國 立 交 通 大 學

電信工程研究所

碩 士 論 文

使用時頻變化調變於
強健語音情緒辨識

Spectro-Temporal Modulations for Robust Speech Emotion Recognition

研究生：葉藍霙

指導教授：冀泰石 博士

中 華 民 國　　99 年 7 月 29 日

使用時頻變化調變於

強健語音情緒辨識

Spectro-Temporal Modulations for Robust Speech Emotion Recognition

研 究 生：葉藍霙　　Student: Lan-Ying Yeh

指導教授：冀泰石 博士　　Advisor: Dr. Tai-Shih Chi

國立交通大學

電信工程研究所

碩士論文

A Thesis
Submitted to Institute of Communication Engineering
College of Electrical and Computer Engineering
National Chiao-Tung University
In Partial Fulfillment of the Requirements
for the Degree of
Master of Science in
Communication Engineering
June 2010
Hsin-Chu, Taiwan, Republic of China

中華民國九十九年七月

使用時頻變化調變於
強健語音情緒辨識

學生：葉藍霖　　指導教授：冀泰石 博士

國立交通大學電信工程研究所

中文摘要

語音情緒的分類是近年來新興的研究題目，目前大多數的研究都著重在乾淨語音中進行分類。在本論文中，我們利用聽覺感知模型提出一種新的時頻變化參數（joint Rate-Scale features, RS features），藉由此參數來處理有雜訊情況下的語音情緒辨識的問題。我們將柏林情緒語料庫(Berlin Emotional Database)以及愛寶情緒語料庫(FAU AIBO Database)加入不同訊雜比的白雜訊(white noise)及人聲雜訊(babble noise)，並且以乾淨語料訓練、有雜訊語料測試的方式評估效能，以模擬真實應用中未能事先預知雜訊程度的狀況。我們也進一步使用循序前進浮動搜尋(Sequential Forward Floating Selection, SFFS)來探討所提出特徵參數的冗餘性，以進一步降低所需參數的維度。實驗於柏林情緒語料庫結果顯示，與傳統音韻參數結合梅爾倒頻率係數參數相比，尤其在低訊雜比的情況下，使用時頻變化參數將有更高的辨識率。實驗結果顯示對於愛寶情緒語料庫，在訊雜比很高的情況下，傳統參數和時頻變化參數皆有過度訓練的情況，需要進一步降低維度及改進參數。

Spectro-Temporal Modulations for Robust Speech Emotion Recognition

Student: Lan-Ying Yeh Advisor: Dr. Tai-Shih Chi

Institute of Communication Engineering

National Chiao-Tung University

English Abstract

Speech emotion recognition is mostly considered in clean speech. In this thesis, joint Rate-Scale features (RS features) are extracted from an auditory model and are applied to detect the emotion status of noisy speech. The noisy speech is derived from the Berlin Emotional Speech database and the FAU AIBO database with added white and babble noises under various SNR levels. The clean train/noisy test scenario is investigated to simulate conditions with unknown noisy sources. The sequential forward floating selection (SFFS) method is adopted to demonstrate the redundancy of RS features and further dimensionality reduction is conducted. Compared with conventional MFCCs plus prosodic features, RS features show higher recognition rates especially in low SNR conditions on Berlin database. However, both conventional and RS features are over-trained in low SNR conditions on AIBO database. Feature selection or reduction techniques are further required.

致謝

這兩年來，最需要感謝的人，就是我的指導教授冀泰石博士。教授親切沒有距離，亦師亦友，並且用心指導每一個學生，一再幫忙修改我們的論文，真的辛苦您了。教授也不斷提醒我們做事的態度和自我管理的重要，對我們未來人生都很有啟發。

感謝實驗室的夥伴，陪我走過從實驗室到餐廳的路大概五、六百次了吧！博班學長大師、阿郎、畢業的 Nick，在我蒙懂無知時給予好多幫助。同屆的大樹、勝哥、叮咚，更是一起從碩一修課、討論作業、研究問題建立的革命情感。在研究卡住時，與實驗室夥伴的討論往往能有很多收穫。也謝謝實驗室的學弟妹，總是帶給我們很多歡樂。

謝謝我的朋友、室友，雖然有些好朋友、好姊妹遠在別的城市、別的國家，還是時時刻刻關心我，陪我聊天打屁分享八卦、吃飯逛街紓解壓力。謝謝親愛的胖達，容忍我的焦躁不安，帶我遊山玩水，把我養的胖胖的。

最後，當然要感謝我的父母、家人，一直相信我能夠做到，給予我支持。與你們分享我努力的成果。

藍龔

2010 年 6 月

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1. Motivation

Speech emotion recognition has been a popular research topic over the last decade. Knowing the emotion status of the speaker is important for human-machine interfaces with better interaction experiences. Many modern applications, such as interactive robots, infant or elder caring systems and speech-recognition based customer service lines, can use such information. However, early studies are often launched on “perfect” conditions, i.e., clean speech with acting emotions, which is far from real-world applications. In this work, we intend to find a robust feature set from a spectro-temporal auditory perceptual model [1, 2] for speech emotion recognition. It is our belief that the new auditory features could be more robust in noisy situations.

1.2. Related Works

Researchers have been devoted to searching novel features and designing powerful classifier to improve the recognition rate. For the emotion recognition, it has been shown that statistic features characterized by GMM models outperform instantaneous features used

in a HMM based recognizer [3]. Moreover, other works intended to combine temporal features with statistic features, such as the one shown in [4]. The best feature sets have been discussed over years, and it is well acknowledged that pitch, energy, and duration contribute the most to emotion recognition [5, 6]. Spectral information or formants are also discussed frequently. Some works, such as in [7], [8], focus on finding the best units of speech segmentation to boost emotion recognition rate. Linguistic information combined with acoustic and prosodic features is also discussed in [9]. As for the recognizers, studies showed the support vector machine (SVM) outperforms and is more robust than K nearest neighbor (K-NN) and neutral network (NN) provided statistic features used [10, 11].

Recently, more and more studies pay attention to natural emotion or real environments with noises [12, 13]. However, these studies often went forward to find thousands of features in order to obtain an optimal set of features with the highest recognition rate for any particular testing environment. These brute-force methods seem working, but these studies only evaluate their performance under the matched condition, where the testing data is under the same noise level as the training data. Undoubtedly, degraded performance is expected with changes of testing environments.

During early days, emotion databases are usually built by either collecting acted speech from television dramas or recording speech by actors performing certain emotions. One problem is that these databases are often built by different research groups; hence, lacking a common ground for fair performance comparisons. Another problem is that these databases only contain “acted emotions,” which may be different to the human emotions expressed in daily lives. To address the first problem, the Berlin emotion speech corpus [14] is used in this thesis since it is widely used by other researchers in recent years. As for the second problem, we use the FAU AIBO database [15], which contains speech with natural emotions and is adopted as the test database in the INTERSPEECH 2009 emotion challenge.

1.3. Experimental Framework

In this thesis, the Berlin emotional speech database and the FAU AIBO database with additive noises is utilized to test the robustness of proposed spectro-temporal auditory features. A linear-kernel SVM [16] is used as the emotion classifier. Recognition rates of our spectro-temporal auditory RS features are evaluated and compared with conventional spectral features (MFCCs) plus additional prosodic features under additive white and babble noises. Furthermore, the dimensionality reduction of our RS features is conducted and corresponding performance is investigated. The flowchart of our proposed method is shown in Figure 1-1.

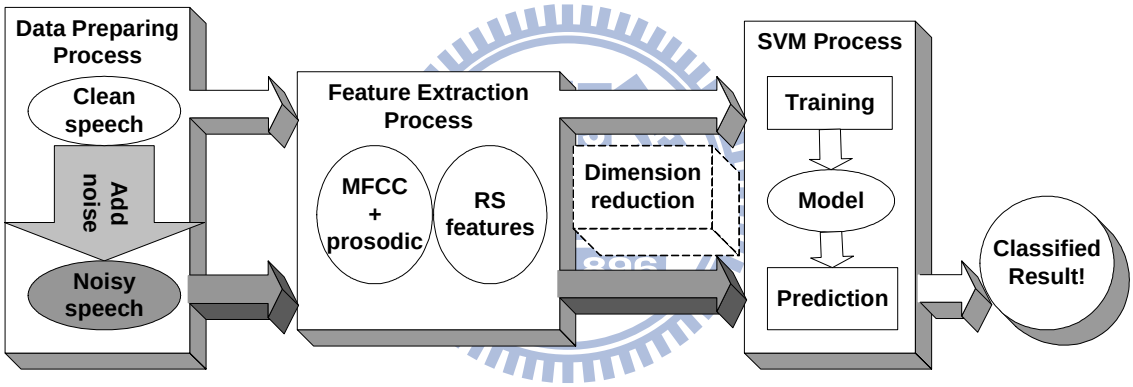


Figure 1- 1 Overall flowchart of proposed method.

1.4. Thesis Organization

This thesis is organized as follows. In section 2, a brief literature review of the two-module spectro-temporal auditory model and support vector machine are given. Two emotion speech databases and four sets of features used in this study are then introduced in section 3. In section 4, experimental setup and recognition results on two databases are demonstrated. We end in section 5 with conclusions and discussions.

Chapter 2

Literature Review

2.1. Auditory Model

The auditory features adopted in this study are extracted from stages of a physiological based auditory model. For better understanding the ideas and reasons of auditory model, some hearing physiology of human perception will be briefly introduced at first. Then, the auditory model which consists of an early cochlear (ear) and a central cortical (A1) module will be discussed in section 2.1.2 and 2.1.3.

2.1.1. Hearing Physiology

The cross-sectional view of the human ear is shown in Figure 2-1. It can be divided into three parts: the outer ear, the middle ear, and the inner ear. Sound waves enter the outer ear and travel through the ear canal to the tympanic membrane (ear drum). The vibrations of the ear drum are transmitted into the inner ear through three ossicles (the malleus, incus and stapes) in the middle ear. The stirrup touches a liquid filled sack and the vibrations travel into the cochlea, which is shaped like a shell. The cochlea attaches to hundreds of nerve fibers, which transmit information along the auditory pathway to the brain. Finally, the brain

processes the information from the ear for various tasks.

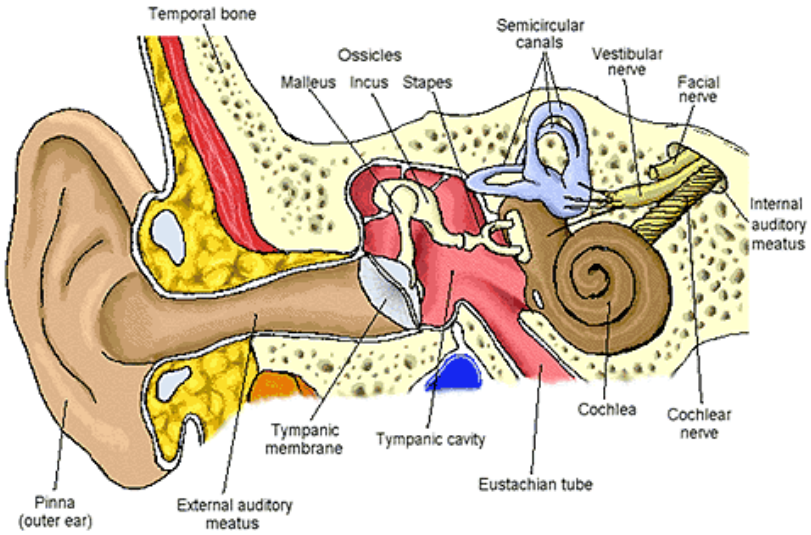


Figure 2- 1 Cross-sectional view of the human ear (http://mail.pittsfield.net/teachersites/Whelihan_Kathleen/).

The major functions of the out ear are localization, amplification and protection. The shape of the outer ear enables people to collect sound waves and judge the direction of sound source easily. The three ossicles transduct the acoustical vibrations into mechanical vibrations and compensate part of the loss of energy due to entering the liquid from the air. The cochlea in the inner ear plays a significant role in the auditory system. The structure of the cochlea is shown in Figure 2-2. The left panel shows the stretched cochlea with the basilar membrane (BM), which is about 35 mm in length with its width increasing and stiffness decreasing both non-uniformly from base to apex. When a mechanical vibration reaches the oval window, a traveling wave is generated and propagates along the basilar membrane of the cochlea. Because of the different stiffness along the BM, the traveling waves caused by different frequencies will reach maximum response and stop at different locations of the BM. The left panel of Figure 2-2 shows the side view and top view of cochlea and the right panel shows a schematic plot about maximum responsive frequencies along the basilar membrane. The lower the frequency is, the further the traveling wave

reaches. A linear relationship was observed between the traveling distance from the cochlear base and the log-frequency of input sounds. The range of resonance frequencies is about 20-20,000 Hz, which is the audible frequency range of human beings. Due to the mechanical property of the traveling wave, the maximum response on a specific frequency would inhibit its neighboring frequencies on the BM. This might explain the well-known "frequency masking" phenomenon of human audition.

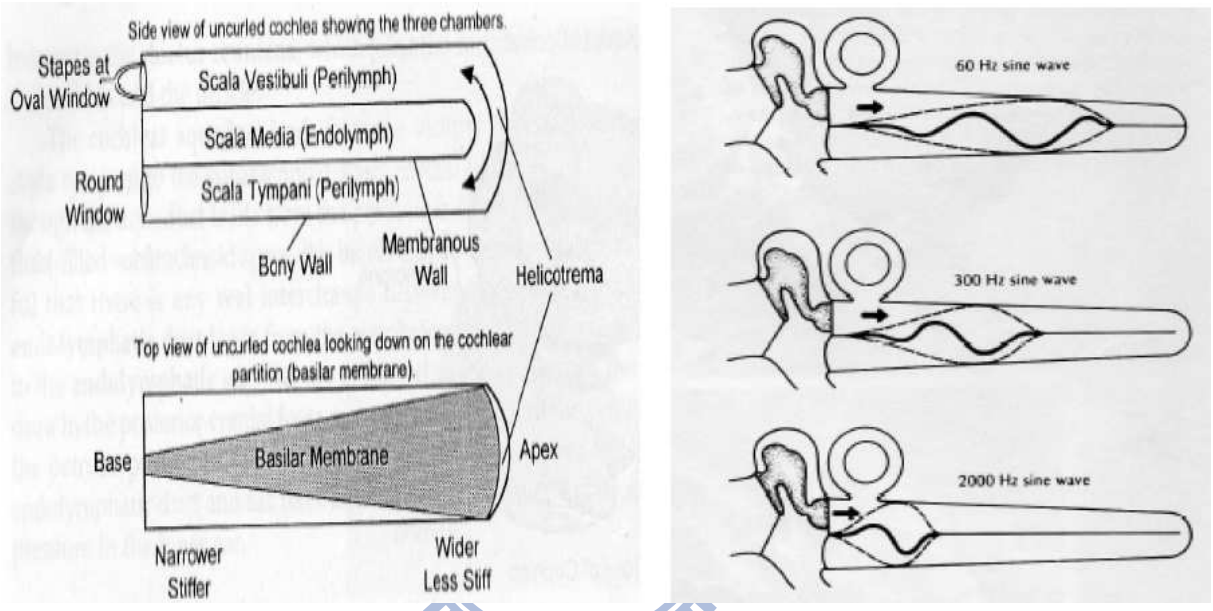


Figure 2- 2 Structure of cochlea (left) and responses for different frequencies (right) (Hearing Physiology Handout, AAIP).

