# 使用時頻變化調變之稀疏表示於<br>強健語音情緒辨認

研究生: 梁詠閎　　　　　　　　　　指導教授: 冀泰石

國立交通大學電信工程研究所碩士班

## 摘　要

語音情緒辨認在這幾年來一直是個熱門的研究題目，目前大多數的研究主要都是著重在分類乾淨語料的情緒類別，在本論文中，我們使用實驗室的感知模型提取出的兩種時頻變化參數 ($ACC$384 以及 $RS$96)，藉由此參數來對參雜了雜訊的語音做情緒上的辨認。我們將柏林情緒語料庫 (Berlin Emotional Speech Database) 以及愛寶情緒語料庫 (Aibo Emotional Corpus) 加入不同訊雜比 (Signal to Noise Ratio) 的白雜訊 (white noise) 以及人聲雜訊 (babble noise), 並比較我們的時頻變化參數與其他知名的參數 ($inter$384) 在不同訊雜比下的優劣。在實驗中，我們使用了兩種不同的分類法，分別為稀疏表示分類法 (Sparse Representation Classification) 與向量支持機 (Support Vector Machine) 進行分析，而實驗的結果顯示，我們實驗室的時頻變化參數再受到雜訊的干擾時仍然有較好的辨認率，也發現使用稀疏分類表示法的實驗解果較優於向量支持機。在本論文中，我們也對稀疏分類表示法做了一些討論。

# Robust Speech Emotion recognition via sparse representation of Spectro-Temporal Modulations

Student : Young-Hong Liang  Advisor : Tai-shih Chi

Institute of Communications Engineering
National Chiao Tung University
Perception Signal Processing Laboratory

## Abstract

Speech emotion recognition is a popular research topic in last decade. However, most researches are always focus on clean speech, in this thesis, we use two kinds of feature sets which are extracted from our auditory model and applied to recognize the emotion categories of both clean and noisy speech. And the noisy utterance is derived from the Berlin Emotional Speech Database and the Aibo emotional Corpus with additive babble noise and additive white noise under different signal to noise ratio (SNR) value. Comparing with the famous feature set which is proposed in the INTERSPEECH 2009 Emotion Challenge and use two kinds of classifiers, which are sparse representation classification (SRC) and support vector machine (SVM). The robustness of our spectro-temporal modulation feature sets are better than the feature set proposed in INTERSPEECH 2009 Emotion Challenge, and the performance of SRC is better than SVM. Some discuss on SRC are given in this thesis.

# 致　謝

光陰似箭，歲月如梭，在交通大學的兩年碩士生活轉眼間就到了尾聲，在這兩年內，我最感謝的人莫過於我的指導教授冀泰石博士。無論是在每周實驗室的新知分享上的叮嚀與指教或是在私底下討論論文題目方向上的建議，都給了我很大的幫助。教授也不時會跟我們提點做人處事應有的態度與自我管理的重要，這些更是在往後的人生道路上受用無窮的。這段時間，真是辛苦您了，謝謝教授。

感謝實驗室的學長，學弟妹們，無論是一起修課時的革命情感或是在課餘時間的嘻笑打罵，都為我的碩士生活增添不少樂趣也紓解了不少壓力，謝謝博班學長大樹，勝哥在一起修最佳化時的切磋，也謝謝阿郎學長在碩一的計畫上的指教；謝謝家銘學長在新知分享上給的建議，也謝謝同是碩二的坤燁、志偉、育群，無論是在研究遇到瓶頸時的討論與建議或是在我想要偷懶時所給予的叮嚀，我都相當的感謝。碩一的學弟妹們所辦的小活動都讓我可以好好的放鬆休息，謝謝你們。在這裡也預祝你們在未來的一年內研究順利。

謝謝我的朋友與室友，雖然有些好朋友現在並不再新竹，也不能時常見面，但是每次見面時所給予的鼓勵，或是點點滴滴的關心都是我繼續努力的動力。謝謝室友在課業上的幫忙與當情緒失控時的容忍，以及在課餘時間一起玩玩遊戲抒發壓力，嚼嚼舌根說說八卦，吃吃美食享受生活。如果沒有你們個鼓勵與包容，這段路走來必定是更加辛苦，謝謝你們。

最後，感謝我的家人，在我返家時總會準備很多好吃的食物，也會不時關心我在新竹生活的情況，還有皮膚的過敏情形，有了你們的關心，我才能夠完正碩士的學業，在此與你們分享這成果。

梁詠閔謹誌　于國立交通大學　新竹
中華民國 102 年 7 月

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

## 1.1 Motivation

As an important form that carries human affective information, speech emotion is too mysterious to be recognized perfectly. Knowing the emotion of the speaker is helpful to build a good human-machine interface, such as the one with an interactive robots, or the custom service line based on speech-recognition, and so on. However, the speech emotion recognition often considers the ”perfect conditions”, for example, the databases we used are recorded in studio-quality and acted by actors. These conditions are quite far from real-world applications. In this work, we will propose a robust feature set extracted from our spectro-temporal auditory model [2] [1] and use the sparse representation classification (SRC) for speech emotion recognition. We believe that our auditory features could be more robust in noisy condition.

## 1.2 Related Works

Speech emotion recognition draws a growing attention in the last decade. The application of speech emotion recognition can be broadly used not only in psychological research, but also in real-life practices like in a call center, in the human-machine interface design or in medical care [3][4]. The neutral speech model were also used in emotional speech analysis and feature normalization [5] [6].

In the last few years, more and more researchers pay attentions to real emotions, i.e., the natural emotions of speech corrupted with environmental noise [7][8]. Trying to cope with the noise in speech emotion recognition, those studies searched a best set of features from thousands of features with the highest recognition rate for each different testing environment. In previous works, some researchers were tried to find the contributions of prosodic features and their dynamics [9] [10]. And some tried to find the best unit that conveys recognizable emotional information [11], and those finding have been successfully used in a study which segmented a speech signal into many sub-sentences with 2-stage classifiers [12].

The databases are important for speech emotion recognition verification. Databases often contain acted speech from television dramas or produced by professional actors. One problem is that the databases are recorded by different groups, and the recording standards of each group probably are different. The new method which considers the reliability-weighted labels was proposed to make the labels more reliable [13]. Another issue is that, the ”acted emotions” are far away from our real daily ones. To address the first problem, we choose the Berlin emotional speech corpus [14], which is widely used by other researchers. For the second problem, we use the Aibo emotion corpus [15], which consists of spontaneous emotional speech and is adopted as the test database in the INTERSPEECH 2009 emotion challenge [16]. The baseline results of the Aibo emotion corpus are the unweighted averaged recall (UA) rate of 38.2% and weighted average recall (WA) rate of 39.2%, respectively. The WA is also refer to as the ’accuracy’.

## 1.3 Experimental Framework

In this thesis, we test the robustness of proposed spectro-temporal auditory feature using the Berlin emotional speech corpus and the Aibo emotion corpus under two kind of additive noise. The performance of our features will be compared with the 384 features proposed in the INTERSPEECH 2009 emotion challenge in the following experiments. The sparse representation classification (SRC) method is used as our classifier [17]. The flowchart of our system is shown in Fig.1.1.

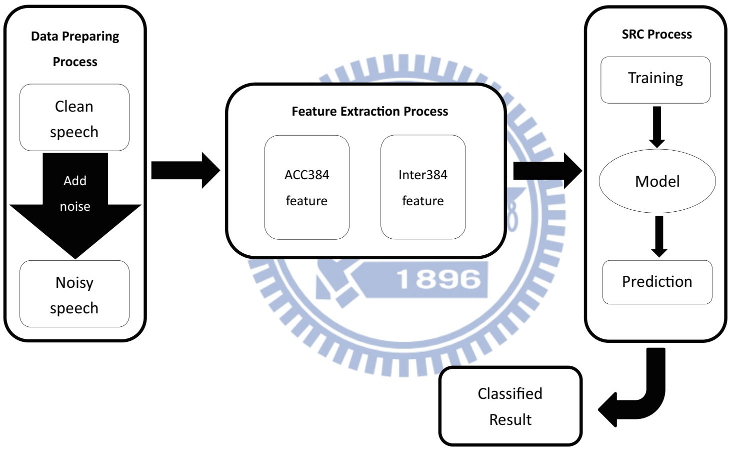


Figure 1.1: Overall flowchart of proposed method

## 1.4 Thesis Organization

The organization of this thesis is described as follow: In chapter 2, the brief introductions of our spectro-temporal auditory model and the sparse representation classification are given. Chapter 3 describes two databases and the feature sets used in our experiment. The experiment setups and recognition results are shown in chapter 4. And we end in chapter 5 with conclusions and discussions.

# Chapter 2

# Literature Review

## 2.1 Auditory Model

The features adopted in this study are extracted from a spectro-temporal auditory model, which is based on physiological evidences and consists of a central cortical (brain) module and an early cochlear (ear) module. We will give a brief introduction of hearing physiology at first, and then discuss the cochlear model and cortical modules in section 2.1.2 and 2.1.3. Detailed descriptions and neuro-physiological evidences of this auditory model can be accessed in [1].

### 2.1.1 Hearing Physiology

The cross-sectional view of the human ear is shown in Fig.2.1. It can be divided into three parts: the outer ear, the middle ear and the inner ear. Sounds waves enter the outer ear, travel through the ear canal ,and then vibrate the ear drum. In the middle ear, the vibrations of the ear drum are transmitted into the inner ear through the three ossicles, which are malleus, incus, and stapes. The stapes touch a liquid filled sack and the vibrations transmit into the cochlea, whose the shape looks like a shell. The cochlea connects with hundreds of nerve fibers, which are transmitted

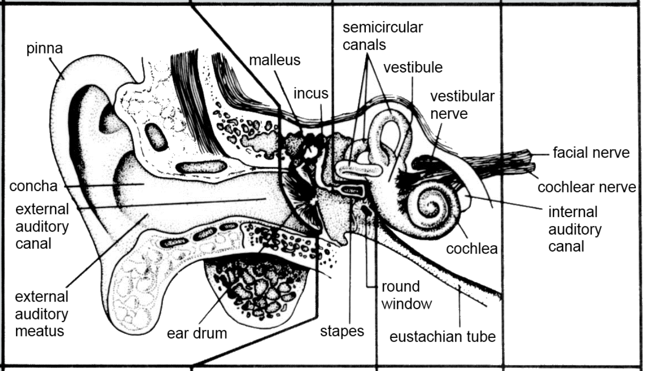
| Gross division | Outer ear | Middle ear | Inner ear | Central auditory nervous system |
|---|---|---|---|---|
| Anatomy |  | | | |
| Mode of operation | Air vibration | Mechanical vibration | Mechanical,<br>Hydrodynamic,<br>Electrochemical | Electrochemical |
| Function | Protection,<br>Amplification,<br>Localization | Impedance matching,<br>Selective oval window stimulation,<br>Pressure equalization | Filtering distribution,<br>Transduction | Information processing |

Figure 2.1: Cross-section view of the functions of human ear (cite from Swiss National Sound Archives Foundation).

information along the auditory pathway to the brain. Finally, the brain processes information from the ear.

