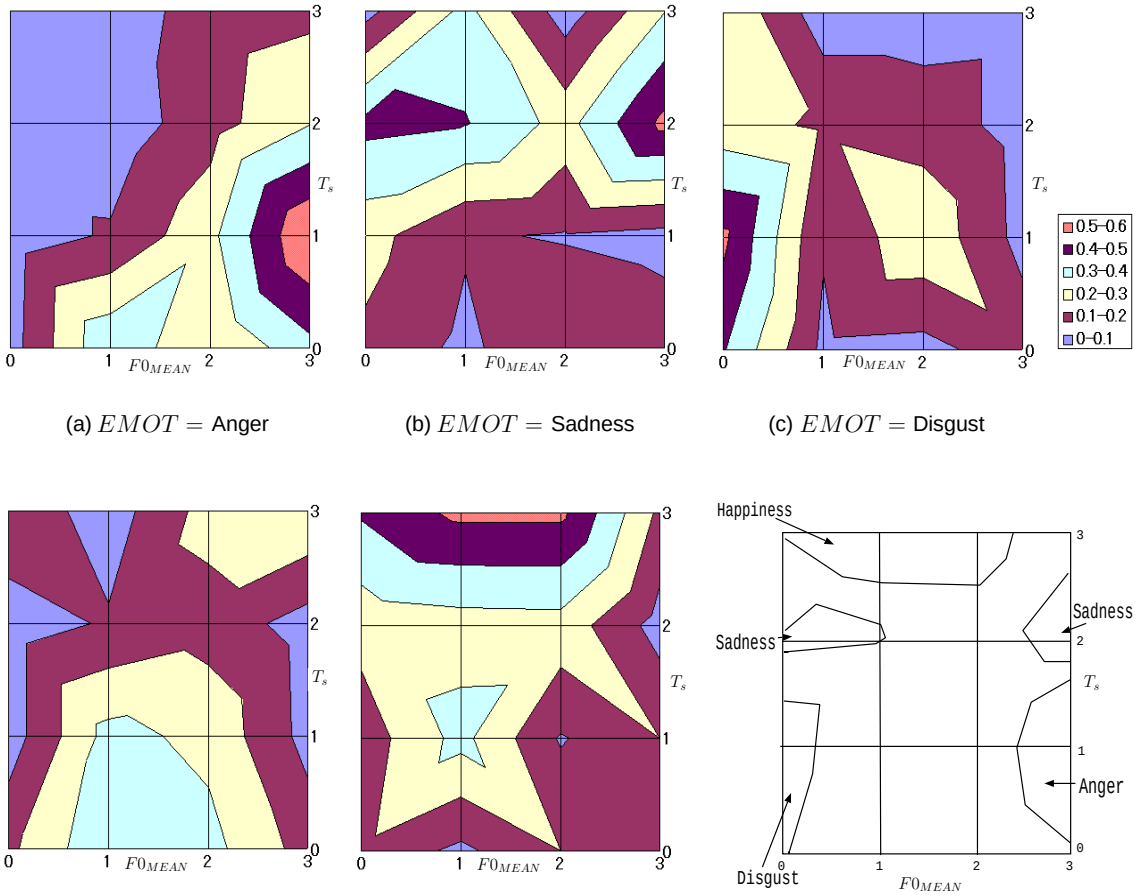
(d) $EMOT =$ Surprise (e) $EMOT =$ Happiness (f) CPD more than 40%

Figure 4.7: CPD of P($EMOT \mid F0_{MAX}, Ts, PW_{MEAN} = 3$) in BN_{JP}



(a) Anger (b) Surprise (c) Sadness

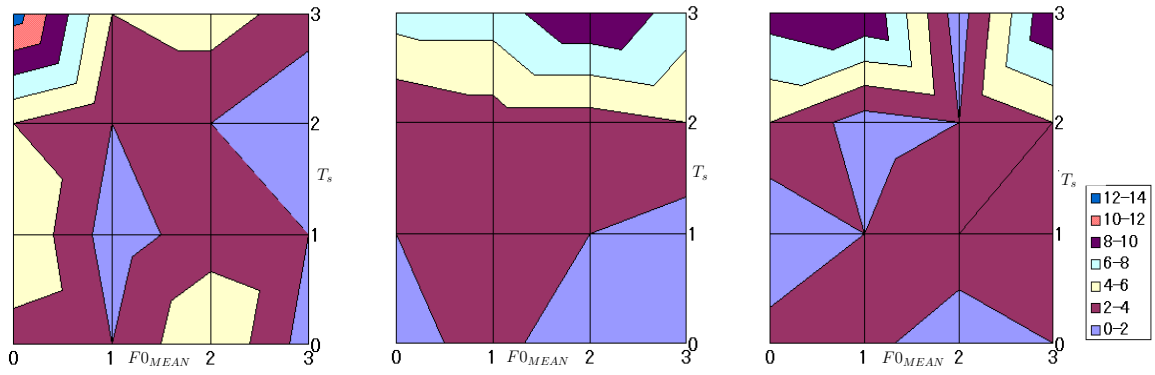
Figure 4.8: The frequency distribution of Korean voice samples in $F0_{MAX}$ and Ts when $PW_{MEAN} = 3$