There are about 3000 inner hair cells distributed along the basilar membrane. When a traveling wave generates displacement on the BM, the hair cells will be stimulated and remit electrical signals via auditory nerves to the midbrain. There are two different hair cells: inner hair cells and outer hair cells. Most of this mechanism of transforming mechanical vibrations into electrical signals is done by inner hair cells. Outer hair cells, on the other hand, are often active in further amplification or reduction in pertaining to extreme sounds. Due to the fact that a relaxation time is needed between consecutive fires of auditory neurons, firing rates can not keep up with high frequency vibrations, as demonstrated in Figure 2-3. Firing rates of inner hair cells are bounded by 4-5k Hz, while the rates of the

midbrain are bounded by about 1k Hz.

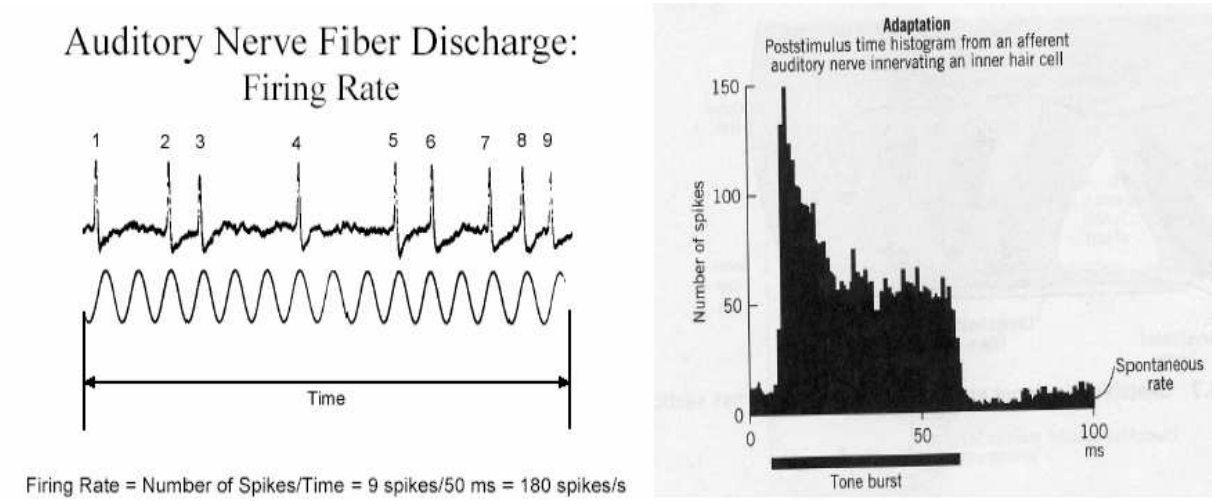


Figure 2- 3 The firing rate of auditory nerve correspond to the single tone input (left) and the adaptation mechanism of auditory nerve (Hearing Physiology Handout, AAIP).

2.1.2. Cochlear Module

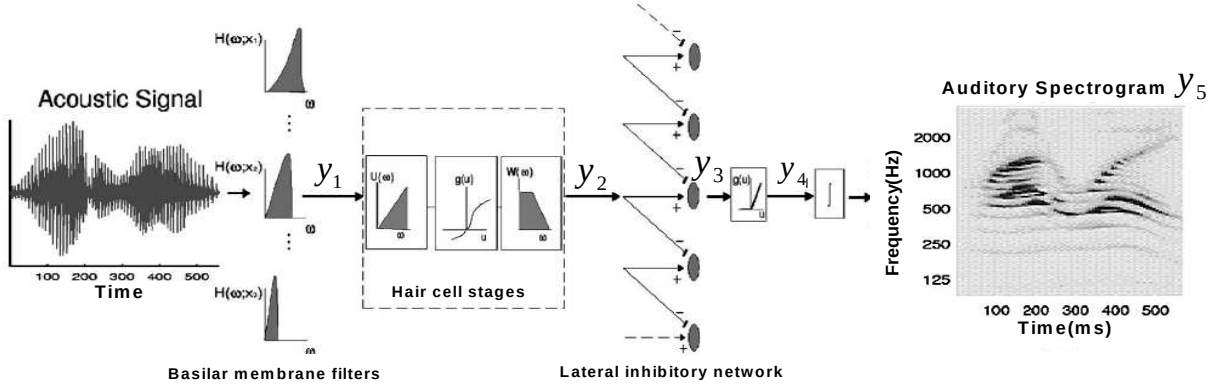


Figure 2- 4 Stages of the early cochlear module (adopted from [2])

The cochlear module models functions of the peripheral auditory system. As shown in Figure 2-4, it first consists of a bank of 128 overlapping asymmetric constant-Q bandpass filters ($Q_{3dB} \approx 4$) which mimic the frequency selectivity of the cochlea. These filters distribute evenly over 5.3 octaves with 24 filters/octave frequency resolution. The output of each filter is fed into a non-linear compression stage and a lateral inhibitory network (LIN), and then processed by an envelope extractor. The non-linear compression is to model the

saturation of the inner hair cells, and the LIN is to model the frequency masking effect. In this study, a simplified linear version of this module without the hair cell stage is used. All tested speech signals are normalized in advance to avoid the high-gain compression done by hair cells. Outputs of different stages of this module can be written as:

$$y_1(t,\omega) = s(t) *_t h(t;\omega) \tag{2-1}$$

$$y_3(t,\omega) = \partial_\omega y_1(t,\omega) \tag{2-2}$$

$$y_4(t,\omega) = \max(y_3(t,\omega), 0) \tag{2-3}$$

$$y_5(t,\omega) = y_4(t,\omega) *_t \mu(t;\tau) \tag{2-4}$$

where $h(t;\omega)$ is the impulse response of the constant-Q cochlear filter with center frequency ω ; $*_t$ depicts the convolution in time; the integration window $\mu(t;\tau) = e^{-t/\tau} \cdot u(t)$ with the time constant τ models the current leakage along the neural pathway to the midbrain; and $u(t)$ is the unit step function.

The output $y_5(t,\omega)$ is referred to as an auditory spectrogram, which represents neuron activities along the time and log-frequency axis. Intuitively, it is similar to the magnitude response of a mel-scaled FFT based spectrogram, where our constant-Q criterion approximates the mel-scale and our local envelope approximates the magnitude of a FFT based spectrogram.

2.1.3. Cortical Module and Rate-Scale Representation

The second module models the spectro-temporal selectivity of neurons of the auditory cortex (A1). Briefly speaking, the auditory spectrogram $y_5(t,\omega)$ is further analyzed by A1's neurons which are modeled by two-dimensional filters tuned to different spectro-temporal

modulation parameters [2]. The rate (or velocity) parameter in Hz reflects how fast the local spectro-temporal envelope varies along the temporal axis. The scale (or density) parameter in cycle/octave characterizes how broad the signal's local spectro-temporal envelope distributed along the log-frequency axis.

In addition to the rate and scale, cortical neurons are also found to be sensitive to the direction of the FM sweep. This directionality is characterized in this module by the sign of the rate (negative for upward sweeping; positive for downward sweeping). From functional point of view, this module models cortical neurons as performing a joint spectro-temporal multi-resolution analysis (due to various rate-scale combinations) on the input auditory spectrogram. The excitation pattern of cortical neurons to a single t-f point in the spectrogram is referred to as the rate-scale representation of that particular t-f point. Each rate-scale representation is labeled by neurons' tuning characteristic of rate, scale, and directionality.

Two averaged rate-scale plots over the frequency axis around 200 and 550 ms are given in Figure 2-5. Two aspects are clearly shown in each rate-scale plot: (1) spectro-temporal modulations of envelopes and (2) resolved pitch below 512 Hz. Take the 550 ms frame as an example. The resolved pitch around 230 Hz excites {high rate, fine scale} neurons, thus produces the corresponding rate-scale representation. On the other hand, envelopes of the almost flat harmonic structure shown at 230, 460 and 1150 Hz excite neurons tuned to {low rate (due to the flatness), low scale (2 cycles within 2.32 octave)} and produce strong rate-scale responses at regions less than 8 Hz and less than 1 cycle/octave. Since flat envelopes do not favor any sweeping directions, symmetric responses to rate are clearly shown in the {low rate, low scale} region. More detailed description and mathematic formulation of this cortical module can be found in [2].

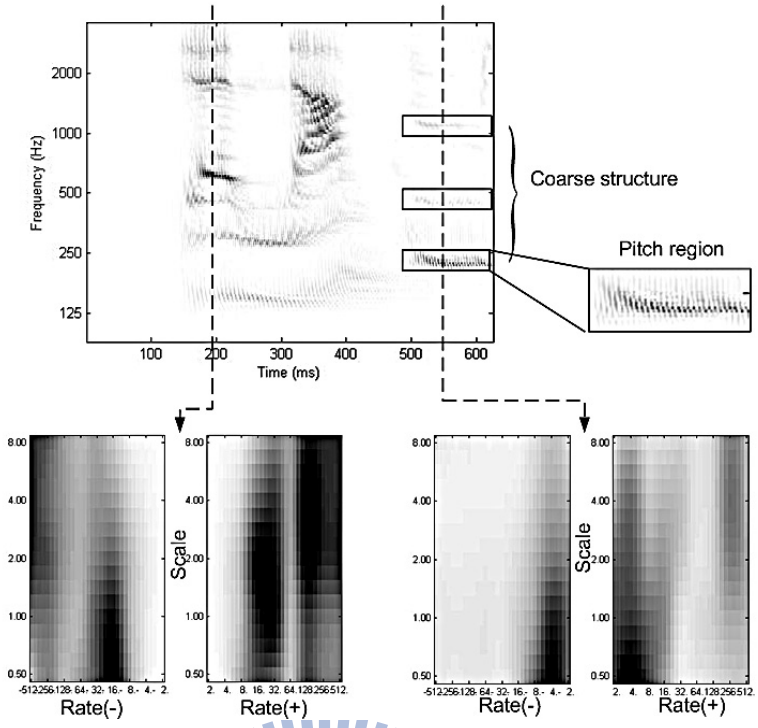


Figure 2- 5 Rate-scale representation produced by the cortical module.

2.2. Support Vector Machine (SVM)

The SVM, a supervised learning algorithm, is usually used for classification and regression. It is very popular in recent years due to its remarkable performance. In this thesis, we adopt the support vector machine as our emotion classifier. In this section, we will give a brief introduction to SVM. Detailed setups for our experiments will then be given in section 4.1.

2.2.1. Separable problem

For a supervised learning algorithm, we consider a set of training samples $\left(\mathbf{x}^{(i)}, y^{(i)}\right)$

where $\mathbf{x^{(i)}} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{bmatrix}$ is the m-dimensional feature vector of the i-th training data, i =1, 2, …,

n and $y^{(i)} \in \{1,-1\}$ represents the class label of the i-th training data for a basic two-class classification problem. As Figure 2-6 indicates, we want to find a hyperplane that can perfectly separate these two classes. The hyperplane can be represented as:

$$g(\mathbf{x}) = \mathbf{w}^{\mathbf{T}}\mathbf{x} + b \tag{2-5}$$

then, the data of the two classes satisfies:

$$\Rightarrow \begin{cases} \mathbf{w}^{\mathbf{T}}\mathbf{x}^{(\mathbf{i})} + b \geq 1, \text{ if } y^{(i)} = 1 \\ \mathbf{w}^{\mathbf{T}}\mathbf{x}^{(\mathbf{i})} + b \leq 1, \text{ if } y^{(i)} = -1 \end{cases} \tag{2-6}$$

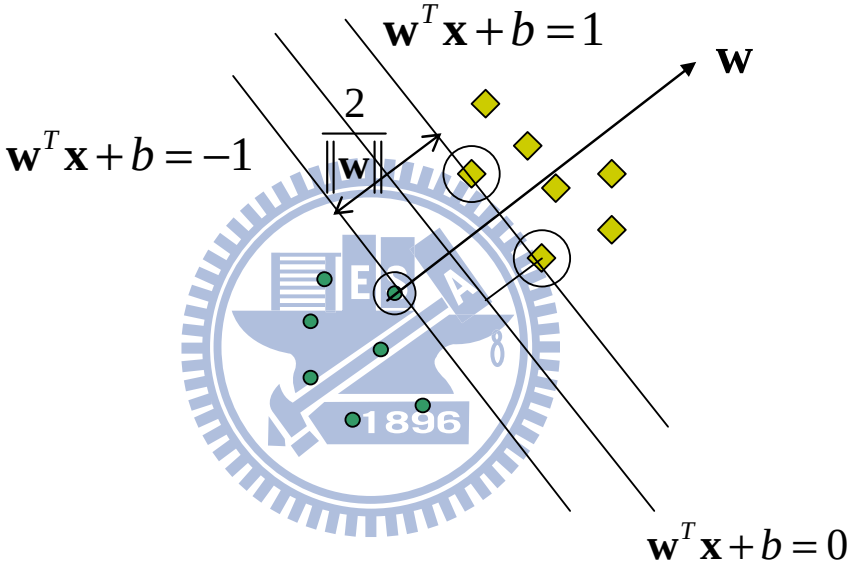


Figure 2- 6 The optimal hyperplane for a separable problem using SVM.