The main function of outer ear are protection, amplification, and localization. The shape of outer ear help people to collect sound waves and find out where the waves come from. The three ossicles transduce the acoustical vibration into mechanical vibration. The energy loss due to entering the liquid from the air is compensated at this part. The cochlea in the inner ear plays an important role in the auditory system. The structure of the cochlea is shown in fig.2.1.2. The panel shows the stretched top view and side view of the cochlea with the basilar membrane (BM), which is about 35 mm in length and its width increases and stiffness decreases non-uniformly from base to apex. While the mechanical vibration gets to the oval window,

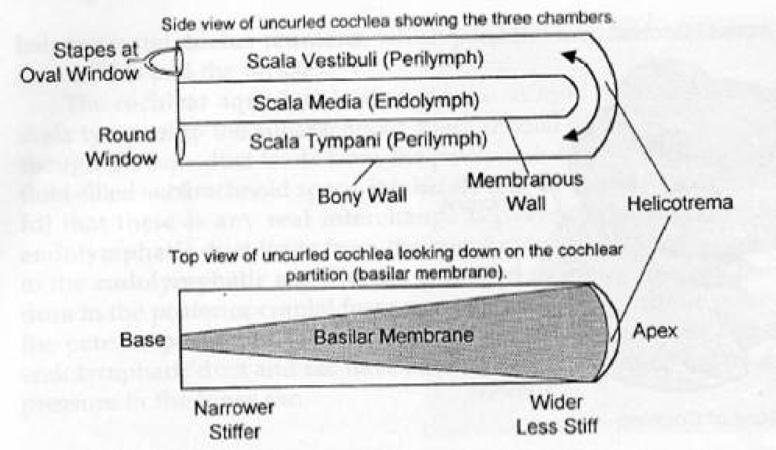

Side view of uncurled cochlea showing the three chambers.
Stapes at Oval Window
Scala Vestibuli (Perilymph)
Scala Media (Endolymph)
Round Window
Scala Tympani (Perilymph)
Bony Wall
Membranous Wall
Helicotrema
Top view of uncurled cochlea looking down on the cochlear partition (basilar membrane).
Base
Basilar Membrane
Apex
Narrower Stiffer
Wider Less Stiff

Figure 2.2: Structure of the cochlea (Hearing Physiology Handout, AAIP).

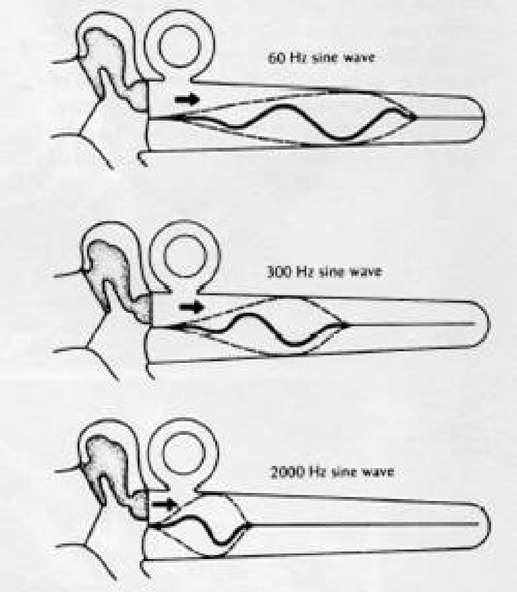


Figure 2.3: The responses for different frequencies (Hearing Physiology Handout, AAIP).

a traveling wave is generated and propagates along the basilar membrane of the cochlea. The traveling waves generated by different frequencies will reach maximum responses and different locations of the BM. And fig.2.3 shows the fact about maximum responsive frequencies along the BM; the lower the frequency is, the farther the traveling wave reaches. A linear relationship was observed between the traveling distance from the base of the cochlea and the log-frequency of input sounds. Due to the mechanical property of the traveling wave, the maximum response on a specific frequency would suppress its surrounding frequencies on the BM, this phenomenon might explain the ”frequency masking” of human audition.

There are lots of hair cells distributed along the BM. When a traveling wave transmit along the BM, the hair cells will be activated and electrical signals will be sent to the midbrain via the auditory nerves. There are two kind of hair cells: the inner hair cell, and the outer hair cell. The total number of inner hair cells is about 3000, and the number of outer hair cells is about 4 times of the inner hair cells. The transforming of mechanical vibration into electrical signals is completed by the inner hair cells, and the outer hair cells often help in further amplification or reduction in pertaining to extreme sounds. Because of some relaxation time is needed between consecutive fires of auditory nerves, firing rates can’t keep up with high frequency vibrations, as show in Fig.2.4. Firing rates of inner hair cells are limited by 4-5 kHz, and the firing rates of the neurons of the midbrain are limited by about 1 kHz.

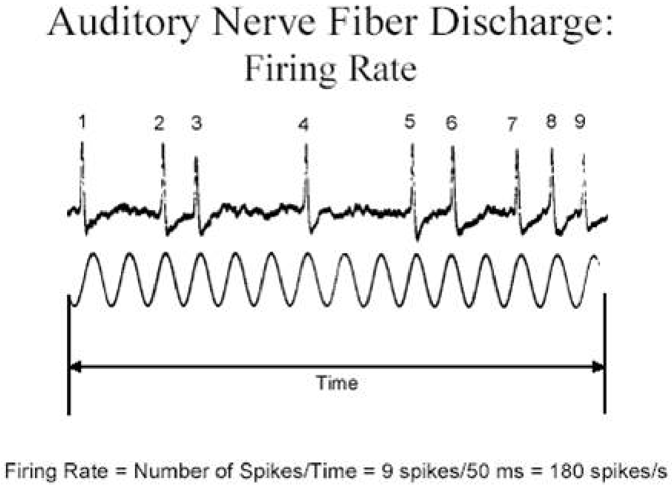


Figure 2.4: The firing rate of an auditory nerve corresponding to a single tone input (Hearing Physiology Handout, AAIP).

### 2.1.2 Cochlear Module

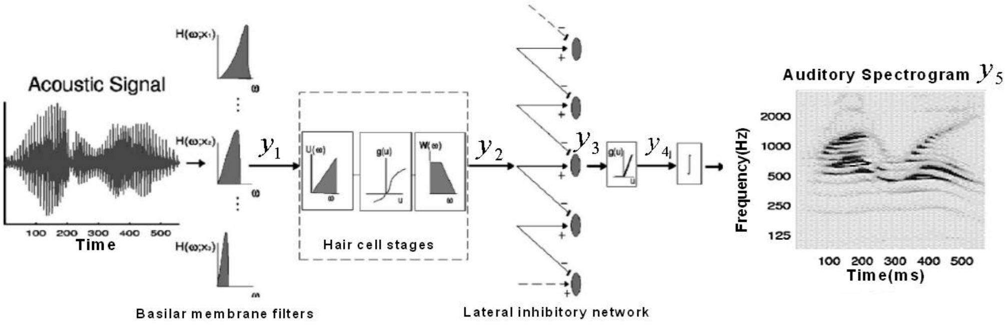


Figure 2.5: Cochlear module of the auditory model [1].

The cochlear module models the functions of the peripheral auditory system. The cochlea works like a frequency analyzer. As Fig.2.5 shows, the cochlear module first consists of a bank of 128 overlapping asymmetric constant-Q band-pass filters ($Q_{3dB} \approx 4$) which mimic the selection of the frequency of the cochlea. The filters are evenly distributed over 5.3 octaves with a 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 used for simulating the saturation of inner hair cells, which transduce the vibrations of the basilar membrane along the cochlea into intracellular hair cell potentials. The following first-order LIN inhibits near-by hair cell potentials along the log-frequency axis. It approximately models the frequency masking of hearing, Then the auditory nerve transmits the hair cell potentials to the cochlear nucleus of the central auditory system, whose dynamic range is largely reduced comparing with the dynamic range of hair cells. In this study, the reduce of dynamic range was modelled by a half-wave rectifier followed with a low-pass filter. This study also considers the case of disabled hair cell stage, i.e., $y_2(t,f) = y_1(t,f)$. To avoid the non-linear high-gain compression of the hair cells, all speech signals are normalized in advance. As in Fig. 2.5, the outputs at different stages of the cochlear module are showed below:

$$y_1(t, f) = s(t) *_t h(t; f) \tag{2.1}$$

$$y_3(t, f) = \partial_f y_1(t, f) \tag{2.2}$$

$$y_4(t, f) = max(y_3(t, f), 0) \tag{2.3}$$

$$y_5(t, f) = y_4(t, f) *_t u(t; \tau) \tag{2.4}$$

where $s(t)$ is the input speech, $h(t, f)$ is the impulse response of the constant-Q cochlear filter with center frequency f, $*_t$ described the convolution in time, $\partial_f$ is the partial derivative along the log-frequency (f) axis and the integration window $u(t; \tau) = e^{-t/\tau} \times u(t)$.

The $y_5(t, f)$ is referred to as an auditory spectrogram, which represents neuron activities of the peripheral auditory system along the time and log-frequency axis. From a functional view, the auditory spectrogram produced by this simplified linear cochlear module is close to the magnitude response of a mel-scaled FFT based spectrogram. The mel-scale warping shares a resembling constant-Q criterion of the filter bank but with different bandwidth and frequency resolution. Note that the LIN accounts for the frequency masking effect provided that the non-linear behaver of hair cells. However, since this study doesn't consider the hair cell stage, the LIN only sharpens the constant-Q cochlear filters effectively.

### 2.1.3 Cortical Module and Rate-Scale Representation

The second module models the spectro-temporal selectivity of neurons of the auditory cortex (A1). In the auditory model, the auditory spectrogram $y_5(t, f)$ is further filtered by cortical neurons which are demonstrated by two-dimensional filters tuned to different spectro-temporal modulation parameters [1]. The *rate* (or velocity) parameter $\omega$ (in Hz) shows how fast the local spectro-temporal envelop changes along the temporal axis. The *scale* (or density) parameter $\Omega$ (in cycle/octave) characterize how dense the the local spectro-tempo envelope distributes along the log-frequency axis. In addition to the rate and the scale, cortical neurons are also found to be sensitive to the sweeping direction of modulation of the sound. This directionality is characterized in this module by sign the rate: negative for upward sweeping direction; positive for backward sweeping direction.