(a) $EMOT$ = Anger (b) $EMOT$ = Sadness (c) $EMOT$ = Disgust

(d) $EMOT$ = Surprise (e) $EMOT$ = Happiness (f) CPD more than 40%

Figure 4.9: CPD of P($EMOT \mid F0_{MEAN}, Ts, PW_{MAX} = 0$) in BN_{KR}



(a) Sadness (b) Surprise (c) Happiness

Figure 4.10: The frequency distribution of Japanese voice samples in $F0_{MEAN}$ and Ts when $PW_{MAX} = 0$

and sadness in $F0_{MAX}$ and Ts when $PW_{MEAN} = 3$. It follows from Figure 4.7 (f) and Figure 4.8 that most of Korean angry voices are recognized correctly by BN_{JP} because most of them range where $F0_{MAX}$ is not low and Ts is fast; lots of Korean surprised voices are mis-recognized as anger by BN_{JP} because lots of them range where $F0_{MAX}$ is a little low and Ts is fast; lots of Korean sad voices are mis-recognized as happiness by BN_{JP} because most of them range where $F0_{MAX}$ is high and Ts is slow.

We then investigated the causal relation in BN_{KR} with Japanese voice. We focus on the relationship between $EMOT$ and its parent nodes $F0_{MEAN}$, PW_{MAX}, and Ts. Table 4.4 shows the itemization of the discretized feature values of Japanese voice samples. For visualization in two dimensions, we selected Korean voice samples such that $PW_{MAX} = 0$ because they had the most of voice samples. Figure 4.9 shows the CPD of $EMOT$ each emotion on $F0_{MEAN}$, Ts, and $PW_{MAX} = 0$. According to the CPD in BN_{KR}, Korean sensibilities probably recognize a speech as anger when $F0_{MEAN}$ is high and Ts is a little fast; they probably recognize it as sadness when $F0_{MEAN}$ is low or high and Ts is a little slow; they probably recognize it as disgust when $F0_{MEAN}$ is low and Ts is fast; they rarely recognize it as surprise; they probably recognize it as happiness when $F0_{MEAN}$ is not high and Ts is slow. Figure 4.9 (f) shows the CPD where $\mathrm{P}(EMOT \mid F0_{MEAN}, Ts, PW_{MAX} = 0) \geq 0.4$. Figure 4.10 shows the frequency distribution of Japanese voice samples of sadness, surprise, and happiness in $F0_{MEAN}$ and Ts when $PW_{MAX} = 0$. It follows from Figure 4.9 (f) and Figure 4.10 that lots of Japanese sad voices are mis-recognized as happiness because lots of them range where $F0_{MEAN}$ is low and Ts is slow; lots of Japanese surprised voices or happy voices are mis-recognized as sadness or happiness because lots of them range where Ts is slow.

These investigations support the results of cross emotion inference shown in Figure 4.5 and Figure 4.6.

4.5 Conclusion

We compared sensibility of recognizing emotions of voice in different cultures based on a Bayesian approach. We modeled the sensibilities of Japanese and Korean by constructing BN from emotional voice. We used K2 algorithm as a method of BN construction. We then compared the sensibilities of Japanese and Koreans, by examining the cross-inference using two BNs with speech in the respective foreign language. Therefore, the experimental results showed that Japanese and Korean use different emotion expressions in voice.

In future work, we will dedicate to the improvement of emotion inference performance for native languages and propose the system that recognizes emotion from voice of different language. Finally, we will aim to develop a support robot for smooth communication with people who have different culture and language.

Chapter 5

Conclusion

This paper focused on Bayesian approach as emotion detection method from voice , and examined emotion detection on native and foreign language.

In Chapter 1, we described related researches of detecting emotion from voice. Then we described emotion expression and extraction from voice, necessary for detecting emotion in the human communication, and communication between people of different culture background. Then, we described BN that is used for detecting emotion method.

In Chapter 2, we proposed BN method for emotional communication robot that detects human emotion. The method is BN model using K2 algorithm, that acquires emotion content from acoustic feature. BN model enabled robots to detect emotions by using probabilistic inference from incomplete evidence and reasoning under uncertainty. In addition, we confirmed influence of specific acoustic features on detection of each emotion.

In Chapter 3, we also proposed pairwise classification for detecting emotion method from human voice. This method is pairwise classification using TAN. Each binary classifier used subset features that are selected on each pair of emotion. Detecting emotion is done by the maximum of weighted probability that is posterior probability of classified emotion from each TAN classifier. In summary, we used a specialized series of binary classifiers, and therefore, we can acquire more acceptable emotion detection performance. In addition, we confirmed that our method is relevant to partial open databases.