There may be a lot of choices for $(\mathbf{w},b)$ that are separable; however, the goal of the SVM is to find a hyperplane which possesses the largest separation, or margin, between the two classes. That is, we want to choose a hyperplane so that the distance from it to the nearest data point on each side is maximized. From Figure 2-6, the margin can be represented as $\frac{2}{\|\mathbf{w}\|}$. The optimal separating hyperplane can be found by solving the following problem:

$$\min \frac{1}{2}\|\mathbf{w}\|^2$$

$$\text{s.t. } y^{(i)}(\mathbf{w}^{\mathbf{T}}\mathbf{x}^{(\mathbf{i})} + b) \geq 1, \text{ for all } i = 1...n$$

(2-7)

Equation (2-7) can be solved by constructing Lagrange multipliers $\alpha_i \geq 0$ in the following primal form:

$$\max_{\alpha} \min_{\mathbf{w},b} L_p = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{n} \alpha_i \left[y^{(i)} \left(\mathbf{w}^T \mathbf{x}^{(i)} + b \right) - 1 \right] \tag{2-8}$$

To find the saddle point, one has to minimize function (2-8) over **w**, *b* and to maximize it over the nonnegative Lagrange multipliers $\alpha_i \geq 0$. At the saddle point, one obtains:

$$\nabla_{\mathbf{w}} L_p = 0 \Rightarrow \mathbf{w} = \sum_{i=1}^{n} \alpha_i y^{(i)} \mathbf{x}^{(i)} \tag{2-9}$$

$$\frac{\partial}{\partial b} L_p = \sum_{i=0}^{n} \alpha_i y^{(i)} = 0 \tag{2-10}$$

Substituting (2.9) and (2.10) into (2.8), it can be further modified into the following dual form:

$$\begin{aligned} &\max_{\alpha} \; L_D = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} y^{(i)} y^{(j)} \alpha_i \alpha_j \mathbf{x}^{(\mathbf{i})} \cdot \mathbf{x}^{(\mathbf{j})} \\ &s.t. \; \alpha_i \geq 0, \; i = 1, \ldots, n \\ &\sum_{i=1}^{m} \alpha_i y^{(i)} = 0 \end{aligned} \tag{2-11}$$

Invoking the Karush-Kuhn-Tucker dual complementary conditions, problems in (2-11) form can further be derived into the following form:

$$\alpha_i \left(y^{(i)} \left(\mathbf{w}^T \mathbf{x}^{(i)} + b \right) - 1 \right) = 0 \tag{2-12}$$

Those points satisfy $y^{(i)} \left(\mathbf{w}^T \mathbf{x}^{(i)} + b \right) - 1 > 0$ (i.e., $\alpha_i = 0$) are called non-support vectors. On the contrary, those points reside on the hyperplane $y^{(i)} \left(\mathbf{w}^T \mathbf{x}^{(i)} + b \right) - 1 = 0$ (i.e., $\alpha_i \geq 0$) are called support vectors. Thus, the computations in equation (2-9) can be further reduced. Finally, α_i are solved by the quadratic programming and parameters **w**, b are then obtained by equation (2-9) and (2-12). Hence, the final classifier $g(\cdot)$ is derived and used to predict a new test point **x** by:

$$g(\mathbf{x}) = sign\left(\mathbf{w}^{\mathbf{T}} \mathbf{x} + b \right) \tag{2-13}$$

2.2.2. Binary non-separable problem

The simplest problem discussed in section 2.2.1 can be extended to a non-separable problem (Figure 2-7) by introducing additional slack variables ξ_i and cost parameter C. We can relax the equation (2-7) into $y^{(i)}(\mathbf{w}^{\mathbf{T}}\mathbf{x}^{(\mathbf{i})}+b)\geq 1-\xi_i$ to tolerate some outliers. The parameter C controls the total relaxation value $\sum_k \xi_k$ in a reasonable small range. The lower the value of C is, the smaller the penalty for outliers is and a softer margin exists.

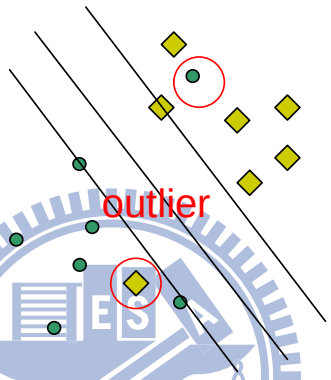


Figure 2- 7 Non-separable problem.

The primal form of equation (2-7) can be reformulated as follows:

$$\begin{aligned} &\min_{w,b} L_P = \frac{1}{2}\|\vec{w}\|^2 + C\sum_{i=1}^{m}\xi_i \\ &\text{s.t. } y^{(i)}\left(w^T x^{(i)}+b\right)\geq 1-\xi_i \\ &\quad \xi_i \geq 0, \text{ for all i=1}\cdots\text{n} \end{aligned} \tag{2-14}$$

and the dual form of equation (2-11) can be modified into:

$$\begin{aligned} &\max_{\alpha}\ L_D = \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{n} y^{(i)}y^{(j)}\alpha_i\alpha_j\mathbf{x}^{(i)}\cdot\mathbf{x}^{(j)} \\ &s.t.\ 0\leq\alpha_i\leq C,\ i=1,...,n \\ &\quad \sum_{i=1}^{n}\alpha_i y^{(i)} = 0 \end{aligned} \tag{2-15}$$

Herein, α_i, $\mathbf{w}$ and b can be solved in a similar way as stated in section 2.2.1.

2.2.3. Nonlinear problem

The formulations of the SVM can also be extended to tackle nonlinear problems. The SVM adopts a way to map the original features into a higher dimensional space, and solves the problem linearly in the new space (see Figure 2-8). If ϕ is our mapping function, the new feature vector can be shown as $\mathbf{x}' = \phi(\mathbf{x})$. Then, the dual form of equation (2-11) can be reformulated as:

$$\begin{aligned}
\max_{\alpha} \; L_D &= \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} y^{(i)} y^{(j)} \alpha_i \alpha_j \mathbf{x}^{(i)} \cdot \mathbf{x}^{(j)} \\
&= \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} y^{(i)} y^{(j)} \alpha_i \alpha_j \phi\left(\mathbf{x}^{(i)}\right) \cdot \phi\left(\mathbf{x}^{(j)}\right) \\
&= \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} y^{(i)} y^{(j)} \alpha_i \alpha_j K\left(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}\right) \qquad \text{(2-11)} \\
s.t. \; \alpha_i &\geq 0, \; i = 1, \ldots, n \\
\sum_{i=1}^{n} \alpha_i y^{(i)} &= 0
\end{aligned}$$

However, the inner product in equation (2-11) would increase the computational load. A key property of SVM recognizers is to use the so-called *kernel function* $K\left(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}\right) = \phi(\mathbf{x}^{(i)})^T \cdot \phi(\mathbf{x}^{(j)})$ to replace the inner product. There is no need to find the mapping function ϕ explicitly, while any function that satisfy Mercer's theorem can be used as kernel functions here. Table 2-1 lists four basic kernel functions used frequently.

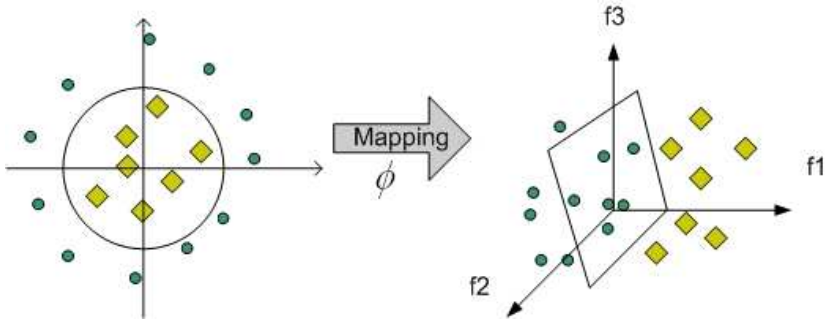


Figure 2- 8 Map nonlinear problem to higher dimensional space.

Table 2- 1 basic kernel functions

linear	$K(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^T \cdot \mathbf{x}_j$
polynomial	$K(\mathbf{x}_i, \mathbf{x}_j) = (\gamma \mathbf{x}_i^T \cdot \mathbf{x}_j + r)$, γ>0
Gaussian (RBF)	$K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma \lVert \mathbf{x}_i - \mathbf{x}_j \rVert^2)$, γ>0
sigmoid	$K(\mathbf{x}_i, \mathbf{x}_j) = \tanh(\gamma \mathbf{x}_i^T \cdot \mathbf{x}_j + r)$

Chapter 3

Database and Feature Extraction

3.1. Berlin Emotional Speech Database (EMO-DB)

The popular Berlin Emotional Speech Database [14] is tested in pilot simulations in this study. Clean speech samples are uttered by five female and five male actors. Each actor speaks ten sentences in German. Each sentence has duration of 2 to 5 seconds. Detail contents are listed in Table 3-1. The database contains emotions of *anger (126)*, *happiness (70)*, *sadness (62)*, *fear (66)*, *disgust (44)*, *boredom (80)*, and *neutral (78)*. Only those utterances scoring higher than 80 emotion recognition rate in a subjective listening test are included in the database. Hence, there are 526 sentences in total with seven classes of emotions. Original speech samples are recorded with 16 kHz sampling frequency under studio condition, and are downsampled to 8 kHz to cover the fundamental frequencies of male speakers when analyzed by our 5.3-octave frequency coverage cochlear filterbank in our auditory model (see section 2.1).

White noise and babble noise are obtained from the NOISEX-92 database [17] and added to clean speech to simulate various SNR conditions. A simple energy-based VAD is first applied to each clean utterance to determine its active regions. Only durations of active regions are considered in calculating SNR.

Table 3- 1 The German content of EMO-DB and its English translation

code	German text	English translation
a01	Der Lappen liegt auf dem Eisschrank.	The tablecloth is lying on the fridge.
a02	Das will sie am Mittwoch abgeben.	She will hand it in on Wednesday.
a04	Heute abend könnte ich es ihm sagen.	Tonight I could tell him.
a05	Das schwarze Stück Papier befindet sich da oben neben dem Holzstück.	The black sheet of paper is located up there besides the piece of timber.
a07	In sieben Stunden wird es soweit sein.	In seven hours it will be.
b01	Was sind denn das für Tüten, die da unter dem Tisch stehen?	What about the bags standing there under the table?
b02	Sie haben es gerade hochgetragen und jetzt gehen sie wieder runter.	They just carried it upstairs and now they are going down again.
b03	An den Wochenenden bin ich jetzt immer nach Hause gefahren und habe Agnes besucht.	Currently at the weekends, I always went home and saw Agnes.
b09	Ich will das eben wegbringen und dann mit Karl was trinken gehen.	I will just discard this and then go for a drink with Karl.
b10	Die wird auf dem Platz sein, wo wir sie immer hinlegen.	It will be in the place where we always store it.

3.2. FAU AIBO database

The FAU AIBO corpus [15] contains recordings from children interacting with SONY's pet robot AIBO. The most important characteristic of these recordings is they are natural with non-acted emotions. The children were invited to play with the AIBO and

asked to guide it through certain missions, such as moving from point A to point B along a particular route. Children believed that the AIBO would have responded to their commands directly, whereas it was actually controlled by a human operator to behave excellently or disobediently, thereby to provoke emotional reactions. The data was collected from two different German schools, **M**ont and **O**hm, from 51 children (of age 10~13; 21 boys and 30 girls). Speaker independence is assured by using the data from one school for training and the data from another school for testing. The original recordings are sampled at 16k Hz. For the same reason as stated in section 3.1, speech samples are downsampled to 8k Hz. The original recordings with pause length over 1 sec were segmented automatically into "turns". Five labelers (advanced students of linguistics) annotated each turns in word-level as neutral (default) or as one of ten other emotion classes. Majority voting (MV) was then used, that is, only those words with three or more than three labelers' agreement were included into the corpus. The classes and number of speech samples in each class were: *joyful* (101), *surprised* (0), *emphatic* (2,528), *helpless* (3), *touchy* (225), *angry* (84), *motherese* (1,260), *bored* (11), *reprimanding* (310), *rest* (3), *neutral* (39,169).

We follow the INTERSPEECH 2009 emotion challenge [18] criterions which differentiate the classification problem into a five-class problem and a two-class problem. For the five-class classification problem, emotions are grouped into **A**nger (*angry*, *touchy*, and *reprimanding*), **E**mphatic, **N**eutral, **P**ositive (*motherese* and *joyful*), and **R**est. The two-class problem deals with **NEG**ative (subsuming *angry*, *touchy*, *reprimanding*, and *emphatic*) and **IDL**e (consisting of all nonnegative states) emotions. More details about numbers of speech samples are listed in Table 3-2 and Table 3-3. Similar to section 3.1, white and babble noises are added to the original clean speech to test the robustness of various features.

Table 3- 2 Number of instances for the 5-class problem

#	**A**	**E**	**N**	**P**	**R**	sum
train	881	2093	5590	674	721	9959
test	611	1508	5377	215	546	8257
sum	1492	3601	10967	889	1267	18216

Table 3- 3 Number of instances for the 2-class problem

#	**NEG**	**IDL**	sum
train	3358	6601	9959
test	2465	5792	8257
sum	5823	12393	18216

3.3. Rate-Scale (RS) Features

As mentioned in section 2.1.3, rate-scale plots reveal joint spectro-temporal modulations of the speech. The slow modulations, which are related to the speaking rate (i.e., the changing rate of the vocal track), are shown in low rate regions. On the other hand, the energy of resolved pitch is captured in high rate regions. In this study, we consider rates at $\pm 2^{1,\cdots,9}$ Hz to cover the complete temporal structures (speaking rate and pitch) of the speech. As for the scale region, we emphasize on the $2^{-1,\cdots,3}$ cycle/octave to cover complete frequency structures, from formants (captured by low scales) to harmonics (captured by high scales). Therefore, 90 rate-scale features (9 rates, 5 scales and both directions) are

extracted per frame. The mean and standard deviation of these 90 RS features are then calculated over the entire utterance. Finally, 180 RS features per utterance are preserved for emotion recognition.

3.4. MFCC Features

The mel-frequency cepstral coefficients (MFCCs) are widely used in the speech analysis field. Here, the first 13 MFCCs (including the zero-order coefficient) are extracted from 25 ms Hamming-windowed frame every 10 ms with the pre-emphasis coefficient 0.97. The mean, standard deviation, skewness, and kurtosis of these 13 MFCCs, their deltas, and double-deltas are computed as 156 features per utterance. It is referred to as MFCC156.