In our auditory model, the 4-dimensional output of this cortical module can be wrote as:

$$r(t, f, \omega, \Omega) = y_5(t, f) *_{tf} STIR(t, f; \omega, \Omega) \tag{2.5}$$

where $STIR(t, f; \omega, \Omega)$ is the joint two-dimensional spectro-temporal impulse response (STIR) of the direction-selective filter tuned to $\omega$ and $\Omega$, and $*_{tf}$ depicts the two-dimensional convolution in the time and log-frequency domain. More detail information is available in [1]. The local energy of the 4-dimensional output is described below:

$$E(t, f, \omega, \Omega) = |r(t, f, \omega, \Omega) + jH(t, f, \omega, \Omega)| \tag{2.6}$$

where $H[\cdot]$ means the Hilbert transform along the log-frequency axis. From a functional point of view, due to various rate-scale combinations, the performance of cortical neurons is a joint spectro-temporal multi-resolution analysis on the input auditory spectrogram. The excitation pattern of cortical neurons to a single time-frequency (T-F) unit at $(t_i, f_j)$ in the spectrogram is referred to as the rate-scale (RS) representation of that particular T-F unit which is represent as $E(\omega, \Omega; t_i, f_j)$. We can get the frame-based RS representation of a speech by taking the mean of the T-F units of a frame over the log-frequency axis showed as follow:

$$P(\omega, \Omega; t_i) = \frac{1}{128} \sum_{j=1}^{128} E(\omega, \Omega; t_i, f_j) \tag{2.7}$$

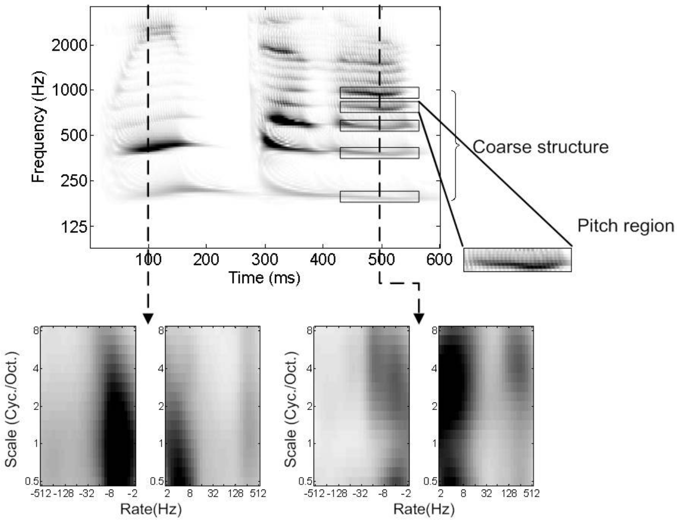


Figure 2.6: rate-scale representations of speech frames.

The panel on the top of Fig.2.6 shows a sample of an auditory spectrogram. The bottom panels show the time-varying RS representation $P(\omega, \Omega; t_i)$ of the sample speech around 100 and 500 ms. Each plot of the RS representation clearly shows two attributes respectively: one is the spectro-temporal modulations of envelopes and the other is the resolved pitch below 512 Hz. Take the 500 ms frame as an example. The resolved pitch around 190 Hz produces a strong response around the high-rate (pitch related) region. On the other hand, the envelopes of the almost flat harmonic structure shown at 190, 380, 570, 760 and 950 Hz produces low-rate (due to the flatness), high-scale (3 cycles within the 500 1000 Hz octave) strong responses at regions less than 8 Hz and around 3 cycle/octave. Since the harmonic structure shows the slightly downward sweeping direction, the low-rate region exhibits an asymmetric response between two directions (represented by the positive rate and the negative rate). From this example, we can observe that the frame-based $P(\omega, \Omega; t_i)$ encodes the information of the spectral-temporal structures, including but not limited to pitch, harmonicity, formant spacing, and modulations of an input sound at each time instant. Some of these structures, such as pitch, amplitude modulation (AM) and frequency modulation (FM), are associated with the prosody of the sound, while others are associated with the spectral characteristics of the sound. Variations of these two types of features (prosodic and spectral features) are commonly used in speech emotion recognition researches [9][10]. Therefore, the time-varying RS representation extracted from the auditory model could be a good candidate for speech emotion recognition.

In Fig.2.7 shows the long-term averaged $p(\omega, \Omega)$ of clean speech from male speakers, female speakers and white noise. The long-term averaged RS representation of clean speech was created by averaging $p(\omega, \Omega; t_i)$ from 30 clean speeches in the NOIZEUS corpus [18]. Obviously, the white noise mainly affects the pitch region ($>$ 128 Hz) of speech. Apart from the pitch region, speech occupies high energies in the low-rate low-scale region ($<$ 16 Hz, $<$ 4 cycle/octave), while white noise activates the high-rate high-scale region ($>$ 16 Hz, $>$ 2 cycle/octave) because of the differences in the structures of their spectrao-temporal envelopes. This figure shows that the local spectro-temporal envelopes of speech are mostly smoother than the envelopes of white noise along both at the time and the frequency axes. These spectro-temporal envelopes represent the AM and the FM of the sound, which are believed important clues for humans to segregate individual sound streams from a sound mixture [19][20]. This segregation process of human hearing perception is quite important to our daily lives, and is referred to as auditory scene analysis (ASA) [21].

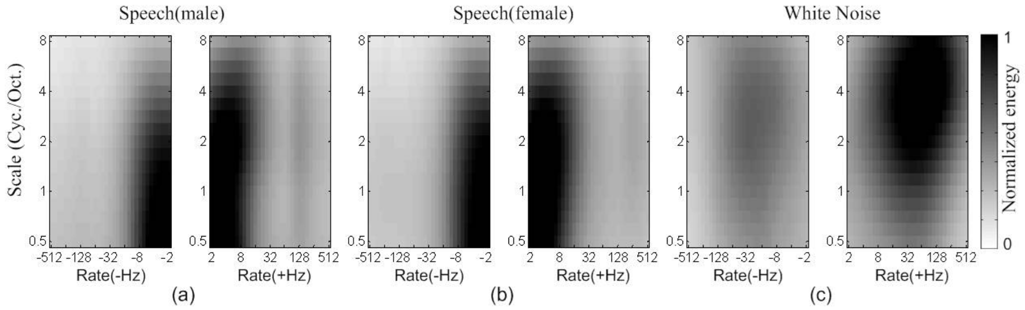


Figure 2.7: Long-term averaged rate-scale representations of speech from male speakers (a) and female speakers (b) and from white noise (c).

## 2.2 Sparse Representation-Based Classification

To determine the class label of a test spectro-temporal modulation representation, given a number of labelled training spectro-temporal modulation representation from $N$ emotional categories, a SRC based classifier is used.

Let $A_i = [a_{i,1}|a_{i,2}|...|a_{i,n_i}] \in \mathbf{R}_+^{m\times n_i}$, represents the dictionary which contain $n_i$ spectro-temporal modulation representation stemming from the ith emotional category as column vectors (i.e. atoms). Given a test spectro-temporal modulation representation $y \in \mathbf{R}_+^{m\times n_i}$ which belongs to the ith class, we assume that $y$ can be expressed as a linear combination of the atoms that belong to the ith class, that is to say,

$$y = \sum_{j=1}^{n_i} a_{i,j}x_{i,j} = A_i x_i \tag{2.8}$$

where $x_{i,j} \in \mathbf{R}$ are coefficient vector $x_i = [x_{i,1}, x_{i,2}, ..., x_{i,n_i}]^T$.

Now, let us define the matrix $A = [A_1|A_2|...|A_N] \in R_+^{m\times n}$ by concatenating the $N$ dictionaries, which stem from N emotional categories. Thus the linear representation of the test spectro-temporal modulation representation $y$ in (2.8) can be rewritten as follows

$$y = Ax \tag{2.9}$$

where $x = [0^T|...|0^T|x_i^T|0^T|...|0^T]^T$ is the coefficient vector whose elements are zeros except those that are associated with the ith emotional category. Thus, the elements of $x$ encode information about the emotional category of the test spectro-temporal modulation representation $y$.

We rewrite our problem below:

$$minimize||Ax - y||_2^2 + \lambda||x||_2^2 \tag{2.10}$$

where $\lambda > 0$ is the regularization parameter. However, the solution of (2.10) is generally dense, with large nonzero entries corresponding to training samples from any different class. In order to find the desired solution $x$, we change our problem from (2.10) to

$$minimize||Ax - y||_2^2 + \lambda||x||_1^2 \tag{2.11}$$

which has been claimed the solution will be sparse [22], that is $x$ has relatively few nonzero coefficients. And we will use the Truncated Newton Interior Point Method [23] to solve (2.11). When we computed the sparse representation $x$, ideally, the nonzero entry in the estimated $x$ will be associated with the column of $A$ from a signal object class $i$, and we can easily deduce the class of the test sample $y$. However, noise and the modeling error may lead to small nonzero entries in other multiple object classes, so we will use the nearest subspace method described below to determine the class of the solution $x$.

For each class $i$, let $\delta_i : \mathbf{R}^n \to \mathbf{R}^n$ be the characteristic function which has the ability to select the class that $y$ is exactly associated with. For $x \in \mathbf{R}^n, \delta_i(x) \in \mathbf{R}^n$ is a new vector whose nonzero elements are the elements of x associated with class i. We can approximate the given test sample $y$ as $\hat{y} = A\delta_i(x)$. Then we can classify $y$ based on the approximation by assigning it to the class i that minimizes the residual between $y$ and $\hat{y}$, as follow:

$$\underset{i}{min}\ r_i(y) = ||y - A\delta_i(x)||_2, \tag{2.12}$$

more detailed information is available in [22].

# Chapter 3

# Database and Feature Extraction

## 3.1 Database

### 3.1.1 Berlin Emotional Speech Database

There are 535 German utterances in this database and each utterance has duration of 3 to 6 seconds. The database contains 7 emotions which are anger, happiness, sadness, fear, disgust, boredom, and neutral. Detailed contents are listed in table 3.1. Those emotions were acted by 5 male and 5 female actors. Only those utterances with higher than 80% emotion recognition rate by human labellers are included in this database. The original speech samples were recorded with the 16 kHz sampling frequency in a studio-quality condition. To cover the fundamental frequencies of regular male speakers when using the 5.3 octave frequency coverage cochlear filterbank in our auditory model described in chapter 2, all speech utterance are downsampled to 8kHz before entering the feature extraction stage.