In Chapter 4, we compared sensibility of recognizing emotions of speech in different cultures based on a Bayesian approach. We chose Japan and Korea as two different cultures, and modeled the sensibilities that Japanese and Koreans have for emotional voices. We used BN as a method of sensibility modeling. We then compared the sensibilities of Japanese and Koreans, by examining the cross-inference using two BNs

with speech in the respective foreign language. From the experiment, we found that Japanese recognize a lot of Korean voices as expressing anger and Korean recognize a lot of Japanese voices as expressing sadness. These results partially corresponds to the national sensibilities of Japanese and Koreans.

Overall, this study proposed detecting emotion method from acoustic features of voice using Bayesian network model, and compared sensibilities of recognizing emotion between Japanese and Korean using BN model. The conclusions which can be drawn from this study are these. Firstly, BN is a available method for emotional communication robot that detects human emotion. We confirmed that BN can perform emotion detection from partial evidence as mentioned in Chapter 2. Secondly, pairwise classification using a specialized series of BN classifiers has a benefit for multi-class classification from diverse data. We confirmed that our method proposed in Chapter 3 obtains more accurate detecting emotion result from voice spoken by unspecified people. Lately, from an examination of Chapter 4, we confirmed possibility of system development that detects emotion between different cultures.

In future works, we will dedicate to the BN modeling of the verbal and facial information for emotion detection, and describe the Bayesian mixture approach: by using a mixture Bayesian networks of voice, verbal, and facial expression. Then, we will aim to implement the emotion detection engine in communication robot. Finally, we believe that this study lays the foundation for future work on the emotional communication robot.