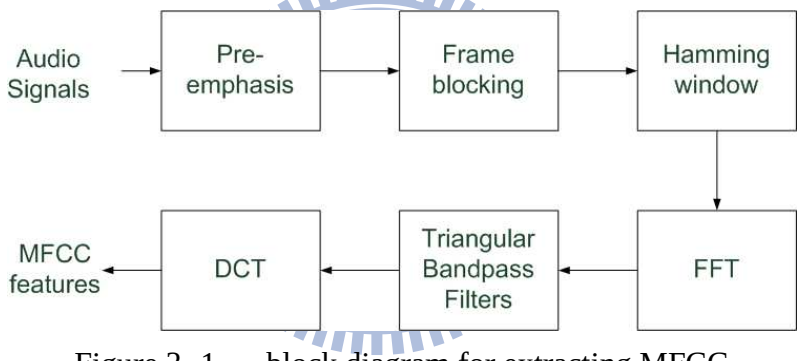


Figure 3- 1 block diagram for extracting MFCC

3.5. Prosodic Features

The 180 RS features mentioned above contain pitch and timbre (i.e., the formant structure) information, however, conventional MFCCs only carry timbre information. To make a fair comparison, prosodic features (pitch, energy and duration) are extracted and combined with MFCC features.

The fundamental frequency (F0) contour is extracted by STRAIGHT [19]. The algorithm estimates the aperiodic power (AP) of each frame. Frames with high AP are assumed unvoiced with zero F0. Only low-AP frames are treated as voiced frames and

return valid F0 estimate. The energy contour is extracted every 10 ms with a 25 ms window. Duration related features are derived from the voiced/unvoiced discrepancy obtained in F0 estimation.

Statistics of these prosodic features used in this study are similar to those used by other researchers [3, 4]. However, not to form a huge feature set with 1000 ~ 4000 parameters, a reasonably small-sized feature set is constructed. As a result, some features are omitted or replaced. For example, the mean of the positive and the negative dF0 are calculated separately to represent the upward and the downward trend, respectively, instead of the mean of all dF0. As for the energy, the minimum value of energy must be close to zero such that the min value, relative position of min, and range would not provide crucial information and hence are dropped from our feature list. Finally, 30 prosodic features are extracted and referred to as the PRO30 feature set. The description of this feature set is given in Table 3-4.

Table 3- 4 30 prosodic features

F0 (8 features)	mean, std, max value, relative position of max, min value, relative position of min, range, number of local max point
dF0 (8 features)	mean of positive, mean of negative, std, max value, relative position of max, min value, relative position of min, ratio of positive
logE (3 features)	std, max value, relative position of max
dlogE (8 features)	mean of positive, mean of negative, std, max value, relative position of max, min value, relative position of min, ratio of positive
Duration (3 features)	speaking rate, std of voiced duration, mean pause time

3.6. INTERSPEECH 2009 Emotion Challenge Acoustic Features

For the AIBO database, we compare the acoustic features adopted in INTERSPEECH 2009 emotion challenge with proposed RS features under noisy conditions. This default feature set provides baseline results for both HMM and linear kernel SVM recognizers in the 2009 challenge and is totally transparent with the accessible open source openSMILE feature extraction toolkit [20]. It includes the most common features in pertaining to prosody, spectral shape, voice quality, as well as their derivatives. In details, the 16 low-level descriptors chosen are: zero-crossing-rate (ZCR) from the time signal, root mean square (RMS) frame energy, pitch frequency (normalized to 500 Hz), harmonics-to-noise ratio (HNR) by autocorrelation function, and mel-frequency cepstral coefficients (MFCC) 1-12 in full accordance to HTK-based computation. To each of these 16 features, the delta coefficients are included as well. Next, as depicted in Table 3-5, the 12 functionals: mean; standard deviation; kurtosis; skewness; minimum and maximum value, relative position, and range; and two linear regression coefficients with their mean square error (MSE); are derived for each low-level and its delta feature on a chunk basis. Thus, the final feature contains 16×2×12 = 384 attributes and is referred to as the Inter384 features. In this thesis, we conduct experiments in section 4.3 to compare Inter384 features with proposed RS features in their robustness.

Table 3- 5 Features used in INTERSPEECH 2009 emotion challenge

LLD (16*2)	Functionals (12)
(Δ) ZCR	mean
(Δ) RMS	Energy standard deviation
(Δ) F0	kurtosis, skewness
(Δ) HNR	extremes: value, rel. position, range
(Δ) MFCC 1-12	linear regression: offset, slope, MSE

Chapter 4

Simulation Result

4.1. Experimental Setup

As mentioned in section 2.2, there are many kinds of kernels available for the SVM to map problems onto higher dimensional spaces. Although the radial basis function (RBF) kernel is suggested to use the first, different choices of parameters C and γ would affect results radically [16]. These parameters need to be fine tuned by the grid search for each training condition. Therefore, a simpler linear kernel is adopted in this study only to investigate the robustness of features. Before building the SVM, all training and testing features are linearly scaled to [0, 1]. To evaluate the robustness of RS features in unknown environments, mismatched tests (clean data for training while noisy data for testing) are performed under various SNR conditions.

To address the problem of insufficient speech samples in the Berlin database, the 10-fold cross-validation procedures are adopted in our test. Speech samples are randomly divided into 10 subsets. In each trial, one subset is used for testing while the other nine subsets are used for training the SVM recognizer. Final recognition rates are obtained by averaging over 10 trials. Features extracted from the Berlin database will further processed through the intra-speaker normalization. That is, for each speaker, features from all

sentences, including seven emotion classes, are normalized by their mean and standard deviation. As for the FAU AIBO database evaluation, the 10-fold method is not utilized due to its sufficient data samples. The data of one school, ***O**hm*, is used for training and the data of another school, ***M**ont*, is used for testing. Therefore, speaker independence is assured for the FAU AIBO database evaluation since there is no overlap between training speakers and testing speakers.

Recognition results are reported in form of the total recognition rate (RR), the mean of class-wise recognition rate (CL) and the harmonic mean F where

$$F = \frac{2 \cdot RR \cdot CL}{RR + CL} \tag{4-1}$$

These three different measures are assessed for cases with unbalanced number of instances among classes. The classes with more instances have more substantial influence on RR than ones with fewer instances. Thus, the RR measure has the tendency of over-estimating the performance. On the contrary, the CL measure increases the influence of minority classes thus under-estimating the performance. Therefore, the F-measure is commonly used to give a fair performance estimate when sizes of classes are not balanced [21]. However, since the FAU AIBO database is severely unbalanced, the classifier loses its detecting ability against those minority classes. To cope with this problem, an under-sampling method is used in majority classes. We randomly down-sample other classes to have the same number of instances as the smallest class, which is the **NEG** in the 2-class problem and the **P** in the 5-class problem. Final recognition rates are obtained by averaging over 10 trials. In this totally balanced condition, the RR and CL measures produce the same results; hence, we only list one measure for the FAU AIBO database.

4.2. Results on Berlin Database

Table 4- 1 Recognition rates (in %) of RS180 under additive white noises

RS180	H	A	S	F	N	B	D	CL	RR	F
clean	42.86	91.35	100.00	52.86	81.96	81.25	45.50	70.83	74.32	72.53
20dB	42.86	88.14	100.00	52.86	80.71	81.25	48.00	70.55	73.57	72.03
15dB	44.29	87.31	98.33	51.43	79.29	81.25	47.50	69.91	72.98	71.41
10dB	44.29	84.87	98.33	52.86	76.61	80.00	54.50	70.21	72.60	71.38
5dB	50.00	83.21	91.90	53.57	72.86	80.00	61.00	70.36	72.22	71.28
0dB	45.71	64.17	75.71	50.71	78.04	65.00	57.00	62.33	62.91	62.62

Table 4- 2 Recognition rates (in %) of MFCC156+PRO30 under additive white noises

MFCC156 +PRO30	H	A	S	F	N	B	D	CL	RR	F
clean	65.71	84.36	98.33	70.95	95.00	90.00	71.00	82.19	83.09	82.64
20dB	40.00	84.36	91.90	26.67	65.36	91.25	68.50	66.86	69.03	67.93
15dB	52.86	73.91	78.81	30.95	74.82	85.00	47.00	63.34	65.99	64.64
10dB	37.14	76.41	77.14	28.57	53.93	83.75	56.50	59.06	61.42	60.22
5dB	64.29	58.91	79.05	10.48	24.46	86.25	42.50	52.28	53.62	52.94
0dB	47.14	51.79	64.05	17.86	42.50	92.50	42.50	51.19	52.44	51.81

Table 4- 3 Recognition rates (in %) of RS180 under additive babble noises

RS180	H	A	S	F	N	B	D	CL	RR	F
clean	44.29	91.28	100.00	50.95	81.07	90.00	47.00	72.08	75.48	73.74
20dB	45.71	90.45	100.00	52.62	82.14	87.50	47.00	72.20	75.49	73.81
15dB	44.29	88.85	100.00	55.48	80.89	85.00	46.50	71.57	74.73	73.12
10dB	45.71	84.17	100.00	54.29	79.64	78.75	44.00	69.51	72.25	70.85
5dB	42.86	61.15	100.00	46.43	92.50	61.25	44.50	64.10	64.62	64.36
0dB	35.71	15.77	100.00	24.29	87.50	52.50	62.00	53.97	49.46	51.62

Table 4- 4 Recognition rates (in %) of MFCC156+PRO30 under additive babble noises

MFCC156 +PRO30	H	A	S	F	N	B	D	CL	RR	F
clean	61.43	83.40	100.00	75.71	92.14	88.75	69.50	81.56	82.54	82.05
20dB	40.00	87.31	81.90	59.29	79.11	87.50	59.00	70.59	73.50	72.01
15dB	45.71	81.79	59.76	46.67	68.57	80.00	66.00	64.07	66.55	65.29
10dB	45.71	76.09	65.71	33.57	49.64	83.75	57.50	58.85	61.38	60.09
5dB	77.14	63.91	50.00	18.33	51.25	73.75	50.00	54.91	56.78	55.83
0dB	68.57	37.50	19.52	13.81	21.43	57.50	40.00	36.90	37.53	37.21

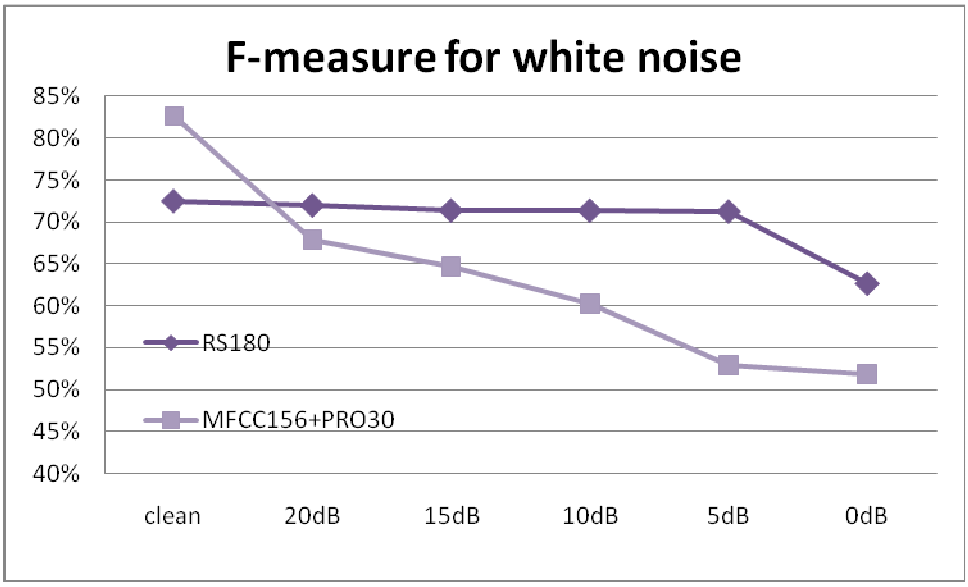


Figure 4- 1 F-measure for additive white noise

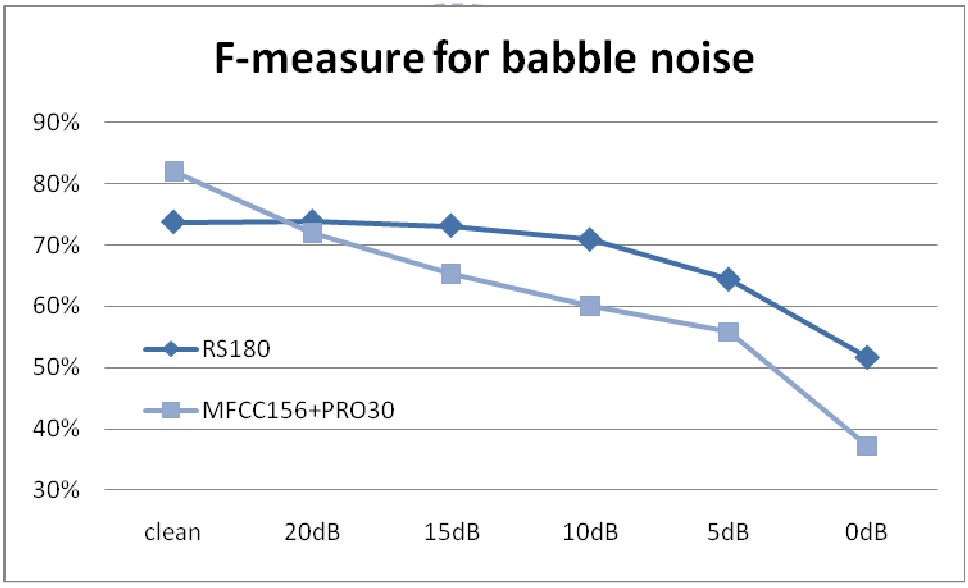


Figure 4- 2 F-measure for additive babble noise

Table 4-1 to 4-4 show detailed performance of using RS180 and MFCC156+PRO30 features in additive white and babble noises, respectively. The class-wise (from H to D) recognition rates are shown in each column. The CL is the mean of class-wise recognition rates and the RR is the total recognition rate. The F-measure, which provides a fair comparison, is given in the last column in each table and summarized in Figure 4-1 and 4-2.