Table 3.1: The German content of Berlin Emotional Speech Database and its English translation

| code | German text | English translation |
|---|---|---|
| a01 | Der Lappen liegt auf dem Eisschrank. | The tablecloth is lting 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. |

### 3.1.2 Aibo Emotion Corpus

There are 18216 speech chunks, which last about 8.9 hours in total in this database. It contains 5 nature emotions, anger, emphatic, neutral, positive and rest, of 51 children (21 boys and 30 girls) aged between 10 and

13. Those children were invited to play with AIBO, a pet robot dog, and asked to guide it through some missions, such as moving along a particular route. The speech samples were recorded with medium quality. Some samples contain serious microphone clicking noise or coughing sounds, and some are clipped due to their loudness. As mentioned before, all speech samples are downsampled to 8 kHz from the original sampling frequency (16 kHz). This corpus was used in INTERSPEECH 2009 Emotion Challenge [16], and the baseline results are the 38.2% unweighted average recall (UA) rate and the 39.2% weighted average recall (WA) rate respectively. More details about the number of speech samples in each emotion class is listed in Table 3.2, where A, E, N, P and R represent are Anger, Emphatic, Neutral, Positive and Rest, respectively.

Table 3.2: Number of instances in Aibo emotion corpus for the 5 class recognition 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 |

Different from the Berlin Emotional Speech Database, which consists of acted emotions, the Aibo Emotion Corpus contains spontaneous emotions, which are much more difficult recognize. Besides the nature of the spontaneous emotion , other reasons also contribute the difficulty of the AIBO Corpus in emotion recognition tasks. First, these samples are also difficult for human labellers to recognize, such that the resulting labels are with low consistency. The low prototypical samples can be considered as a kind of internal noise which would damage the classification results either for the labellers or the machine. Second, the ”rest” emotion is undefined, such that speech sample in this class demonstrate highly variable characteristics. Third, the Aibo Corpus is highly unbalanced. The ratio between the largest class ”neutral” and the smallest class ”positive” is 12.33. To cope with this problem, we use the method in [24] which randomly down-sample other classes to meet the size of the smallest class.

## 3.2 Feature Extraction

### 3.2.1 spectro-temporal modulation representation

As mention in 2.1.3, the full output of our spectro-temporal modulation analysis model is a 4-dimensional (time, frequency, rate, scale) representation. Originally in our model, we set the range of rate from $2^1$ to $2^8$ and $-2^1$ to $-2^8$ (to cover the complete temporal information from the slow amplitude modulation to the fine pitch structure of the speaker), and the range of scale from $2^{-2}$ to $2^3$ (to cover the complete frequency structures such as formants and harmonicity). To get 384 rate-scale-frequency features, the 4-dimensional full reprensentation is first downsized to 3-dimension by taking an average along the time axis. Second, due to the fact that the rate filter are highly overlapped, only $2^2, 2^4, 2^6, 2^8$ and $-2^2, -2^4, -2^6, -2^8$ rate filters are selected to reduce the dimensionality. Third, the 128 constant-Q filter cochlear filters are evenly divided into 8 groups, each of which has 16 cochlear filter. Total energy in each group roughly catches the energy profile of the speech signal along the log frequency axis. Hence, the total number of features we used is 2 (directions) $\times$ 4 (number of rates) $\times$ 6 (number of scales) $\times$ 8 (output energies along the log-frequency axis)$=$ 384. These 384 features is referred to as the **ACC384** feature set. We also consider the rate-scale features by averaging over the time and the frequency axes, and the total number of RS features is 2 (directions) $\times$ 8 (number of rates) $\times$ 6 (number of scales) $=$ 96. This feature set is referred to as the **RS96**.

### 3.2.2 INTERSPEECH 2009 Emotion Challenge Features

The feature set adopted in INTERSPEECH 2009 emotion challenge [16] is used for performance comparison. It can be extracted by the accessible open source openSMILE feature extraction toolkit [25]. This feature set consists of 16 low-level descriptors (LLDs) and their 1st-order derivatives. The 16 LLDs include zero-crossing-rate (ZRC) of the time signal, root mean square (RMS) frame energy, pitch frequency (normalized to 500 Hz), harmonics-to-noise (HNR) ratio by autocorrelation function, and 1-12 mel-cepstral coefficients (MFCC) derived using HTK as shown in table 3.3. In additional, 12 functionals, including mean, standard deviation, kurtosis, skewness, the value and position of minimum and maximum, the range between minimum and maximum, and two linear regression coefficients with their mean square error (MSE) are considered. Thus, this feature set contains $16 \times 2 \times 12 = 384$ features, which is referred to as the **inter384** feature set.

Table 3.3: Features used in INTERSPEECH 2009 Emotion Challenge

| LLD (16*2) | Functionals (12) |
|---|---|
| (Δ)RMS energy | max, min, range, max position, min position, temporal mean, standard deviation, skewness, kurtosis<br>linear regression: offset, slope, MSE |
| (Δ)MFCC 1-12 | |
| (Δ)ZCR | |
| (Δ)HNR | |
| (Δ)F0 | |

# Chapter 4

# Experimental Results and Discussions

## 4.1 Experimental Setup

In our system, a speech utterance is processed through following stages in order: pre-processing, feature extraction, and classification stage. In the per-processing stage, as mentioned before, the speech is down-sampled to 8k Hz to meet the specifications of our cochlear module. Then a voice activity detector (VAD) is applied to mark the voice regions. After those steps, all utterances are normalized to have zero mean and unit variance. In the next stage, the auditory model extracts the modulation features. In this experiment, we not only use the RS features which are derived from (2.7), but also use the outputs of the constant-Q cochlear filters, i.e., **ACC384**. And the **inter384** which was proposed in [16] and is extracted by openSMILE feature extraction toolkit [25] is used as the feature set for performance comparison. In the Final stage, the SRC [17] which was introduced in section 2.2, and the well known SVM [26] are our classifiers.

For Aibo, the training set contains 9959 clean speech chunks which were uttered by children in Ohm school and the test set contains 8527 clean speech chunks which were uttered by children in Mont school (Notice that, as mentioned before, to deal with the data imbalance issue, we randomly down-size the other classes to have the same size as the smallest class). For Emo-DB, 10-fold cross validation is applied due to the limitation of sample numbers. All samples are randomly divided into 10 set, one of them is used for testing, and the others are used for training. Final simulation results are obtained by averaging over 10 trails. In order to create different signal-to-noise (SNR) ratio environments, the white Gaussian noise (AWGN) and the babble noise (ABN) in NOISEX-92 database [27] are added in our following experiment. The SNR conditions we test are 20dB (only for Emo-DB), 15dB, 10dB, 5dB, and 0dB.

## 4.2 Results on Berlin Database

Table 4.1: Simulation results on Berlin Database under babble noise using SRC

| | clean | 20dB | 15dB | 10dB | 5dB | 0dB |
|---|---|---|---|---|---|---|
| $ACC384$ | 76.54% | 76.43% | 74.66% | 72.33% | 69.09% | 63.45% |
| $inter384$ | 65.31% | 59.65% | 57.51% | 55.95% | 52.48% | 44.46% |
| $RS96$ | 68.45% | 68.27% | 66.22% | 63.87% | 62.87% | 51.27% |

Table 4.2: Simulation results on Berlin Database under white noise using SRC

| | clean | 20dB | 15dB | 10dB | 5dB | 0dB |
|---|---|---|---|---|---|---|
| $ACC384$ | 76.54% | 76.43% | 74.37% | 74.30% | 68.95% | 65.26% |
| $inter384$ | 65.31% | 57.08% | 54.91% | 50.37% | 48.47% | 46.37% |
| $RS96$ | 68.45% | 68.58% | 66.60% | 64.34% | 62.13% | 54.36% |

Table 4.3: Simulation results on Berlin Database under babble noise using SVM

| | clean | 20dB | 15dB | 10dB | 5dB | 0dB |
|---|---|---|---|---|---|---|
| $ACC$384 | 75.28% | 75.69% | 74.07% | 72.13% | 69.00% | 59.78% |
| $inter$384 | 78.6% | 69.42% | 67.83% | 63.60% | 62.76% | 54.61% |
| $RS$96 | 71.04% | 70.16% | 69.27% | 69.66% | 62.15% | 49.70% |

Table 4.4: Simulation results on Berlin Database under white noise using SVM

| | clean | 20dB | 15dB | 10dB | 5dB | 0dB |
|---|---|---|---|---|---|---|
| $ACC$384 | 75.28% | 74.22% | 72.74% | 71.71% | 68.72% | 60.11% |
| $inter$384 | 78.6% | 72.25% | 71.74% | 67.05% | 64.90% | 58.78% |
| $RS$96 | 71.04% | 68.65% | 65.24% | 65.42% | 61.54% | 55.95% |

Table 4.1 to 4.4 show the un-weighted average recall rate (UR) on Berlin Database in different SNRs with different types of noise. $SRC$ is the sparse representation classifier discussed in 2.2, $SVM$ is the support vector machine, $b$ is additive babble noise, and $w$ is additive white noise. Results of different classifiers are shown in Fig.4.1 and Fig.4.2, respectively.

Clearly, the $ACC$384 outperforms other two feature sets using $SRC$ in different SNRs, and the robustness of our RS feature set has been demonstrated in our previous works [24]. Fig.4.3 and Fig.4.4 shows the outputs of a pitch-related RS region (rate=256 Hz, scale=4 cycle/octave) and an AM-related RS region (rate=4 Hz, scale=0.5 cycle/octave) along the time axis. Although the pitch-related response is affected by the white noise, the trend along the time axis is not seriously damaged, and that's why our RS feature set is robustness under the white noise condition.

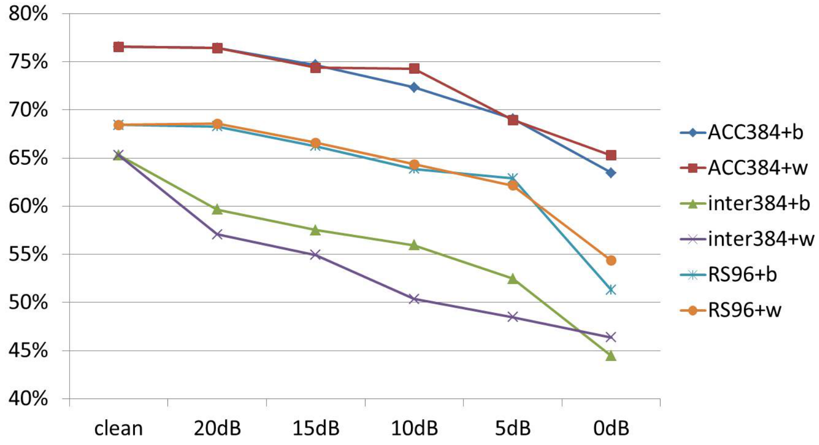


Figure 4.1: Simulation result on Berlin Database using SRC.