Appendix A

Second Phase BN Model in Section 2

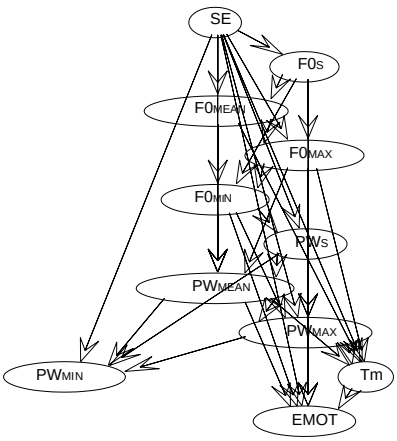


Figure A.1: Second phase BN model of anger and sadness

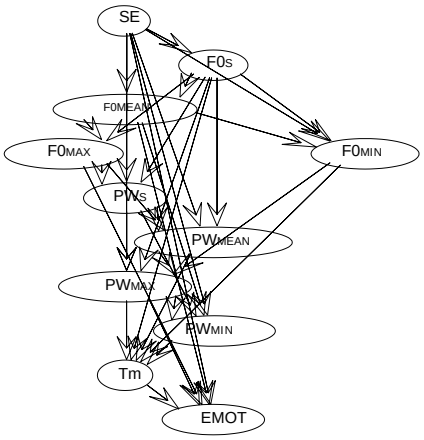


Figure A.2: Second phase BN model of anger and disgust

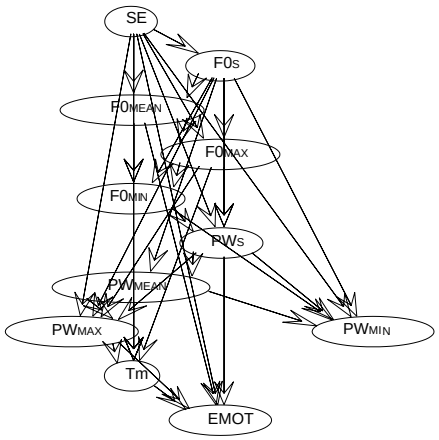


Figure A.3: Second phase BN model of anger and fear

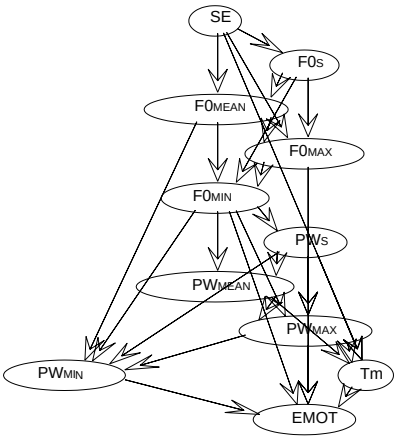


Figure A.4: Second phase BN model of anger and surprise

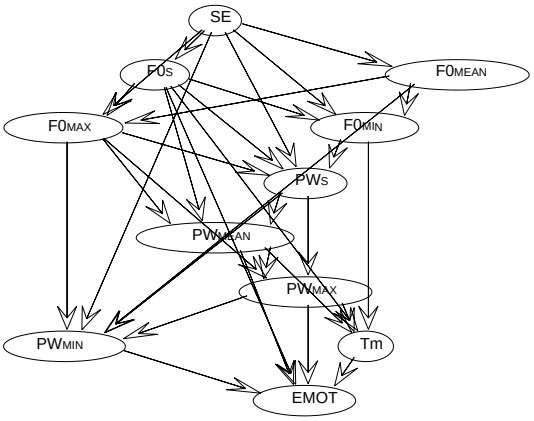


Figure A.5: Second phase BN model of anger and happiness

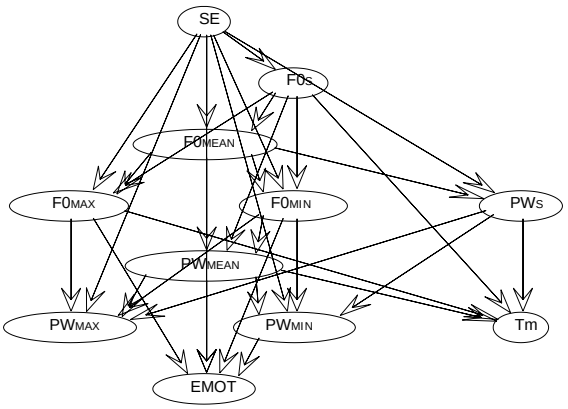


Figure A.6: Second phase BN model of sadness and disgust

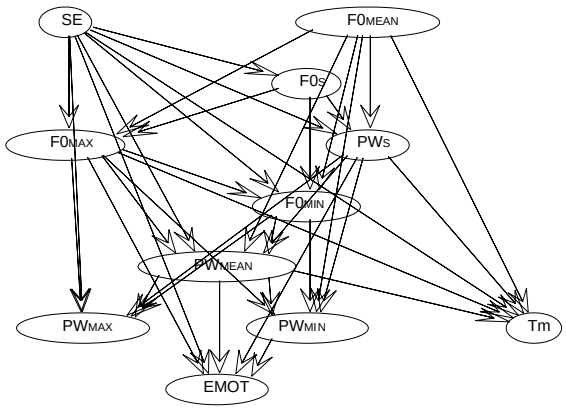


Figure A.7: Second phase BN model of sadness and fear

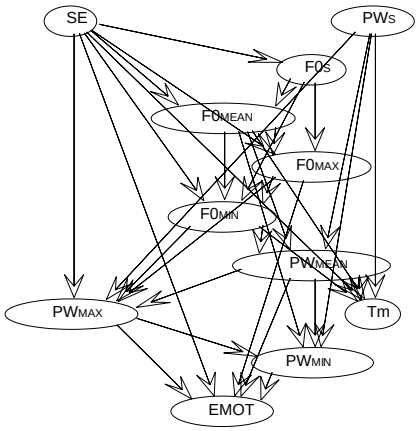


Figure A.8: Second phase BN model of sadness and surprise

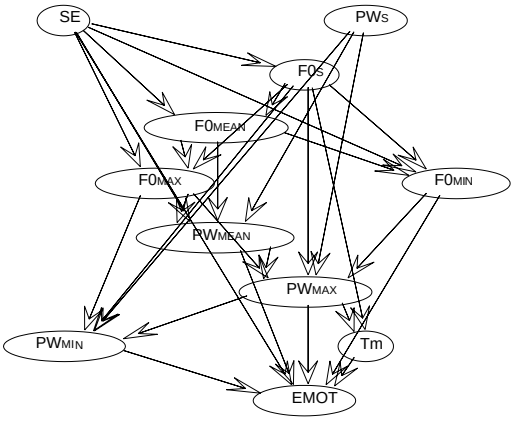


Figure A.9: Second phase BN model of sadness and happiness

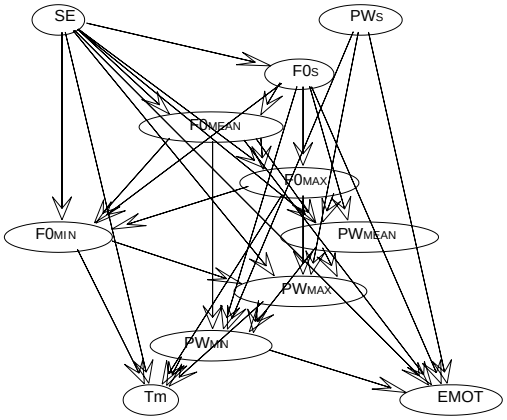


Figure A.10: Second phase BN model of disgust and fear

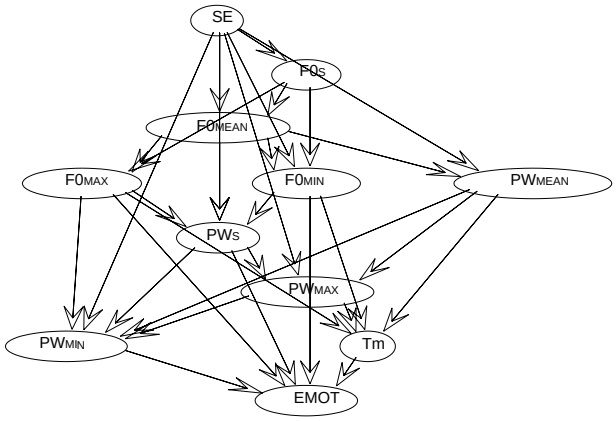


Figure A.11: Second phase BN model of disgust and surprise

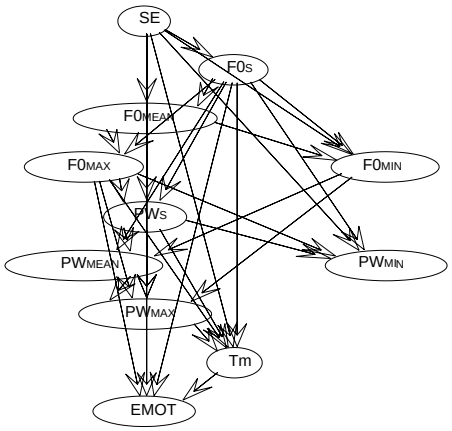


Figure A.12: Second phase BN model of disgust and happiness

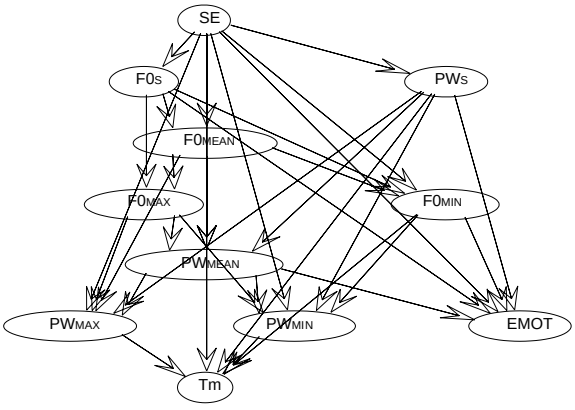


Figure A.13: Second phase BN model of fear and surprise

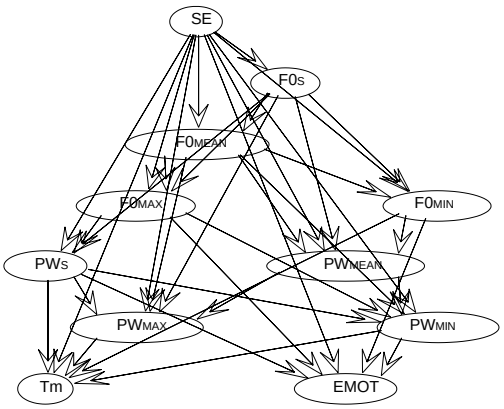