Clearly, the RS180 outperforms the MFCC156+PRO30 in all SNR conditions (20dB~0dB), except in the clean condition. The MFCC156+PRO30 features from training samples depict magnitude spectra and pitch values with high precision. Such precise representations would produce good matches in clean condition, but are also prone to degradations by noises. On the other hand, RS features only carry the information of spectro-temporal amplitude modulations, which is equivalent to the spectro-temporal envelopes without carriers' fine structure (phase) information. While not providing accurate matches in the clean condition, RS features are more resistant to deteriorations from spectro-temporal envelopes of noises.

Using RS features, Anger and Sadness are the two most recognizable emotions, whereas Fear, Disgust and Happiness are more difficult to be classified. With additive background noises, Fear emotion is particularly prone to be deteriorated in both feature domains of MFCC plus prosodic and RS features. Moreover, recognition rates of the emotion of Neutral are severely degraded under noisy conditions when using conventional features. Nevertheless, it is very well preserved by the RS features.

Figure 4-3 shows an example of sentence spoken by the same speaker with Anger and Neutral emotions. Panel (a) and (c) are the auditory spectrograms of utterances with Anger and Neutral emotions, respectively. Panel (b) and (d) are their corresponding rate-scale plots. As seen in these figures, the pitch-related response (high rate, high scale) of Neutral is more intense than that of Anger. The reason for this phenomenon is that the speaker's pitch is moving up-and-down more dramatically in Anger emotion than in Neutral emotion. Hence, the mean response at each specific pitch-related rate-scale point in Anger emotion is weaker than that in Neutral emotion. On the other hand, the low rate region encodes the coarse temporal AM structure of the utterance. In Neutral speech, pitch and formant contours are usually with smooth declination toward the end of sentence. This declination trend is revealed as a notable positive rate (downward) response as mentioned in section 2.1.3. However, no declination trend in Anger speech produces comparable response in positive

and negative rates.

For better demonstrating the robustness of our RS features, Figure 4-4 shows the response curves of a pitch-related region (rate=256 Hz, scale=4 cycle/octave) and an AM-related region (rate=4 Hz, scale=0.25 cycle/octave) along the time axis under clean (panel (a) and (b)) and 5 dB noisy conditions (panel (c) and (d)). Both curves are derived from the Anger sentence used in Figure 4-3. Figure 4-5 shows the spectrogram and RS plot of white noise alone. As observed in Figure 4-5 (b), the white noise activates a high rate/high scale response, which is quite different from the response of speech. For speech with added white noise, the low rate/low scale regions are less affected (see Figure 4-4 (b) and (d) for clean and 5 dB SNR condition) while the pitch-related RS regions are more affected. However, comparing Figure 4-4 (a) and (c), distortions are roughly as from a dynamic range compression. The original trend along the time axis is not damaged. On the contrary, conventional ways of extracting pitch may totally become invalid with low SNR noise. The similar trend can also be observed in one Neutral sentence as shown in Figure 4-6.

Figure 4-7 and 4-8 show the distributions of the specific pitch-related RS feature (rate=256 Hz, scale=4 cycle/octave) under clean and 5dB noisy conditions, respectively. The distributions are derived from the same sentences used in Figure 4-4 and 4-6. Response for Neutral is greater than that for Anger as we mentioned earlier. The effect by white noise does not cause dramatic damage but only a slight shift to the distributions. These figures give ideas about the superior performance of our RS features to conventional MFCCs plus prosodic features in low SNR conditions.

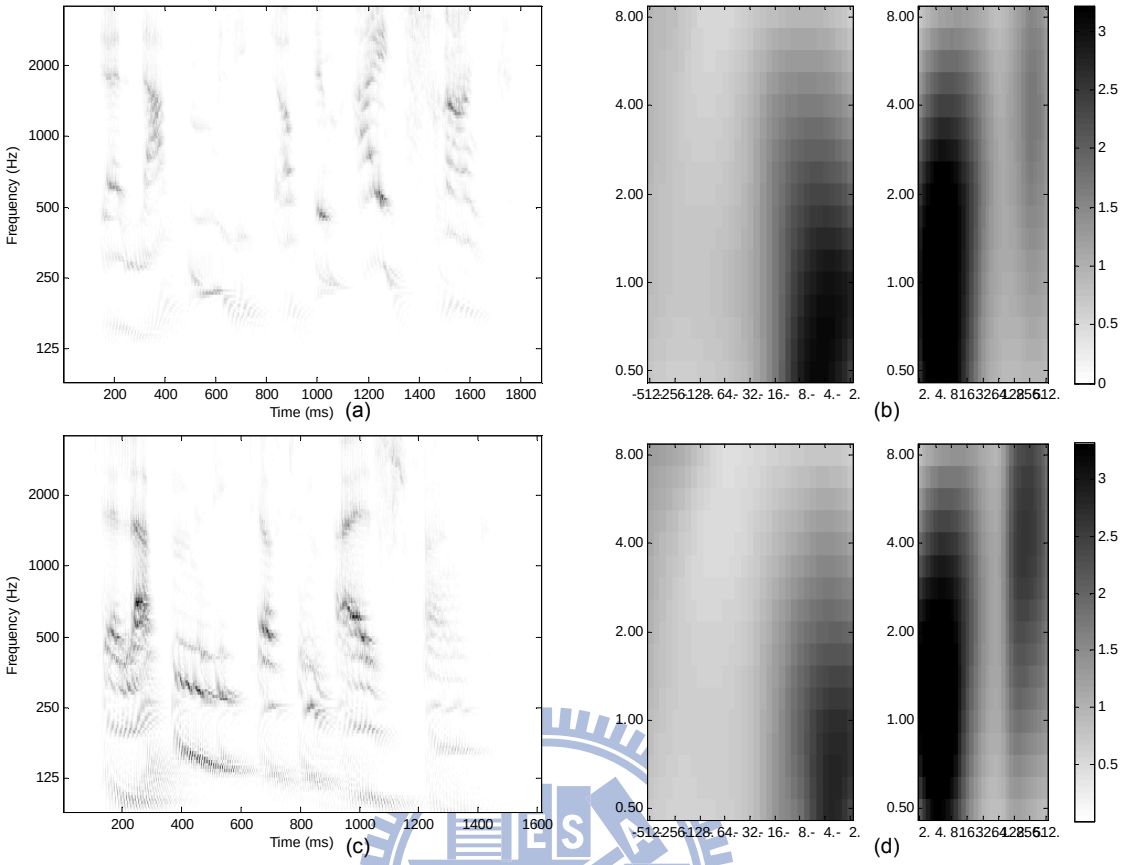


Figure 4- 3 (a), (b) spectrogram and RS plot of a Berlin Anger sentence; (c), (d) spectrogram and RS plot of the same sentence with Neutral emotion.

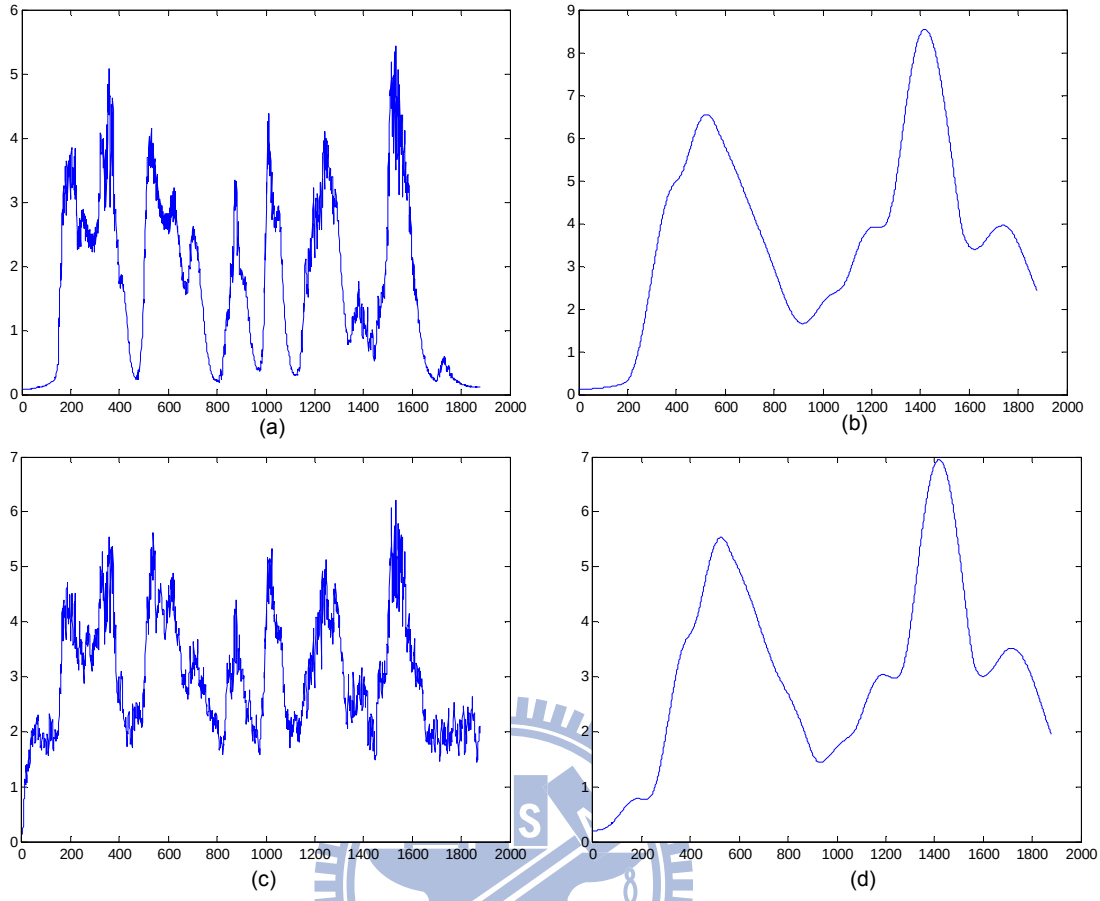


Figure 4- 4 One Berlin Anger sentence: (a) and (b) depict a high rate/high scale (pitch-related) response and a low rate/low scale (AM-related) response plotted along the time axis under clean condition; (c) and (d) depict the responses of the same rate-scale combinations as in (a) and (b) under 5dB noisy condition

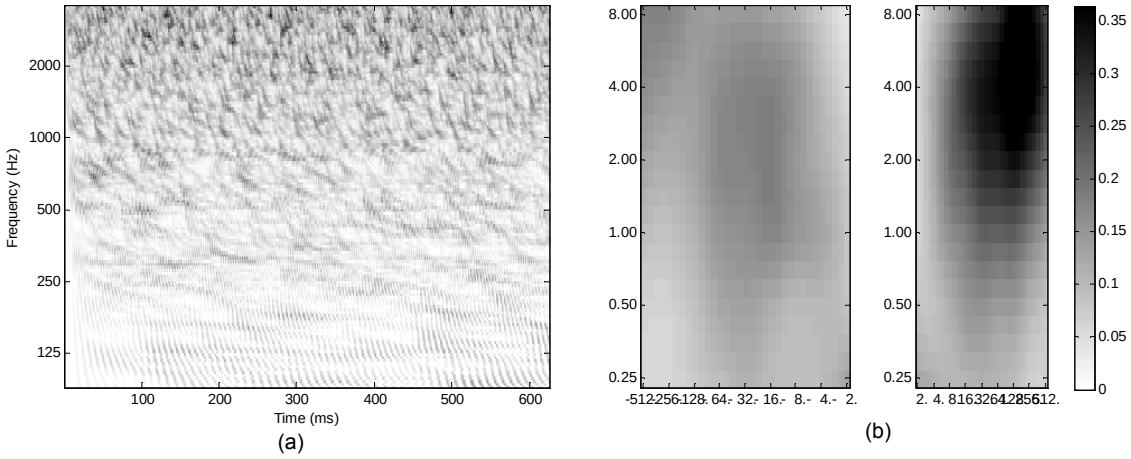


Figure 4- 5 white noise: (a) spectrogram, (b) RS plot

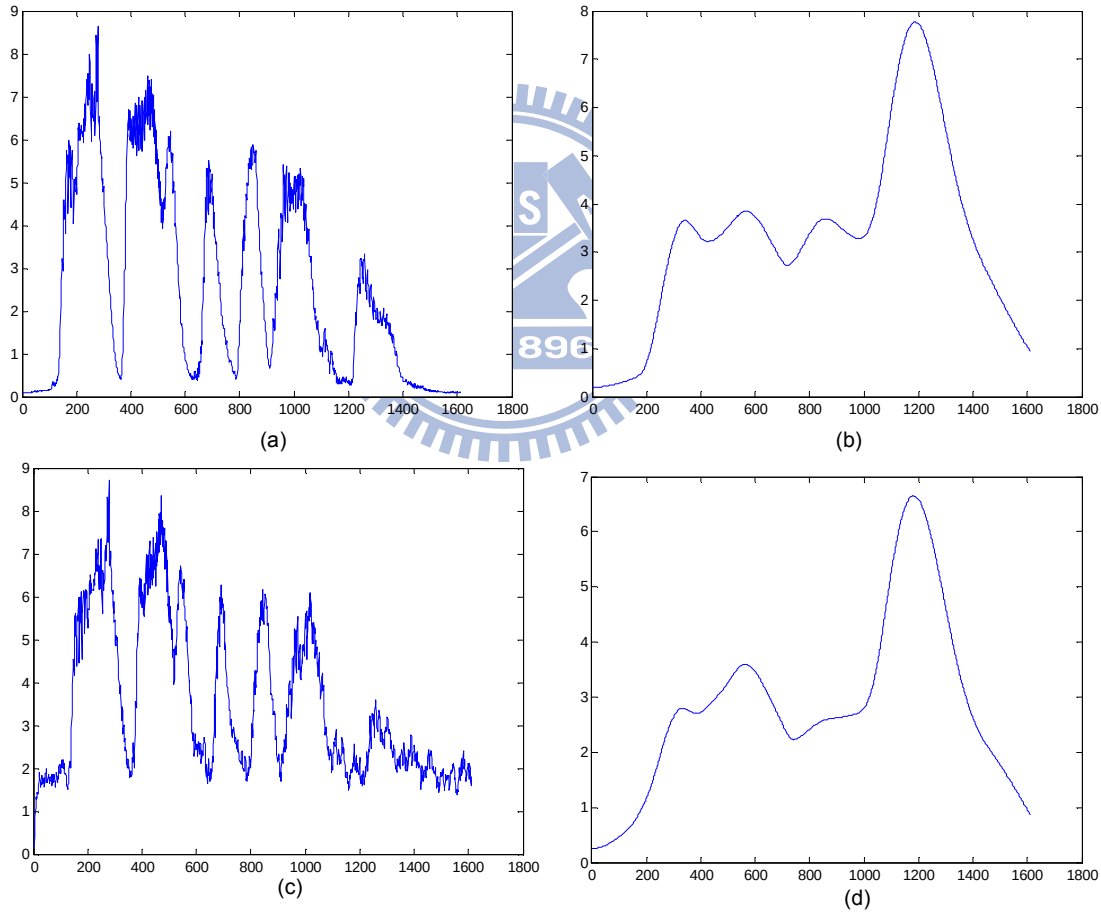


Figure 4- 6 One Berlin Neutral sentence: (a) and (b) depict a high rate/high scale (pitch-related) response and a low rate/low scale (AM-related) response plotted along the time axis under clean condition; (c) and (d) depict the responses of the same rate-scale combinations as in (a) and (b) under 5dB noisy condition