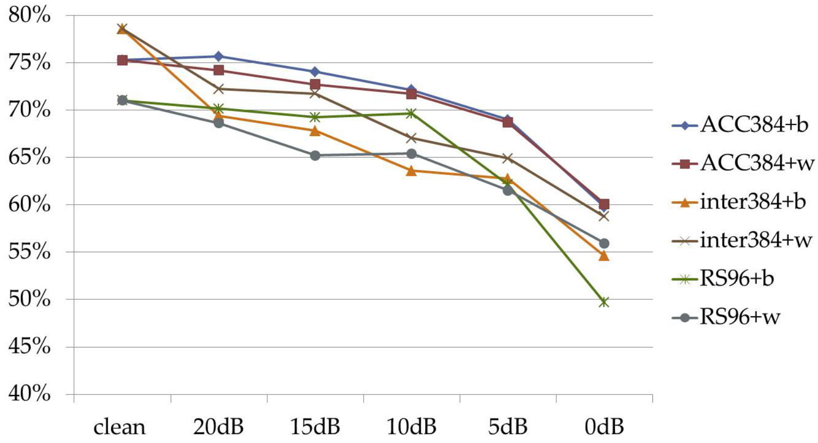


Figure 4.2: Simulation result on Berlin Database using SVM.

The difference between our $ACC384$ and $RS96$ is the inclusion of modulation phase information. The $RS96$ only carries the information of spectro-temporal amplitude modulations, which is equivalent to the spectro-temporal envelopes without the phase information. That's the reason why our RS feature set can not provide an accurate prediction in clean condition. The $ACC384$ feature set not only carries the information of spectro-temporal amplitude modulations, but also keeps the phase information by including the original waveform out of the constant-Q filters. Combining those informa-

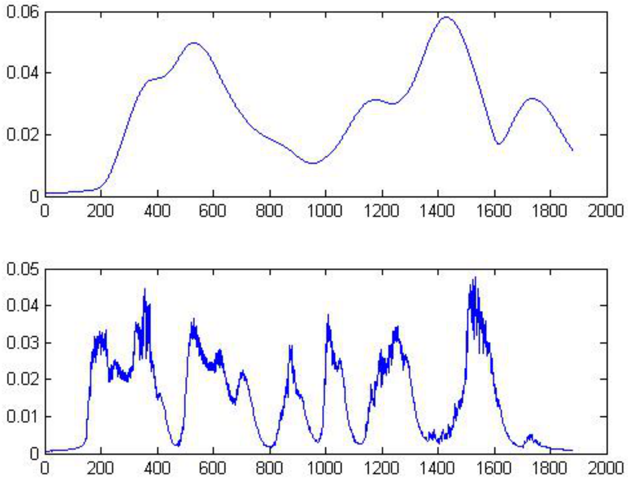


Figure 4.3: One clean Berlin Anger sentence: the bottom panel shows a high rate/high scale (pitch-related) response and the top panel shows a low rate/ low scale (AM-related) response plotted along the time axis.

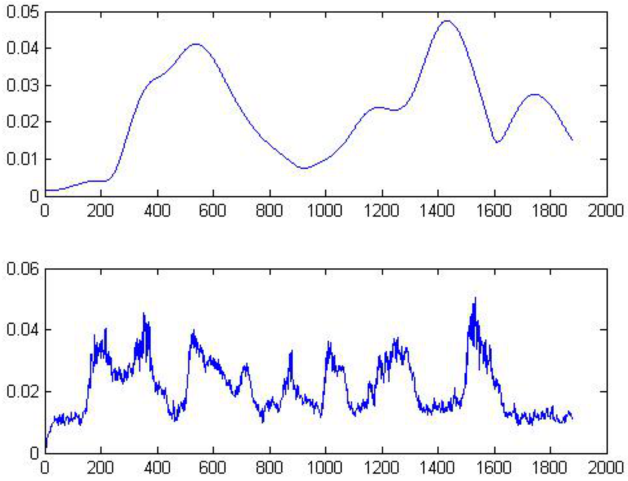


Figure 4.4: One 5dB Berlin Anger sentence under white noise: the bottom panel shows a high rate/high scale (pitch-related) response and the top panel shows a low rate/low scale (AM-related) response plotted along the time axis.

tion, we can have a better match under clean condition. Fig.4.5 shows the RS96 for a ”fear” sentence and an ”anger” sentence. As the figure shows, the rate-scale pattern is very similar and hard to distinguish. However, if we consider the output information obtained from each different groups of constant-Q filters, the difficulty might be reduce. Fig.4.6 to Fig.4.13 show the RS96 in each group. The difference between each RS plot is now more observable, and that is the reason why our *ACC*384 feature set outperforms the *RS*96 feature set. We also show the confusion matrices of clean speech with 0dB babble noise and with 0dB white noise in table.4.5 to table.4.10.

Table 4.5: Confusion matrix of clean speech for ACC384 feature on Berlin Database

| | F | D | J | N | S | A | B | |
|---|---|---|---|---|---|---|---|---|
| F | 81.16% | 1.45% | 8.7% | 0 | 0 | 8.7% | 0 | |
| D | 15.22% | 65.22% | 0 | 4.35% | 4.35% | 4.35% | 6.52% | |
| J | 11.27% | 4.23% | 43.66% | 7.04% | 0 | 33.8% | 0 | |
| N | 1.27% | 7.59% | 2.53% | 81.01 | 0 | 0 | 7.59% | UR |
| S | 0 | 0 | 0 | 3.23% | 88.71% | 0 | 8.06% | 76.68% |
| A | 0.79% | 0.79% | 7.09% | 0 | 0 | 90.55% | 0.79% | WR |
| B | 0 | 1.23% | 0 | 6.17% | 6.17% | 0 | 86.42% | 78.69% |

Table 4.6: Confusion matrix of clean speech for RS96 feature on Berlin Database

| | F | D | J | N | S | A | B | |
|---|---|---|---|---|---|---|---|---|
| F | 72.46% | 0 | 5.8% | 1.45% | 0 | 20.29% | 0 | |
| D | 21.74 | 36.96% | 4.35% | 8.7% | 6.52% | 10.87% | 10.87% | |
| J | 7.04% | 7.04% | 38.03% | 4.23% | 0 | 40.85% | 2.82% | |
| N | 11.39% | 0 | 3.8% | 72.15% | 1.27% | 0 | 11.39% | UR |
| S | 0 | 0 | 0 | 3.23% | 90.32% | 0 | 6.45% | 68.35% |
| A | 1.57% | 0.79% | 11.02% | 0 | 0 | 85.83% | 0.79% | WR |
| B | 0 | 1.23% | 0 | 9.88% | 6.17% | 0 | 82.72% | 71.59% |

Table 4.7: Confusion matrix of 0dB (babble noise) speech for ACC384 feature on Berlin Database

| | F | D | J | N | S | A | B | |
|---|---|---|---|---|---|---|---|---|
| F | 69.57% | 2.90% | 10.14% | 5.8% | 0 | 10.14% | 1.45% | |
| D | 4.35% | 43.48% | 17.39% | 13.04% | 2.17% | 8.7% | 10.87% | |
| J | 18.31% | 12.68% | 28.17% | 8.45% | 0 | 29.58% | 2.82% | |
| N | 5.06% | 2.53% | 1.27% | 62.03 | 10.13% | 0 | 18.99% | UR |
| S | 1.61% | 0 | 0 | 6.45% | 87.10% | 0 | 4.84% | 63.57% |
| A | 0 | 2.36% | 9.45% | 0 | 0 | 84.25% | 3.94% | WR |
| B | 0 | 3.7% | 1.23% | 11.11% | 12.35% | 1.23% | 70.37% | 66.36% |

Table 4.8: Confusion matrix of 0dB (babble noise) speech for RS96 feature on Berlin Database

| | F | D | J | N | S | A | B | |
|---|---|---|---|---|---|---|---|---|
| F | 43.48% | 4.35% | 8.7% | 10.14% | 8.7% | 18.84% | 5.8% | |
| D | 23.91% | 21.74% | 13.04% | 4.35% | 0 | 15.22% | 21.74% | |
| J | 16.90% | 4.23% | 28.17% | 2.82% | 0 | 45.07% | 2.82% | |
| N | 21.56% | 6.33% | 1.27% | 35.44% | 13.92% | 6.33% | 15.19% | UR |
| S | 0 | 0 | 0 | 6.45% | 83.87% | 0 | 9.68% | 51.64% |
| A | 0.79% | 1.57% | 11.81% | 0 | 0 | 85.83% | 0 | WR |
| B | 4.94% | 2.47% | 2.47% | 12.35% | 13.58% | 1.23% | 62.93% | 56.07% |

Table 4.9: Confusion matrix of 0dB (white noise) speech for ACC384 feature on Berlin Database

| | F | D | J | N | S | A | B | |
|---|---|---|---|---|---|---|---|---|
| F | 72.46% | 0 | 11.59% | 0 | 2.9% | 13.04% | 0 | |
| D | 8.7% | 41.3% | 2.17% | 23.91% | 0 | 10.87% | 13.04% | |
| J | 9.86% | 9.86% | 38.03% | 5.63% | 0 | 35.21% | 1.41% | |
| N | 10.13% | 2.53% | 2.53% | 53.16% | 2.53% | 0 | 29.11% | UR |
| S | 0 | 0 | 0 | 4.84% | 87.10% | 0 | 8.06% | 68.22% |
| A | 0.79% | 3.15% | 9.45% | 0 | 0 | 85.83% | 0.79% | WR |
| B | 1.23% | 0 | 0 | 13.56% | 6.17% | 0 | 79.01% | 65.27% |

Table 4.10: Confusion matrix of 0dB (white noise) speech for RS96 feature on Berlin Database

| | F | D | J | N | S | A | B | |
|---|---|---|---|---|---|---|---|---|
| F | 57.97% | 4.35% | 8.7% | 8.7% | 5.8% | 11.59% | 2.9% | |
| D | 15.22% | 21.74% | 15.22% | 15.22% | 6.52% | 13.04% | 13.04% | |
| J | 7.04% | 5.63% | 29.58% | 4.23% | 0 | 49.30% | 4.23% | |
| N | 12.66% | 3.8% | 6.33% | 46.84% | 11.39% | 1.27% | 17.72% | UR |
| S | 0 | 0 | 0 | 8.06% | 74.19% | 0 | 17.74% | 54.29% |
| A | 3.15% | 0.79% | 11.81% | 0 | 0 | 84.25% | 0 | WR |
| B | 0 | 6.17% | 2.47% | 8.64% | 17.28% | 0 | 65.43% | 58.69% |