Figure A.14: Second phase BN model of fear and happiness

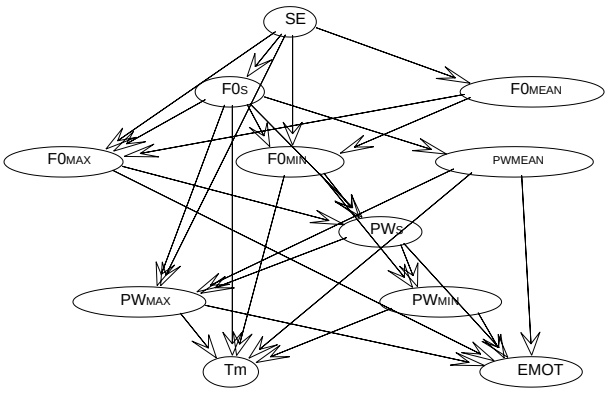


Figure A.15: Second phase BN model of surprise and happiness

Acknowledgement

From 2006 to 2008, at department of computer science and engineering Nagoya Institute of Technology, I was affiliated with Itoh Hidenori laboratory as research student and master student, and I have been affiliated with Kato Shohei laboratory as doctor student since 2009, I have engaged in a research about “Detecting emotion from voice based on a Bayesian approach”. I have been taught and supported by many respecters. I would like to show my acknowledgement to them in this section.

First, I would like to show my acknowledgement to Professor Emeritus Hidenori Itoh. When I came to Japan, I was interested in his research about “Kansei robotics”. He gave me a research chance about “Kansei robotics” in Japan. In my research, he gave me much useful advice and research concept. I remain deeply indebted to him.

And, I would like to express my sincerest gratitude to Associate Professor Shohei Kato. He has advised research concept, techniques, how to write paper. Furthermore, He also has taught manners of Japanese and supported living abroad. Due to his teaching and support, I am able to have a successful research and life in Japan. I wish to record my best thanks to him.