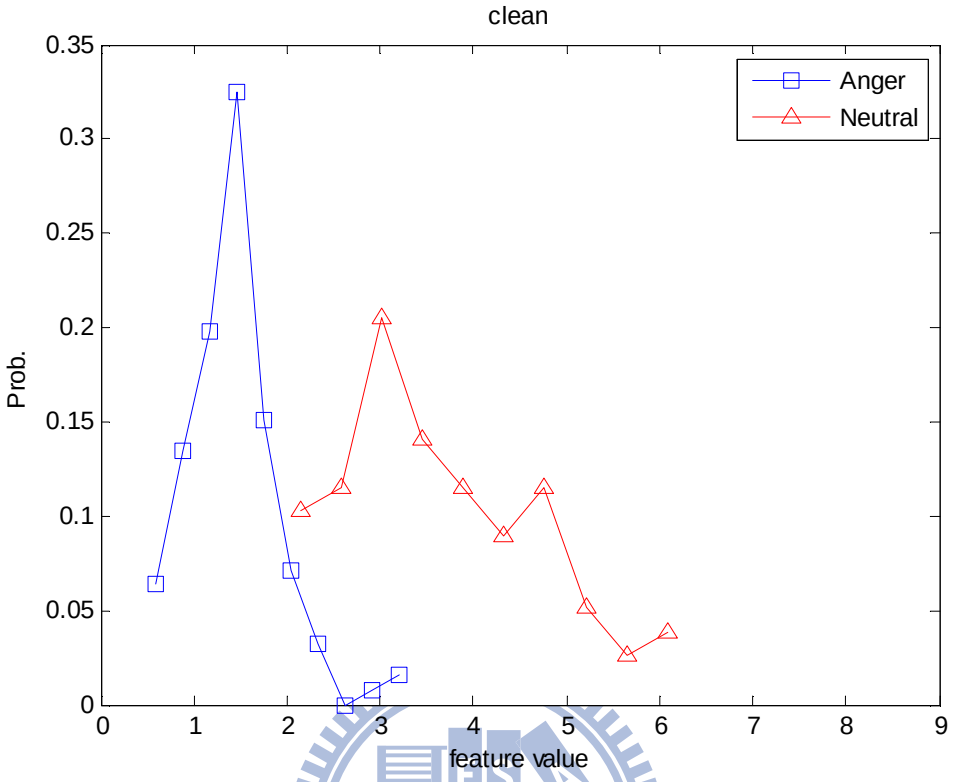


Figure 4- 7 The distribution of a pitch-related feature (rate=256 Hz, scale=4 cycle/octave) under clean condition

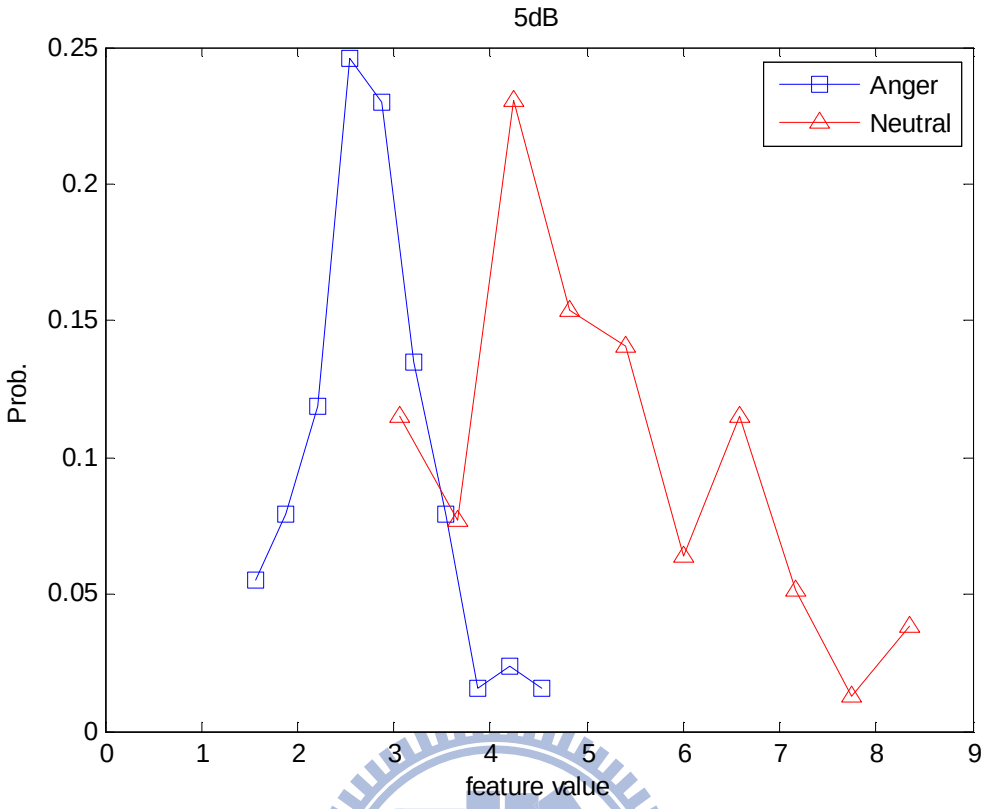


Figure 4- 8 The distribution of a pitch-related feature (rate=256 Hz, scale=4 cycle/octave) under 5dB noisy condition

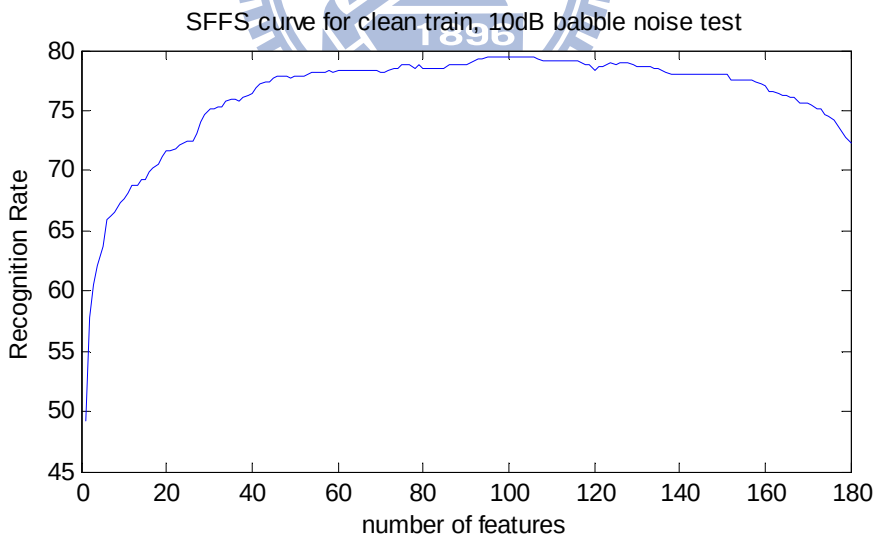


Figure 4- 9 Recognition rate (in %) of RS180 by SFFS method

A feature selection method, sequential forward floating selection (SFFS) [22], is used to examine contributions within RS180 features. It starts from an empty feature set and sequentially includes (or excludes) a feature into the selected set, then evaluates the

performance of newly constructed feature set. As shown in Figure 4-9, the performance peaks around using 100 features and does not vary a lot from using 60 to 140 features. Tests on other SNR conditions have the similar trend. These results simply imply our RS features are highly redundant, which is not unexpected due to the highly overlapped two-dimensional filters in the cortical module [2]. Therefore, RS180 can be further downsampled to RS92 by choosing rate-scale combinations of gray spots in Figure 4-10. Note, only downward direction (positive rate) is shown in the figure.

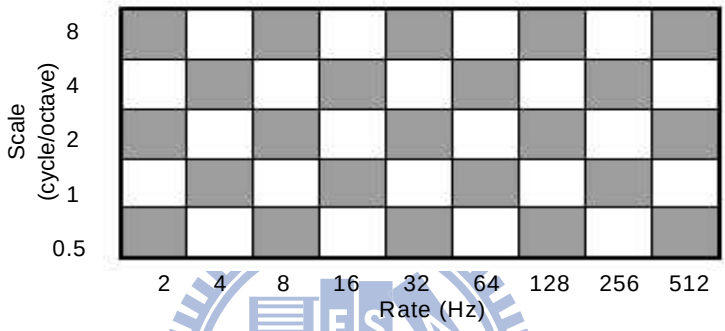


Figure 4- 10 Rate-scale selections (gray areas) of RS92

Two subsets of MFCC156 are selected to compare with our reduced RS92 features. The first subset (MFCC78) contains the mean and standard deviation of 13 MFCCs, 13 ΔMFCCs and 13 ΔΔMFCCs. The second subset (MFCC52) contains the mean, standard deviation, skewness, and kurtosis of 13 MFCCs. Both subsets are then combined with PRO30 features to have comparable feature numbers as RS92. Detailed recognition rates are shown from Table 4-5 to 4-10, and the overall F-measures are given in Figure 4-11 and Figure 4-12. Results show that RS92 has almost the same performance as RS180 in white noise while performs slightly worse in high-SNR (20, 15 dB) babble noise. The reason for that is white noise has vastly different spectro-temporal modulations from speech, while the babble noise has similar modulations to speech. Hence, a higher resolution in the RS domain is preferred for babble noise. Nevertheless, RS92 outperforms MFCC78+PRO30 and MFCC52+PRO30 in almost all SNR conditions, especially in low SNR conditions.

Table 4- 5 Recognition rates (in %) of RS92 under additive white noises

RS92	H	A	S	F	N	B	D	CL	RR	F
clean	40.00	92.12	98.57	54.76	75.36	81.25	41.00	69.01	72.79	70.85
20dB	45.71	90.58	97.14	53.10	76.61	82.50	41.00	69.52	73.18	71.30
15dB	48.57	88.97	97.14	51.67	75.18	82.50	41.00	69.29	72.80	71.00
10dB	45.71	86.54	95.24	53.57	73.93	80.00	45.50	68.64	71.66	70.12
5dB	50.00	83.27	90.24	53.81	73.93	80.00	47.50	68.39	71.06	69.70
0dB	50.00	72.88	76.19	50.71	80.54	60.00	50.50	62.97	64.63	63.79

Table 4- 6 Recognition rates (in %) of MFCC78+PRO30 under additive white noises

MFCC78 +PRO30	H	A	S	F	N	B	D	CL	RR	F
clean	70.00	88.97	96.67	74.76	91.25	91.25	71.00	83.41	84.81	84.11
20dB	51.43	78.65	88.81	47.86	70.71	86.25	65.00	69.82	71.27	70.54
15dB	67.14	68.33	86.90	39.05	74.64	85.00	50.50	67.37	68.83	68.09
10dB	50.00	58.97	86.90	38.81	56.61	86.25	58.00	62.22	62.36	62.29
5dB	67.14	50.00	86.90	16.67	28.04	81.25	61.00	55.86	54.94	55.39
0dB	51.43	39.17	78.81	21.19	46.25	91.25	36.50	52.09	51.93	52.01

Table 4- 7 Recognition rates (in%) of MFCC52+PRO30 under additive white noises

MFCC52 +PRO30	H	A	S	F	N	B	D	CL	RR	F
clean	67.14	89.10	92.14	80.48	83.57	86.25	68.00	80.95	82.83	81.88
20dB	45.71	78.91	85.24	50.00	81.07	90.00	70.00	71.56	72.88	72.21
15dB	57.14	66.79	91.90	37.38	81.96	90.00	53.50	68.38	69.59	68.98
10dB	51.43	62.95	98.33	42.14	83.21	80.00	66.50	69.22	68.98	69.10
5dB	54.29	65.32	90.24	15.00	61.96	72.50	61.00	60.04	60.65	60.35
0dB	54.29	48.59	87.38	24.29	68.21	82.50	45.00	58.61	58.56	58.58

Table 4- 8 Recognition rates (in %) of RS92 under additive babble noises

RS92	H	A	S	F	N	B	D	CL	RR	F
clean	35.71	92.82	100.00	52.62	71.96	87.50	37.50	68.30	72.44	70.31
20dB	34.29	91.22	100.00	55.71	73.21	85.00	40.00	68.49	72.26	70.32
15dB	31.43	90.45	100.00	58.81	71.96	85.00	42.50	68.59	72.07	70.29
10dB	37.14	88.14	100.00	55.95	78.21	72.50	46.50	68.35	71.30	69.79
5dB	31.43	72.12	100.00	46.67	91.25	60.00	37.00	62.64	64.80	63.70
0dB	24.29	28.46	100.00	22.62	91.25	50.00	50.50	52.45	50.01	51.20

Table 4- 9 Recognition rates (in %) of MFCC78+PRO30 under additive babble noises

MFCC78 +PRO30	H	A	S	F	N	B	D	CL	RR	F
clean	70.00	85.83	96.90	77.86	88.21	87.50	68.00	82.04	83.11	82.57
20dB	51.43	78.46	88.57	62.38	82.86	87.50	47.00	71.17	73.57	72.35
15dB	54.29	74.74	76.43	53.10	84.82	86.25	57.50	69.59	71.29	70.43
10dB	75.71	49.23	77.38	44.05	59.82	83.75	54.50	63.49	62.73	63.11
5dB	80.00	39.94	80.48	30.95	45.89	72.50	49.50	57.04	55.54	56.28
0dB	65.71	29.36	57.86	20.24	29.46	56.25	54.50	44.77	42.58	43.65