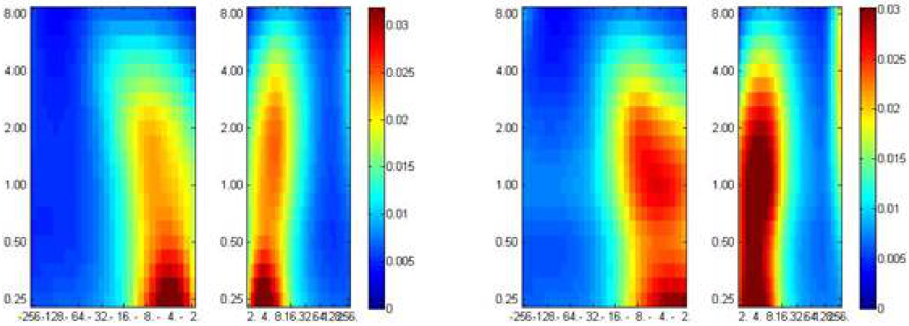


Figure 4.5: An overall clean speech RS plot for a fear sentence (right panel) and an angry sentence(left panel)

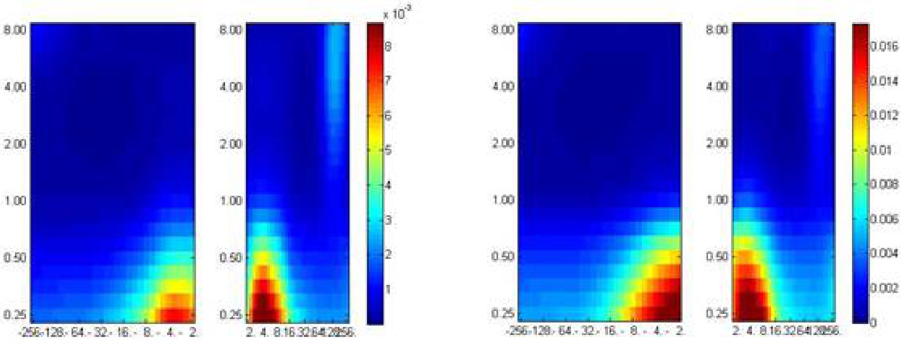


Figure 4.6: A clean speech RS plot of the 1st group (G1) for a fear sentence (right panel) and an angry sentence(left panel)

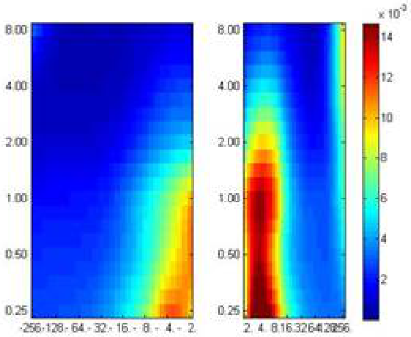


Figure 4.7: A clean speech RS plot of the 2nd group (G2) for a fear sentence (right panel) and an angry sentence(left panel)

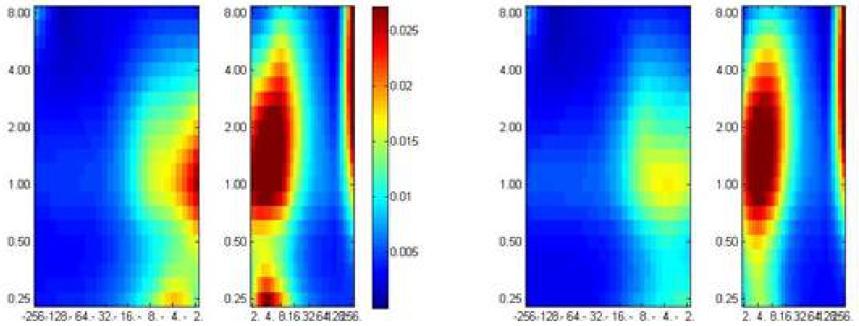


Figure 4.8: A clean speech RS plot of the 3rd group (G3) for a joy sentence (right panel) and an angry sentence(left panel)

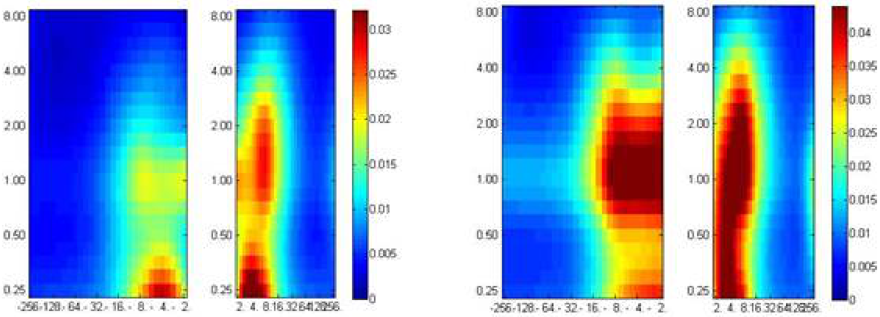


Figure 4.9: A clean speech RS plot of the 4th group (G4) for a fear sentence (right panel) and an angry sentence(left panel)

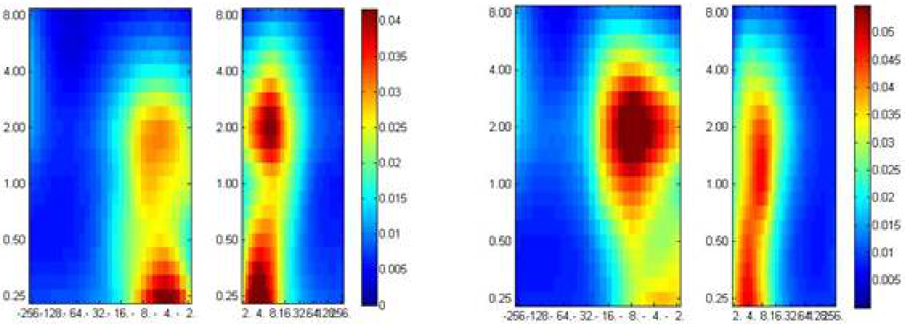


Figure 4.10: A clean speech RS plot of the 5th group (G5) for a fear sentence (right panel) and an angry sentence(left panel)

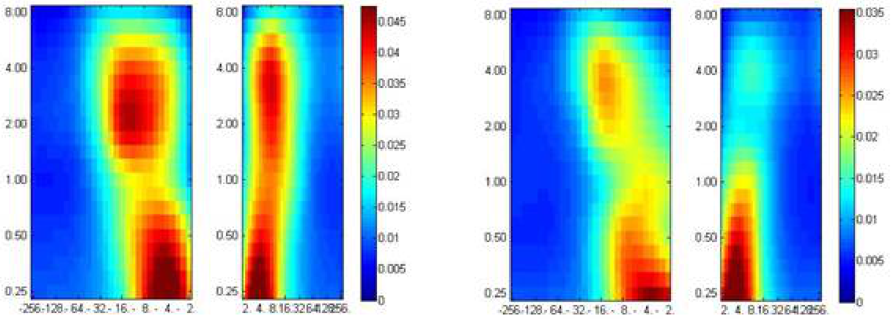


Figure 4.11: A clean speech RS plot of the 6th group (G6) for a fear sentence (right panel) and an angry sentence(left panel)

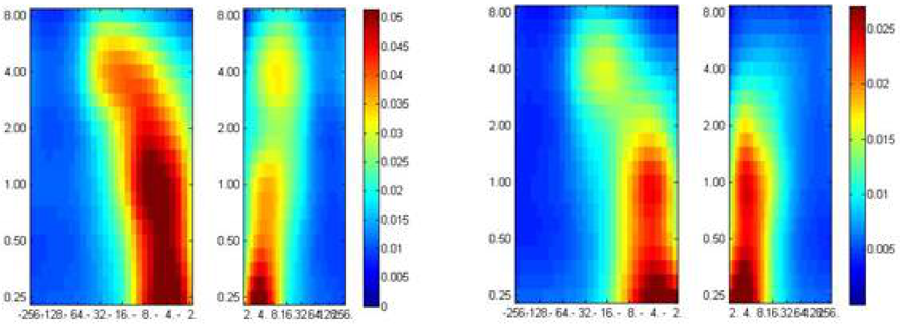


Figure 4.12: A clean speech RS plot of the 7th group (G7) for a fear sentence (right panel) and an angry sentence(left panel)

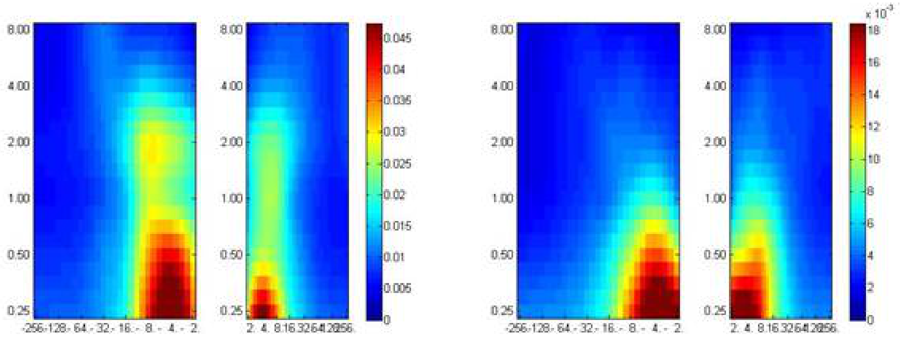


Figure 4.13: A clean speech RS plot of the 8th group (G8) for a fear sentence (right panel) and an angry sentence(left panel)

The detailed frequency range of each group is shown in table.4.11. The overlap of frequency range between adjacent groups is quite small. This inter-independence might be the reason that sub-group testing results show in table.4.12 are not better than the original $ACC384$ feature set, which cover all frequency regions.

Table 4.11: Frequency range for each groups

| | G1 | G2 | G3 | G4 | G5 | G6 | G7 | G8 |
|---|---|---|---|---|---|---|---|---|
| Freq(kHz) | 0.17-0.26 | 0.27-0.42 | 0.43-0.66 | 0.68-1 | 1-1.7 | 1.7-2.6 | 2.7-4.2 | 4.3-6.6 |

Table 4.12: Sub-group testing result

| | G1-G8 | G2-G8 | G3-G8 | G1-G7 | G2-G7 | G3-G7 |
|---|---|---|---|---|---|---|
| result | 76.54% | 76.37% | 75.64% | 76.50% | 76.49% | 74.92% |

The *ACC*384 feature set contain 8 frequency groups, which uniformly cover the frequency range of the 128 constant-Q cochlear filters. In other words, each frequency group covers 16 cochlear filters. To investigate the optimality of the frequency coverage, we divide the constant-Q filters into different number of groups, for example, 1 group, 2 groups, 4 groups, ..., and up to 128 groups, and use the *SRC* to recognize these emotions. The simulation results are shows in table 4.13 and Fig.4.14, and we find out that the more groups out of the 128 constant-Q filters, the higher the recognition rate we can get. Considering the computational load duo to high dimensionality, 8-group seems a reasonable choice.