And I wish to record my thanks to Professor Taizo Umezaki and Professor Ichi Takumi. I was able to improve the quality of this dissertation thanks to their benefit comments and advises. I am grateful to them.

I also thank all of members in Kato Shohei laboratory.

Finally, I would like to thank my family and my friends for supporting my research.

Bibliograghy

[1] M. Shigenaga, "Features of emotionally uttered speech revealed by discriminant analysis," *The Transactions of the Institute of Electronics, Information and Communication Engineers*, vol. J83-A, no. 6, pp. 726–735, 2000. (in Japanese).

[2] T. Shirasawa, T. Yamamura, T. Tanaka, and N. Ohnishi, "Discriminating emotions intended in speech," *Technical report of IEICE. HIP*, vol. 96, no. 499, pp. 79–84, 1997. (in Japanese).

[3] Y. Kinjo, Y. Tsuchimoto, and I. Nagayama, "Feeling recognition of spoken words for advanced communication with emotional information processing," *Technical report of IEICE. HIP*, vol. 101, no. 594, pp. 49–54, 2002.

[4] T. Moriyama and S. Ozawa, "A measurement of human vocal emotion using fuzzy control," *The transactions of the Institute of Electronics, Information, and Communication Engineers*, vol. J82-D-11, no. 10, pp. 1710–1720, 1999. (in Japanese).

[5] D. Ververidis, C. Kotropoulos, and I. Pitas, "Automatic emotional speech classification," *IEEE International Conference on Acoustic, Speech, and Signal Processing*, vol. 1, pp. I–593–596, 2004.

[6] C. M. Lee and S. S. Narayanan, "Toward detecting emotions in spoken dialogs," *IEEE Trans. on Speech and Audio Processing*, vol. 13, pp. 293–303, 2005.

[7] W. James, "What is emotion?," *Mind*, vol. 9, pp. 188–205, 1884.

[8] S. Schachter and J. Singer, "Cognitive, social, and physiological determinants of emotional state," *Psychological Review*, vol. 69, pp. 379–399, 1962.

[9] P. Ekman and W. V. Friesen, *Unmasking the Face*. Pretice-Hall, 1975.

[10] B. Mesquita and R.Walker, "Cultural differences in emotions: a context for interpreting emotion experiences," *Behaviour Research and Therapy*, vol. 41, no. 7, pp. 777–793, 2003.

[11] Y. Fujita, "Development of personal robot papero," *Journal of the Society of Instrument and Control Engineers*, vol. 42, no. 6, pp. 521–526, 2003. (in Japanese).

[12] C. Breazeal, "Designing socialable robots," *MIT Press*, 2002.

[13] S. Kanda, Y. Murase, and K. Fujioka, "Internet-based robot: Mobile agent robot of next-generation (maron-1)," *Fujitsu*, vol. 54, pp. 285–292, 2003. (in Japanese).

[14] Business Design Laboratory Co. LTD; The Extremely Expressive Communication Robot, Ifbot, http://www. business-design.co.jp/en/product/001/index.html.

[15] S. Kato, S. Ohsiro, H. Itoh, and K. Kimura, "Development of a communication robot ifbot," *In IEEE International Conference on Robotics and Automation*, pp. 697–702, 2004.

[16] M. Kanoh, S. Kato, and H. Itoh, "Facial expressions using emotional space in sensitivity communication robot ifbot," *In IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2004)*, pp. 1586–1591, 2004.

[17] S. Kato, S. Ohsiro, K. Watabe, H. Itoh, and K. Kimura, "A domestic robot with sensitive communication and its vision system for talker distinction," *Proceedings of the 8th Conference on Intelligent Autonomous Systems*, pp. 1162–1168, 2004.

[18] T. Akiba and H. Tanaka, "A bayesian approach for user modelling in dialog systems," *In 15th Internation Conference of Computational Linguistics*, pp. 1212–1218, 1994.

[19] G. F. Cooper and E. Herskovits, "A bayesian method for the induction of probabilistic networks from data," *Machine Learning*, vol. 9, pp. 309–347, 1992.

[20] K. B. Korb and N. Ann E, *Bayesian Artificial Intelligence*. Chapman & Hall/CRC, 1997.

[21] J. Pearl, *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, 1998.

[22] F. V. Jensen, *Bayesian Networks and Decision Graphs*. Springer-Verlag, 2001.

[23] M. Henrion, "Propagating uncertainty in bayesian networks by logic sampling," *Uncertainty in Artificial Intelligence*, vol. 2, pp. 149–163, 1988.

[24] K. P. Murphy, Y. Weiss, and M. I. Jordan, "Loopy belief propagation for approximate inference, an empirical study," *In Proceedings of Uncertainty in AI*, pp. 467–475, 1999.