Table 4- 10 Recognition rates (in %) of MFCC52+PRO30 under additive white noises

MFCC52 +PRO30	H	A	S	F	N	B	D	CL	RR	F
clean	70.00	90.51	89.29	79.52	84.46	82.50	73.50	81.40	82.52	81.95
20dB	42.86	85.71	49.76	69.29	84.64	88.75	68.50	69.93	72.44	71.16
15dB	52.86	77.82	41.67	58.81	79.29	90.00	79.50	68.56	70.00	69.27
10dB	61.43	66.73	48.81	51.43	56.25	86.25	80.00	64.41	64.45	64.43
5dB	77.14	58.97	38.33	36.90	70.36	77.50	64.00	60.46	61.01	60.73
0dB	67.14	41.35	28.57	24.52	44.29	40.00	54.50	42.91	42.58	42.74

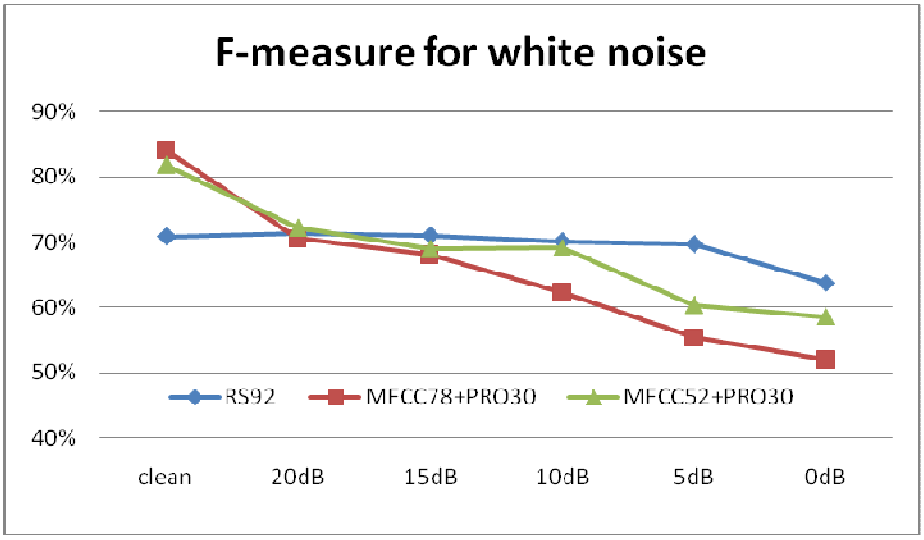


Figure 4- 11 F-measure for additive white noise (reduced features)

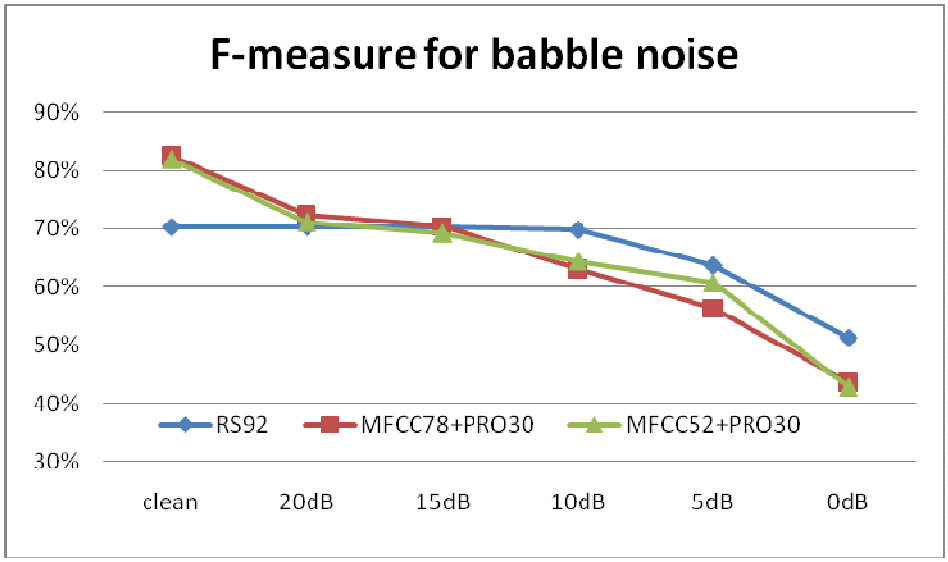


Figure 4- 12 F-measure for additive babble noise (reduced features)

4.3. Results on FAU AIBO Database

Table 4- 11 FAU AIBO database recognition rates (in %) under additive white noise: matched condition (noisy train/noisy test)

2-class	clean	15dB	10dB	5dB	0dB
Inter384	68.74	67.23	66.66	66.33	66.00
RS180	66.59	66.35	66.58	66.24	65.99
Inter384+RS180	68.49	66.67	66.90	66.72	66.32
5-class	clean	15dB	10dB	5dB	0dB
Inter384	38.19	39.37	38.61	37.41	34.99
RS180	38.65	38.74	38.43	37.66	37.55
Inter384+RS180	39.40	39.90	39.90	38.12	35.13

Table 4- 12 FAU AIBO database recognition rates (in %) under additive babble noise: matched condition (noisy train/noisy test)

2-class	clean	15dB	10dB	5dB	0dB
Inter384	68.74	67.77	67.45	67.24	66.41
RS180	66.59	66.17	66.20	66.34	65.50
Inter384+RS180	68.49	67.89	67.36	66.84	66.32
5-class	clean	15dB	10dB	5dB	0dB
Inter384	38.19	38.33	37.86	37.28	35.87
RS180	38.65	38.50	37.64	37.53	36.42
Inter384+RS180	39.40	38.35	38.73	37.27	36.81

For the FAU AIBO database, table 4-11 and 4-12 show recognition rates under additive white and babble noises, respectively. These results are measured in the matched (noisy train/noisy test) condition, which is commonly inspected in other speech emotion recognition researches [12, 13]. Under such condition, the performance is measured by training and testing the classifier with the same SNR noisy speech samples. As seen in Table 4-11 and 4-12, any tested feature sets produce similar performance in the matched condition since characteristics of noises are considered to be well trained into the classifier. However, recognizers built for matched conditions are not practical since the SNR is usually varying in testing environments.

Table 4- 13 FAU AIBO database recognition rates (in %) under additive white noises: mismatched condition (clean train/noisy test)

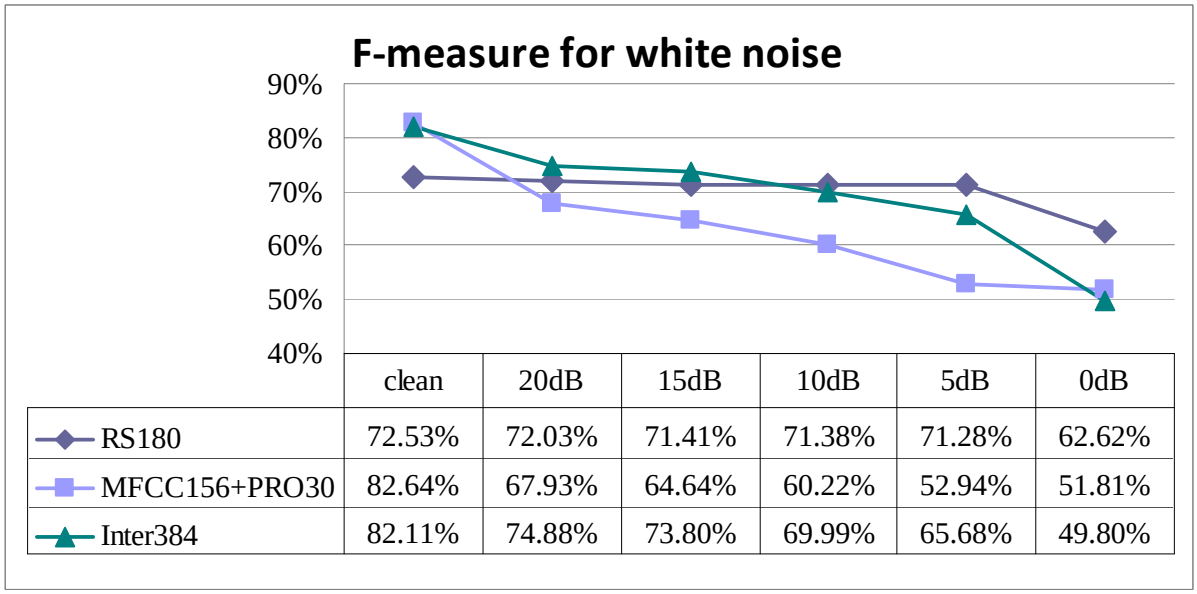
2-class	clean	15dB	10dB	5dB	0dB
Inter384	68.74	63.78	60.92	59.07	55.00
RS180	66.59	60.16	52.82	50.37	50.00
RS20	66.59	66.66	66.35	64.65	61.49
Inter20	66.65	66.41	67.73	66.91	62.24
(RS+Inter)20	67.27	67.36	67.54	66.92	66.10
5-class	clean	15dB	10dB	5dB	0dB
Inter384	38.19	31.14	29.05	23.86	22.13
RS180	38.65	24.58	20.14	20.00	20.00
RS20	37.50	37.80	37.53	32.66	31.18
Inter20	36.58	38.50	38.92	36.47	33.09
(RS+Inter)20	36.60	38.00	37.15	36.45	32.84

Table 4- 14 FAU AIBO database recognition rates (in %) under additive babble noises: mismatched condition (clean train/noisy test)

2-class	clean	15dB	10dB	5dB	0dB
Inter384	68.74	57.50	53.12	50.92	50.00
RS180	66.59	51.81	50.06	50.00	50.00
RS20	66.59	63.65	59.17	52.64	50.39
Inter20	66.65	67.00	65.96	65.70	63.51
(RS+Inter)20	67.27	67.32	66.84	65.97	59.05
5-class	clean	15dB	10dB	5dB	0dB
Inter384	38.19	28.17	25.84	23.45	20.81
RS180	38.65	21.16	20.07	20.00	20.00
RS20	37.50	35.92	32.22	26.38	21.15
Inter20	36.58	34.18	32.42	28.87	25.82
(RS+Inter)20	36.60	37.00	35.56	29.88	23.63

Recognition rates under mismatched (clean train/noisy test) condition for the AIBO database are shown in Table 4-13 and 4-14. The Inter384 and RS180 feature sets are tested first. From the first glance, it seems right to conclude the Inter384 feature set outperforms our proposed RS180 feature sets in noisy conditions. However, after taking a deeper

inspection, one can see the class-wise recognition rates of RS180 are more skewed with lower SNR. Briefly speaking, the classifier predicts all testing data as from a certain class, causing the mean class-wise recognition rate approaching 50% in the 2-class problem and 20% in the 5-class problem. Not surprisingly, the number of support vectors of the SVM is almost the same as the number of training samples, which indicates that the SVM classifier is over-trained thus performs badly. In order to have a fair comparison, we further test the Inter384 feature set on the Berlin database, where our RS180 feature set does not cause the over-training of the recognizer, and show its F-measure in Figure 4-13 (extended from Figure 4-1). As can be seen, our proposed RS180 feature set still outperforms the Inter384 feature set in low SNR conditions (≤ 10 dB) when tested on the Berlin database.



	clean	20dB	15dB	10dB	5dB	0dB
RS180	72.53%	72.03%	71.41%	71.38%	71.28%	62.62%
MFCC156+PRO30	82.64%	67.93%	64.64%	60.22%	52.94%	51.81%
Inter384	82.11%	74.88%	73.80%	69.99%	65.68%	49.80%

Figure 4- 13 Performance of Inter384 on Berlin database (extended from Figure 4-1)

While both RS180 and Inter384 more or less over-train the SVM recognizer on the AIBO database, neither one of them makes the SVM over-trained on the Berlin database. One possible reason for this overtraining problem may be that the FAU AIBO database contains natural emotions in much shorter command-like sentences while the Berlin database contains acting emotions in longer sentences. Obviously, natural emotions are

much more difficult to recognize than acted emotions in our simulations especially with short sentences. Therefore, the RS180 calculated from the mean and standard deviation of the whole sentence might not carry consistent information among very short sentences, which causes an over-trained recognizer and produces poor performance especially in low SNR conditions. Figure 4-14 shows the histograms of sentence length of Anger and Neutral emotion samples in Berlin database and AIBO database respectively. As we can see, almost all sentences in the Berlin database last about two to five seconds (detail content shown in Table 3-1). On the other hand, while most sentences in AIBO are short commands lasting about zero to three seconds, some extremely long sentences exist. Most of these long sentences contain bad clips with long period of silence (Figure 4-15 (c), (d)). Moreover, some serious interference caused by microphone collision at the beginning or end of a sentence exists in long sentences. This kind of interference is treated as speech by the VAD system and included in the database (Figure 4-15 (a) 0 sec ~4 sec, (b) 5 sec ~ 9 sec). All of these artificial distortions shown in Figure 4-15 seriously affect the mean and standard deviation over the entire sentence, thus causing troubles to our proposed RS features.

On the other hand, we use the downsampling instead of the upsampling approach to balance the database. Such random downsampling, which ignores many speech sentences, might aggravate the over-training problem.

To tackle this problem, SFFS is further used to reduce the dimensionality of feature sets. The SFFS is conducted in the condition of using clean samples for training and 10dB white noisy samples for testing. The RS20 feature set contains the top 20 features selected from RS180 while the Inter20 feature set contains the top 20 features selected from Inter384. Additionally, the (RS+Inter)20 feature set holds the top 20 features selected from 564 features of combining Inter384 with RS180. With these three reduced feature sets, substantially improved performance in all SNR conditions is shown in gray areas in Table 4-13. Meanwhile, using the same reduced feature sets selected from 10 dB white noisy

samples, recognition rates under babble noise conditions are also improved as seen in gray areas in Table 4-14. These results infer that in real applications where some background noises are expected, the brute-force selection on large feature sets for each noise condition (as seen in [12, 13]) may not be the right approach because not the same features will be selected under different noisy conditions. Instead, selecting a smaller feature set under a *certain* noise condition beforehand may be a less time-consuming way to enhance the performance.