Table 4.13: Simulation results in different number of groups of the constant-Q filters

| | 1 | 2 | 4 | 8 | 16 | 32 | 64 | 128 |
|---|---|---|---|---|---|---|---|---|
| UR | 62.09% | 69.84% | 74.31% | 76.56% | 76.54% | 77.22% | 77.26% | 77.42% |

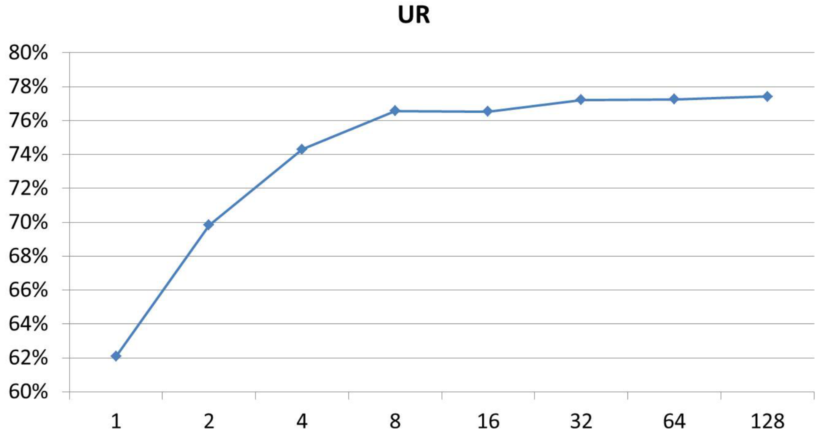


Figure 4.14: Simulation results in different number of groups of the constant-Q filters.

We notice that the simulation result of $inter384$ feature set is not very good while using the $SRC$ as the classifier. As introduced in section 2.2, the sparse representation classifier uses the training set as the basis and tries to find the best coefficients of each testing data; then the nearest subspace method is applied to determine the class of each test data. The $inter384$ feature set is built from a lot of functional features, such like mean, standard deviation, maximum value, maximum position, minimum value, and minimum position,..., and so on. In order to have a high recognition rates, a functional feature should have a similar pattern for a specific emotional class. However some of those functionals are not suitable for building a good subspace, for example, the maximum and minimum position. Fig.4.15 and Fig. 4.16 show the variance of functional features of energy and the 1st MFCC. The first 2 peak in each figure are maximum position and minimum position, respectively. The larger variance it is, the more uncertainty it has, and that is not good for our sparse representation classifier.

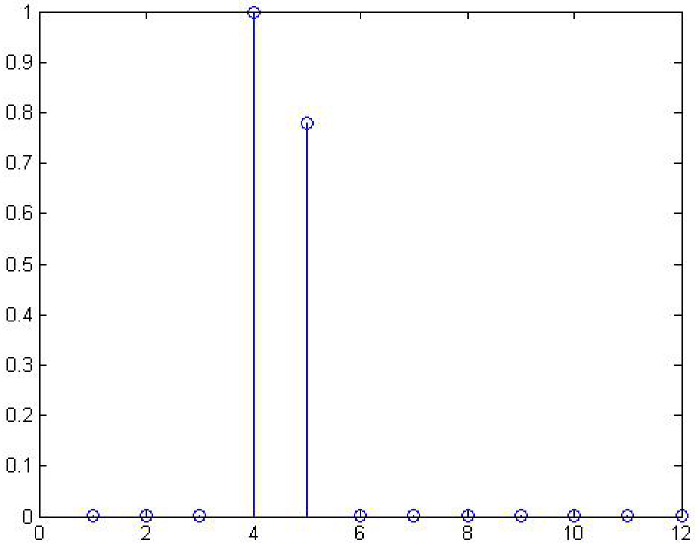


Figure 4.15: Variance of functional features of energy.

We then removed the position related feature from $inter384$ and com-

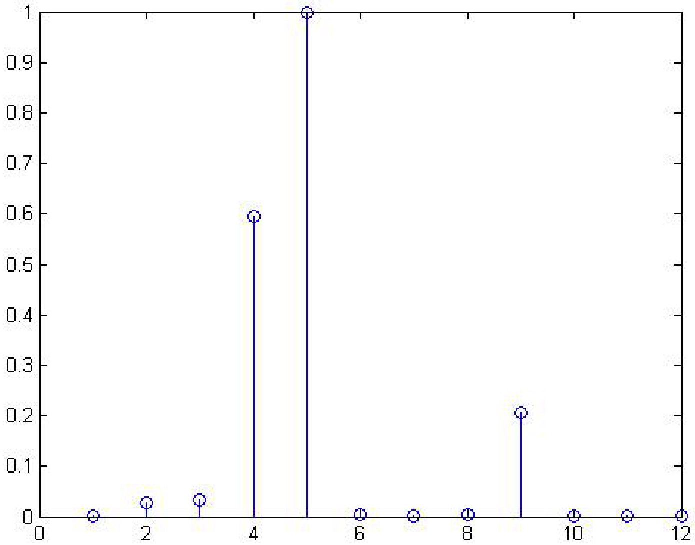


Figure 4.16: Variance of functional features of 1st MFCC.

pared with the original *inter*384 set, and the simulation result for one sample ”boredom” sentence is shows in Fig.4.17 and Fig.4.18. We can observe that Fig.4.17 has two groups of peaks which probably cause the error when using nearest subspace method and Fig.4.18 has only one group of peaks which is much more precise. The recognition rate of using the *inter*384 feature set without position related functional is 67.96%, which is higher than the original result 65.31%. Furthermore, we remove the features with the normalized variance large than $10^{-5}$, and the recognition rate goes up to 72.24%. Above discussion tell us that the variance should be small while using the sparse representation classifier. Similar result also applied to our *ACC*384 feature set, when we removed the features whose variances are larger than others, the recognition result goes up to 78.89%, which is better than the original result of 76.54%.

In order to compare with other published works, we also consider classifying four different emotion classes which are H, (happy), A, (Anger), S, (Sad), and N, (Neutral). To our knowledge, the best performance was demonstrated

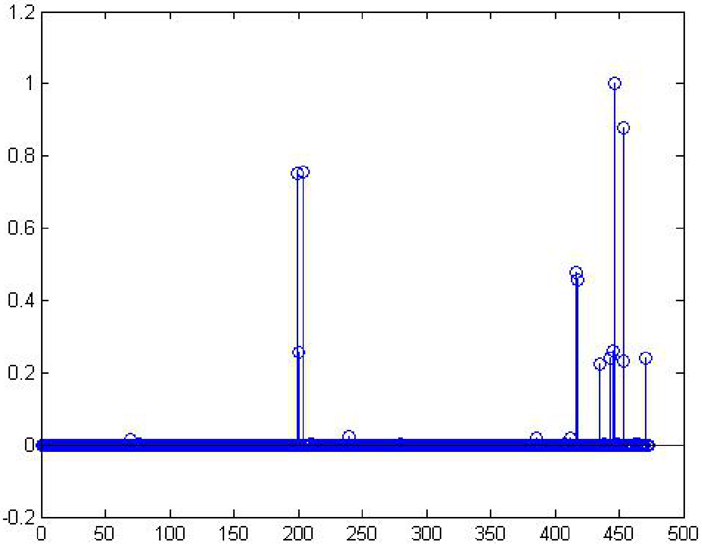


Figure 4.17: SRC coefficients of one boredom sentence using all *inter*384 feature set

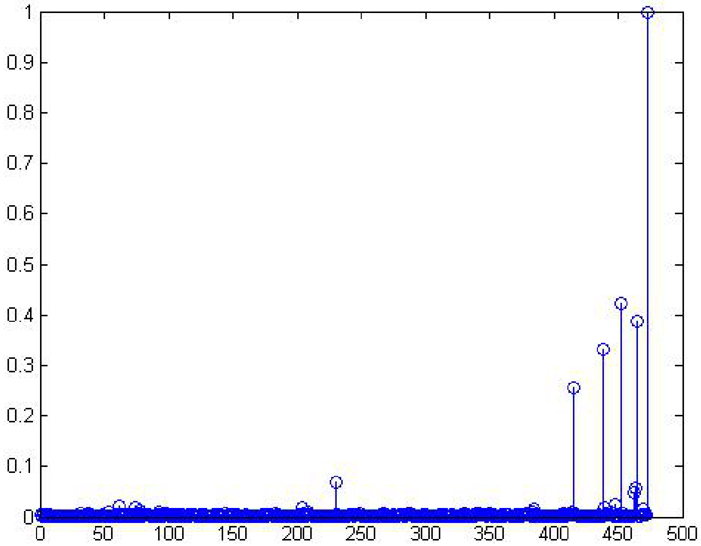


Figure 4.18: SRC coefficients of one boredom sentence using *inter*384 feature set without any position related features.

is [12], which is segmented an utterance into many segments and used the support vector machine as the first order classifier to calculate the probability for each class. The probability for each class was than used as the new feature set with the conventional Gaussian Mixture Model (GMM) as the second order classifier. The 4-class recognition rate reported in [12] is 84.7%, and the result of our $SRC + ACC384$ is 85.05%. We also test the result without the l1-norm term, and the result is 53.11%, which is much more worse than our proposed result.