[25] G. F. Cooper and E. Herskovits, "A bayesian method for constructing bayesian belief networks from databases," *In Proceedings of the 17th conference on Uncertainty in Artificial intelligence*, pp. 86–94, 1991.

[26] N. Friedman, D. Geiger, and M. Goldszmind, "Bayesian network classifiers," *Machine Learning*, vol. 29, pp. 131–163, 1997.

[27] L. Jiang, H. Zhang, Z. Cai, and J. Su, "Learning tree augmented naive bayes for ranking," *Lecture Notes in Computer Science*, vol. 3453, pp. 688–698, 2005.

[28] J. Cheng and R. Greiner, "Comparing bayesian network classifiers," *In Proceeedings of the 15th Conference on Uncertainty in Artificial Intelligence(UAI'99)*, pp. 101–107, 1999.

[29] G. Endo, J. Nakanishi, J. Morimoto, and G. Cheng, "Experimental studies of a neural oscillator for biped locomotion with qrio," *In IEEE International Conference on Robotics and Automation*, pp. 598–604, 2005.

[30] Y. Murase, Y. Yasukawa, K. Sakai, and M. Ueki, "Design of a compact humanoid robot as a platform," *In Proc. of the 19th conference of Robotics Society of Japan*, pp. 789–790, 2001. (in Japanese).

[31] S. Takeuchi, A. Sakai, S. Kato, and H. Itoh, "An emotion generation model based on the dialogist likability for sensitivity communication robot," *Journal of the Robotics Society of Japan*, no. 7, pp. 1125–1133, 2007. (in Japanese).

[32] M. Fujita, "Development of an autonomous quadruped robot for robot entertainment," *Autonomous Robots*, vol. 5, pp. 7–18, 1998.

[33] S. Kato, Y. Sugino, and H. Itoh, "A bayesian approach to emotion detection in dialogist's voice for human robot interaction," *Lecture Notes in Computer Science*, vol. 4252, pp. 961–968, 2006.

[34] H. Shibata, M. Kanoh, S. Kato, and H. Itoh, "A system for converting robot 'emotion' into facial expressions," *IEEE International Conference on Robotics and Automation*, pp. 3660–3665, 2006.

[35] C. Itoh, S. Kato, and H. Itoh, "Mood-transition-based emotion generation model for the robot's personality," *IEEE International Conference on Systems, Man, and Cybernetics*, pp. 2957–2962, 2009.

[36] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, and et al., "Emotion recognition in human-computer interaction," *IEEE Trans. on IEEE Signal Processing Magazine*, vol. 18, pp. 32–80, 2001.

[37] K. R. Scherer, T. Johnstone, and G. Klasmeyar, *Vocal expression of emotion.* Handbook of the Affective Science, Oxford University Press, 2003.

[38] G. Schwarz, "Estimating the dimension of a model," *Annals of Statistics*, vol. 6, no. 2, pp. 461–464, 1978.

[39] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royals Statistical Society*, vol. 39, pp. 1–38, 1977.

[40] K. Sjölander The Snack Sound Toolkit, http://www.speech.kth.se/snack.

[41] K. P. Murphy Bayes Net Toolbox, http://www.cs.ubc.ca/˜murphyk/Soft ware/BNT/ bnt.html.

[42] J. Cho, S. Kato, and H. Itoh, "Bayesian-based inference of dialogist's emotion for sensitivity robots," *IEEE International Conference on Robot & Human Interactive Communication*, pp. 792–797, 2007.

[43] J. Cho, S. Kato, M. Kanoh, and H. Itoh, "Bayesian method for detecting emotion from voice for kansei robots," *JSKE Journal of Kansei Engineering International*, vol. 8, no. 1, pp. 15–22, 2009.

[44] J. Cho, S. Kato, M. Kanoh, and H. Itoh, "A method of inferring dialogist's emotion for sensitivity robots using bayesian network," *Information Technology Letters*, vol. 6, pp. 327–330, 2007. (in Japanese).

[45] J. Cho, S. Kato, and H. Itoh, "A biphase-bayesian-based method of emotion detection from talking voice," *Lecture Notes in Artificial Intelligence*, vol. 5179, pp. 50–57–22, 2009.

[46] S. H. Pack and J. Fürnkranz, "Efficient pairwise classification," *Machine Learning:ECML 2007*, vol. 4701, pp. 658–665, 2007.

[47] E. Hüllermeier and S. Vanderlooy, "Combining predictions in pairwise classification: An optimal adaptive voting strategy and its relation to weighted voting," *Pattern Recognition*, vol. 43, pp. 128–142, 2010.

[48] D. Price, S. Knerr, L. Personnaz, and G. Dreyfus, "Pairwise neural network classifiers with probabilistic outputs," *in Advances in Neural Information Processing Systems*, vol. 7, pp. 1109–1116, 1995.