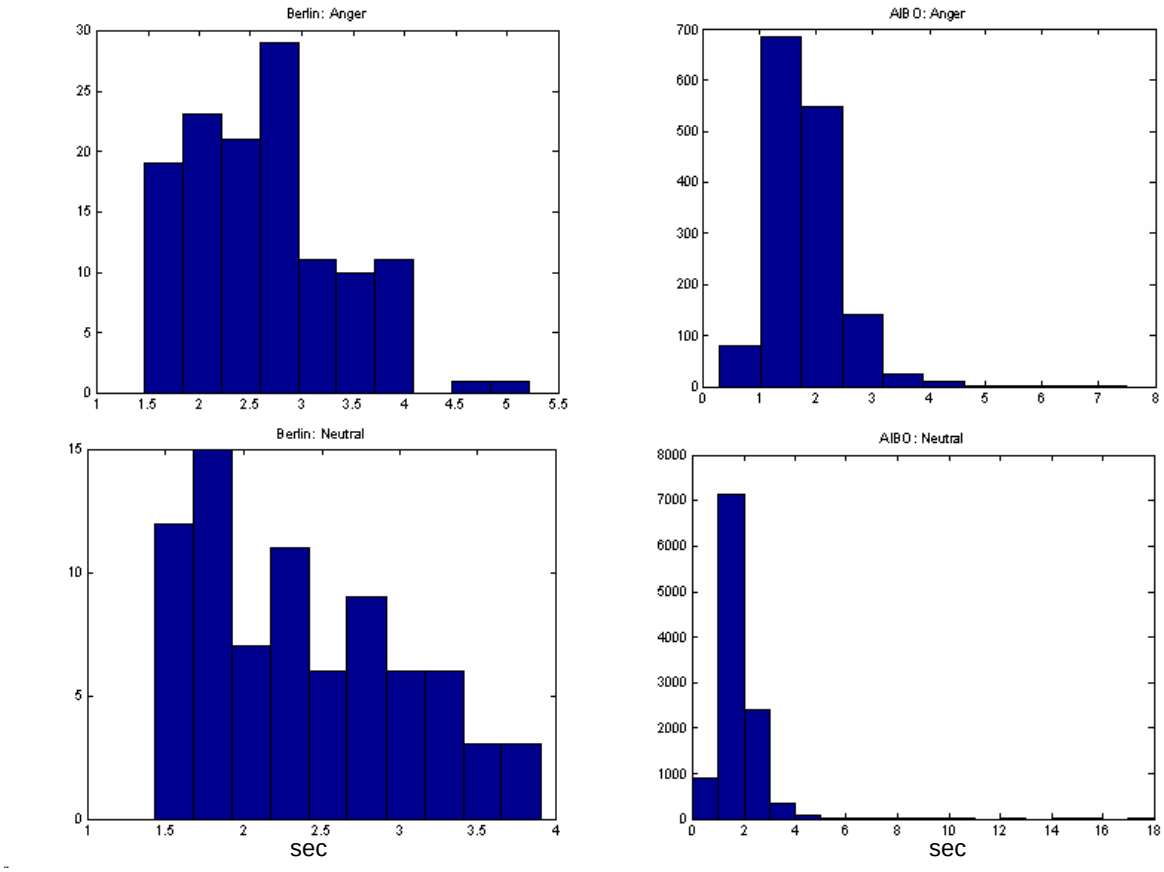


Figure 4- 14 sentence length of AIBO database

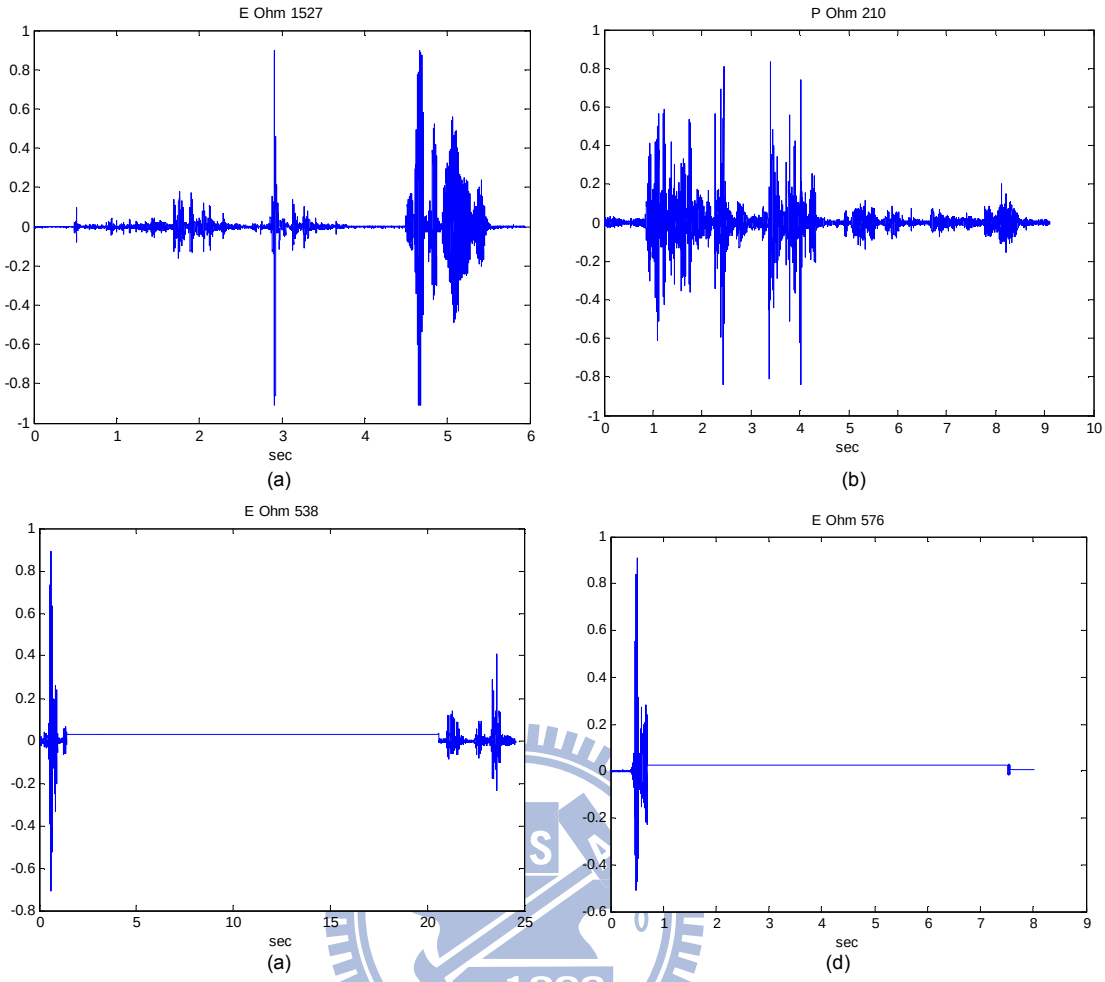


Figure 4- 15 Examples of bad clips in AIBO database

Chapter 5

Conclusion and Future Works

5.1. Conclusion

In the first part of this thesis, experiments on the Berlin database show that features from spectro-temporal modulations are more robust to additive white and babble noises than conventional MFCCs plus prosodic features in mismatched emotion recognition simulations, especially in low SNR (≤ 10 dB) conditions. Simulation results also show our proposed RS180 feature set outperforms the Inter384 feature set proposed in the INTERSPEECH 2009 emotion challenge in low SNR (≤ 10 dB) conditions.

For the AIBO database, experiments show that a serious over-fitting recognizer is constructed by using our proposed RS180 features and produces unsatisfactory performance especially in low SNR conditions. Feature selection method is further adopted to reduce the dimensionality in the hope to mitigate the over-fitting problem. Preliminary results by using reduced sets of 20 RS features or 20 Inter384 features demonstrate improved recognition rates although the best selection of features and the corresponding perceptual meaning of each feature are to be explored in the future.

Overall speaking, the spectro-temporal modulations effectively capture high percentage of emotion characteristics embedded in speech as demonstrated in our

simulation results.

5.2. Future Works

To further improve recognition rates for speech samples from the FAU AIBO database, several aspects are to be considered in the future. First, those bad clips should be excluded, and an effective VAD should be considered to better extract voice parts. Secondly, the balancing method is a major concern. Instead of under-sampling the majority as we have done in this study, the SMOTE (Synthetic Minority Over-sampling TEchnique) or other combined methods [23, 24] may provide effective ways for balancing data sets without causing the over-fitting problem from insufficient data. Moreover, kernel functions, which are used by the SVM to map a nonlinear problem to a higher dimensional space, can be changed. Some commonly used kernel functions are listed in Table 2-1. Finally, other feature selection methods besides the SFFS, which is time-consuming and its convergence is not guaranteed, can be adopted for further improvement.

The current RS180 only contains the first and the second order statistics. Usage of the higher-order statistics (as in the Inter384) of the spectro-temporal modulations is expected to boost recognition rates for long speech, but may inevitably cause a more serious over-fitting problem for short speech. We are most interested in the performance by combining over-sampling techniques with higher-order statistics and will pursue it in the future.

Reference

[1] T. Chi, Y. Gao, M. C. Guyton, P. Ru, and S. Shamma, "Spectro-temporal modulation transfer functions and speech intelligibility," *The Journal of the Acoustical Society of America*, vol. 106, p. 2719, 1999.

[2] T. Chi, P. Ru, and S.A. Shamma, "Multi-resolution spectro-temporal analysis of complex sounds," *J. Acoust. Soc. Am.*, vol. 118, no. 2, pp. 887-906, 2005.

[3] B. Schuller, G. Rigoll, and M. Lang, "Hidden Markov Model-Based Speech Emotion Recognition," *Proc. ICASSP*, 2003, vol. 2, pp. 1-4.

[4] Dan-Ning Jiang, and Lian-Hong Cai, "Speech Emotion Classification with the Combination of Statistic Features and Temporal Features", *ICME*, *2004*, pp. 1967-1970.

[5] V. Ververidis and C. Kotropoulos, "Emotional speech recognition: Resources, features, and methods," *Speech Comm.*, vol. 48, no. 9, pp. 1162–1181, September 2006.

[6] Z. Zeng, M. Pantic, G. I. Rosiman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 31, no. 1, pp. 39–58, 2009.

[7] B. Schuller, and G. Rigoll, "Timing Levels in Segment-Based Speech Emotion Recognition," *Proc. INTERSPEECH 2006*, ICSLP, ISCA, pp.1818-1821, Pittsburgh, PA, 2006.

[8] F Ringeval, and M Chetouani, "A vowel based approach for acted emotion recognition," *Proc. Interspeech*, 2008.

[9] B. Schuller, G. Rigoll, and M. Lang, “Speech Emotion Recognition Combining Acoustic Features and linguistic information in a hybrid support vector machine-belief network architecture,” *Proc. ICASSP*, 2004, Vol. I, pp. 577-580.

[10] Feng Yu, Eric Chang, Ying-Qing Xu, and Heung-Yeung Shum, “Emotion detection from speech to enrich multimedia content,” *Proc. IEEE Pacific-Rim Conf. on Multimedia 2001*, Vol. 1, pp. 550–557. 2001.

[11] Tsang-Long Pao, Yu-Te Chen, Jun-Heng Yeh, and Pei-Jia Li, “Mandarin emotional speech recognition based on SVM and NN,” *Proc. of the 18th International Conference on Pattern Recognition (ICPR’06)*, vol. 1, September 2006, p. 1096-0.

[12] B. Schuller, D. Arsić, F. Wallhoff, and G. Rigoll, “Emotion Recognition in the Noise Applying Large Acoustic Feature Sets,” in *Proc. Speech Prosody*, 2006.

[13] B. Schuller, D. Seppi, A. Batliner, A. Maier, and S. Steidl, “Towards More Reality in the Recognition of Emotional Speech,” *Proc. ICASSP*, 2007, Vol. IV, pp. 941-944.

[14] Felix Burkhardt, Astrid Paeschke, Miriam Rolfes, Walter Sendlmeier und Benjamin Weiss, “A Database of German Emotional Speech”, *Proc. Interspeech*, Lissabon, Portugal, 2005, pp. 489-492.

[15] S. Steidl, “Automatic Classification of Emotion-Related User States in Spontaneous Children’s Speech,” Logos Verlag, Berlin, 2009.

[16] Chih-Chung Chang and Chih-Jen Lin, LIBSVM : a library for support vector machines, 2001. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm

[17] A. Varga and H.J.M. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Comm.*, vol.12(3), pp. 247-251, 1993.

[18] B. Schuller, S. Steidl, and A. Batliner, “The INTERSPEESH 2009 Emotion Challenge,” *Proc. Interspeech*, 2009, pp. 312-315.

[19] H. Kawahara, Alain de Cheveign´e, H. Banno, T. Takahashi and T. Irino, “Nearly Defect-free F0 Trajectory Extraction for Expressive Speech Modifications based on STRAIGHT,” *Proc. Interspeech*, 2005, pp. 537-540.

[20] F. Eyben, M. Wollmer, B. Schuller (2009): Speech and Music Interpretation by Large-Space Extraction, http://sourceforge.net/projects/openSMILE.

[21] B. Schuller, M. Wöllmer, F. Eyben, and G. Rigoll, *"Spectral or Voice Quality? Feature Type Relevance for the Discrimination of Emotion Pairs,"* in The Role of Prosody in Affective Speech, Linguistic Insights, Studies in Language and Communication, Vol. 97, Slyvie Hancil (ed.), Peter Lang Publishing Group, ISBN 978-3-03911-696-6, pp. 285-307, 2009.

[22] P. Pudil, F.J. Ferri, J. Novovicova, and J. Kittler, “Floating search methods for feature selection with nonmonotonic criterion functions,” *Proc. international Conference on Computer Vision & Image Processing*, pp. 279-283, 1994.

[23] N. V. Chawla, L. O. Hall, K. W. Bowyer, and W. P. Kegelmeyer, “SMOTE: Synthetic Minority Oversampling Technique,” Journal of Artificial Intelligence Research 16, pp. 321-357, 2002.

[24] G. E. A. P. A. Batista, R. C. Prati, and M. C. Monard, “A study of the behavior of several methods for balancing machine learning training data,” ACM SIGKDD Explorations Newsletter, vol. 6 , issue 1, pp. 20 – 29, 2004.