[49] J. Cho and S. Kato, "Detecting emotion from voice using selective bayesian pairwise classification," *IEEE Symposium on Computer & Informatics*, pp. 90–95, 2011.

[50] J. Cho and S. Kato, "Emotion detection from voice using bayesian pairwise classification," *The 28th Annual Conference of the Robotics Society of Japan*, pp. 2I2–2 (4 pages), 2011. (in Japanese).

[51] D. Ververidis and C. Kotropoulos, "Emotion speech recognition:resources, features, and methods," *Speech Communication*, vol. 48, pp. 1162–1181, 2006.

[52] H. Liu and H. Motoda, *Feature Extraction Construction and Selection: A Data Mining Perspective*. Kluwer Academic Publishers, 1998.

[53] R. Kohavi and G. H. Jone, "Wrappers for feature subset selection," *Artificial Intelligence*, vol. 97, pp. 273–324, 1997.

[54] C. L. Blake and C. J. Merz, "Uci repository of machine learning databases," Http://www.ics.uci.edu/ mlearn.MLRepository.html.

[55] J. Cho and S. Kato, "Probabilistic pairwise classification using selective tree augmented naive bayes," *12th International Symposium on Advanced Intelligent Systems*, pp. 291–294, 2011.

[56] Y. Kitahara and Y. Tohkura, "Prosodic control to express emotions for man-machine speech interaction," *IEICE Trans. Fundamentals*, vol. E75-A, no. 2, pp. 151–163, 1992.

[57] T. L. Nwe, S. W. Foo, and L. C. D. Silva, "Speech emotion recognition using hidden markov models," *Speech Communication*, vol. 41, no. 4, pp. 603–623, 2003.

[58] J. Cho, S. Kato, and H. Itoh, "Comparison of sensibilities of japanese and korean in recognizing emotions from speech by using bayesian networks," *IEEE International Conference on Systems, Man, and Cybernetics*, pp. 2866–2871, 2009.

[59] J. Cho, S. Kato, and H. Itoh, "The comparison of kansei on the emotion recognition from talking voice between japan and korea using bayesian approach,"

Journal of Japan Society of Kansei Engineering, vol. 8, no. 3, pp. 913–919, 2009. (in Japanese).

List of Publications

Journal

1. Jangsik Cho, Shohei Kato, Masayoshi Kanoh and Hidenori Itoh, A Method of Inferfing Dialogist's Emotion for Sensitivity Robots using Bayesian Network, Information Technology Letters, Vol.6, pp.327-330, 2007 (short paper) (in Japanese).

2. Jangsik Cho, Shohei Kato, Masayoshi Kanoh and Hidenori Itoh, Bayesian Method for Detection Emotion from Voice for Kansei Robots, JSKE journal of Kansei Engineering International, Vol.8, No.1 pp.15-22, 2009.

3. Jangsik Cho, Shohei Kato, and Hidenori Itoh, The Comparison of Kansei on the Emotion Recognition from Talking Voice between Japan and Korea using Bayesian Approach, Journal of Japan Society of Kansei Engineering, Vol.8, No.3, pp.913-919, 2009 (in Japanese).

International Conference

1. Jangsik Cho, Shohei Kato, and Hidenori Itoh, Bayesian-Based Inference of Dialogist's Emotion for Sensitivity Robots, 16th IEEE International Symposium on Robot & Human Interactive Communication, pp.792-797,2007.

2. Jangsik Cho, Shohei Kato, and Hidenori Itoh, A Biphase-Bayesian-Based Method of Emotion Detection from Talking Voice, Lecture notes in Artificial Intelligence (12th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems), Vol.5179, pp.50-57, 2008.

3. Jangsik Cho, Shohei Kato, and Hidenori Itoh, Comparison of Sensibilities of Japanese and Koreans in Recognizing Emotions from Speech by using Bayesian Networks, 2009 IEEE International Conference on Systems, Man, and Sybernetics, pp.2866-2871, 2009.

4. Jangsik Cho and Shohei Kato, Detecting Emotion from Voice using Selective Bayesian Pairwise Classifiers, IEEE Symposium on computer & Informatics, pp.90-95, 2011.

5. Jangsik Cho and Shohei Kato, Probabilistic Pairwise Classification using Selective Tree Augmented Naive Bayes, 12th International Symposium on Advanced Intelligent Systems, pp.291-294, 2011.

The Others

1. Jangsik Cho, Shohei Kato, and Hidenori Itoh, Bayesian-Based Inference of Dialogist's Emotion for Sensitivity Robots, The 74th National Convention of IPSJ, Vol.4 pp. 235-236, 2007 (in Japanese).

2. Jangsik Cho and Shohei Kato, Emotion Detection from Voice using Bayesian Pairwise Classification, The 28th Annual Conference of the Robotics Society of Japan, 2I2-2 (4 pages), 2010 (in Japanese).