## 4.3 Results on Aibo Corpus

Table 4.14: Simulation results on Aibo Corpus under babble noise using SRC

| | clean | 15dB | 10dB | 5dB | 0dB |
|---|---|---|---|---|---|
| $ACC384$ | 40.84% | 40.52% | 39.82% | 39.07% | 37.40% |
| $inter384$ | 38.79% | 35.62% | 34.26% | 33.32% | 32.30% |
| $RS96$ | 33.69% | 33.43% | 32.54% | 31.67% | 29.58% |

Table 4.15: Simulation results on Aibo Corpus under white noise using SRC

| | clean | 15dB | 10dB | 5dB | 0dB |
|---|---|---|---|---|---|
| $ACC384$ | 40.84% | 40.47% | 39.16% | 38.60% | 37.86% |
| $inter384$ | 38.79% | 36.67% | 34.14% | 33.92% | 32.53% |
| $RS96$ | 33.69% | 33.44% | 31.91% | 30.09% | 28.44 % |

Table 4.16: Simulation results on Aibo Corpus under babble noise using SVM

| | clean | 15dB | 10dB | 5dB | 0dB |
|---|---|---|---|---|---|
| $ACC384$ | 40.37% | 39.73% | 39.47% | 38.80% | 37.81% |
| $inter384$ | 38.19% | 38.33% | 37.86% | 37.28% | 35.87% |
| $RS96$ | 36.84% | 37.00% | 35.46% | 35.19% | 33.91% |

Table 4.17: Simulation results on Aibo Corpus under white noise using SVM

|  | clean | 15dB | 10dB | 5dB | 0dB |
|---|---|---|---|---|---|
| $ACC384$ | 40.37% | 40.07% | 39.08% | 38.45% | 37.56% |
| $inter384$ | 38.19% | 39.37% | 38.61% | 37.41% | 34.99% |
| $RS96$ | 36.84% | 36.56% | 36.13% | 34.71% | 33.54% |

Table 4.14 to table 4.17 shows the recognition rates on Aibo Corpus with different recognizer different types of noise and in difference SNRs. In order to cope with the data imbalance issue, we have randomly down-size other classes to have the same size as the smallest class. This method was used in our previous work [24], such that the weighted average recall (WR), i.e. recognition rate, and the un-weighted recall (UR) are equal. For better demonstration, the results of using different classifiers are shown in Fig.4.19 and Fig.4.20, respectively.

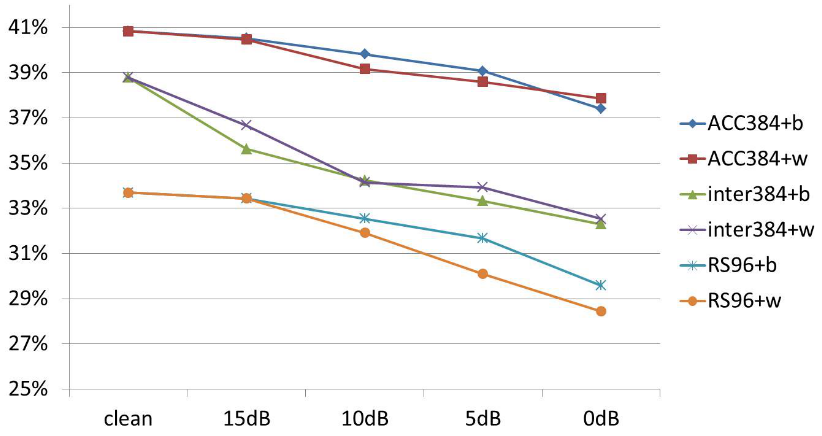


Figure 4.19: Simulation result on Aibo Corpus using SRC.

Fig.4.19 shows the simulation results for three different feature sets when using the $SRC$ classifier. For sparse representation classification, the best result here is $SRC + ACC384$. And Fig.4.20 also shows that the best feature set when using the $SVM$ classifier is still $ACC384$. We believe that if we use more frequency groups, the recognition result will improve, which has been

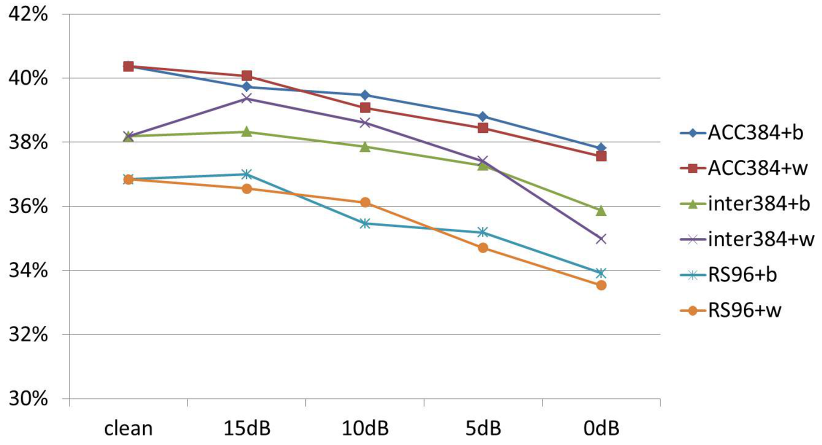


Figure 4.20: Simulation result on Aibo Corpus using SVM.

shown in the previous section with the Berlin Database.

It is worth noting that we used the downsizing instead of the up-sizing approach to balance the database. Such random downsizing, which ignores many speech sentences, might aggravate the over-training problem.

The Aibo Corpus contains instantaneous emotions, which are very different from the acted emotions of the Berlin Database, and its recording environment is not as good as one of the Berlin Database, so in other words, the Aibo corpus is considered much close to real life situations, and our experiment results show that the $ACC384$ feature set is much more suitable for real life applications. We also notice that some sentences contain bad clips which caused by microphone collision at the beginning or end of a sentence as shown in Fig.4.21. These bad clips are probably treated as voice by the published VAD system and seriously degraded our feature sets.

Comparing with the work published in The INTERSPEECH 2009 Emotion Challenge [16] (the baseline results of the $inter384$ feature set with $SVM$ are un-weighted average recall (UR) 38.2% and weighted average recall (WR) 39.2%), both our results ($SRC + ACC384$ is 40.84% and $SVM + ACC384$ is 40.37%) are better than the published baseline with the same number of feature dimensions.

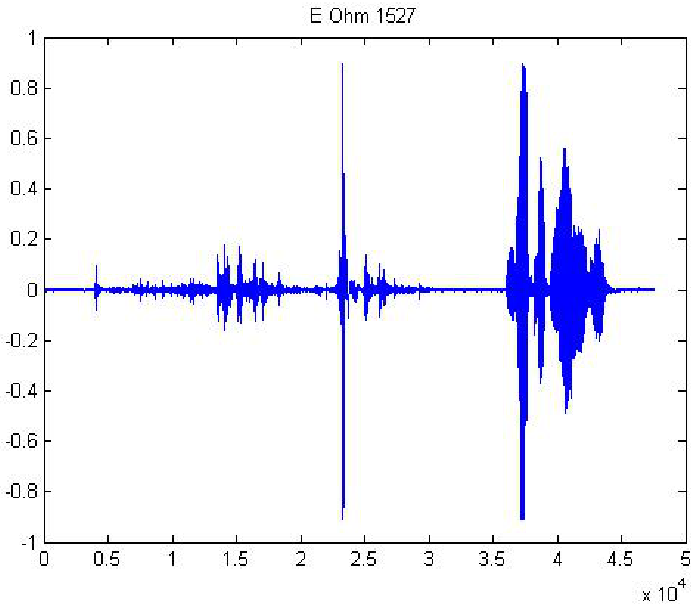


Figure 4.21: An example for bad clips happened at the range 2 to 2.5 in Abio corpus.

# Chapter 5

# Conclusion and Future Works

## 5.1 Conclusion

In the first part of this thesis, the experiment results on the Berlin Database show that our feature sets derived from the spectro-temporal modulation are much more robustness than the feature set proposed in the INTERSPEECH 2009 Emotion Challenge. The simulation results of two classifiers, $SVM$ and $SRC$, show that the $SRC$ performs well if the variance of feature is small enough, at this part, our $ACC384$ feature set is better than $inter384$ feature set. And the difference between our $RS96$ feature set and our $ACC384$ feature set is that the $ACC384$ feature set carries phase information of the modulation envelops. This phase information is beneficial to the classification such that the recognition rates go up . We discussed the feature dimensions of our spectro-temporal modulation, the results show that as the feature dimension increases, the un-weighted average recall increases, too. And about the sparse representation classification, the smaller variance of features, the higher recognition rates we can obtain. In the second part of this thesis, we used the Aibo Corpus as our test database. Again, the experiment results show that our feature sets are much more robust than the $inter384$ feature set. And we compare our experiment results in this two databases with other published works. For Berlin Database, our system performs better than the one in [12], and for the Aibo corpus, the simulation results of our proposed system are better than the baseline provied by the INTERSPEECH 2009 Emotion Challenge.

Overall speaking, the feature sets derived from our spectro-temporal modulation are much more robust than the *inter*384 feature set, and our feature sets are more suitable for the sparse representation classifier because of the small variance of features.

## 5.2 Future Works

In order to improve the recognition results for both Aibo Corpus and Berlin Database, some points can be take into consideration. First, a good feature selection method is needed. In our previous work, the sequential forward floating selection (SFFS) method is used to find the best subset in our feature sets, or we can use principal component analysis (PCA) to down-size our feature sets and investigate the performance in each different dimension, just like [28]. Second, the data balance issue is also important. Instead of down-sizing the majority as we have done in this study, the SMOTE (Synthetic Minority Over-Sampling TEchnique) applied in the INTERSPEECH 2009 Emotion Challenge may provide effective ways for balancing data sets.

As mentioned before, our spectro-temporal modulations provided 4 dimensional (time, frequency, rate, and scale) features. In this work, we use 3 dimensional (frequency, rate, and scale) features with the sparse representation classifier, and the performance is better than other published systems. Since the performance of our 3 dimensional feature set, *ACC*384, is better than our 2 dimensional feature set, *RS*96, we believe that the overall 4 dimensional feature set can capture much more detailed time-varying information and provide higher performance than our *ACC*384 feature set.

# Bibliography

[1] T. Chi, P. Ru, and S. A. Shamma, "Multiresolution spectrotemporal analysis of complex sounds," *The Journal of the Acoustical Society of America*, vol. 118, no. 2, pp. 887–906, 2005.

[2] 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, no. 5, pp. 2719–2732, 1999.

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

[4] D. Ververdis and C. Kotropoulos, "A state of the art review on emotional speech databases," in *Interspeech, pp.2249-2252*, 2003.

[5] C. Busso, S. Lee, and S. Narayanan, "Using neutral speech models for emotional speech analysis," in *Interspeech 2007 - Eurospeech*, Antwerp, Belgium, August 2007, pp. 2225–2228.

[6] C. Busso, A. Metallinou, and S. Narayanan, "Iterative feature normalization for emotional speech detection," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2011)*, Prague, Czech Republic, May 2011, pp. 5692–5695.

[7] B. Schuller, A. Batliner, S. Steidl, and D. Seppi, "Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge," *Speech Communication*, vol. 53, pp. 1062 – 1087, 2011.

[8] F. W. . G. R. Bjorn Schuller, Dejan Arsic, "Emotion recognition in the noise applying large acoustic feature sets," in *Proc. Speech Prosody*, 2006.

[9] B. Schuller, A. Batliner, D. Seppi, S. Steidl, T. Vogt, J. Wagner, L. Devillers, L. Vidrascu, N. Amir, L. Kessous, and V. Aharonson, “The relevance of feature type for the automatic classification of emotional user states: low level descriptors and functionals,” in *INTERSPEECH*, 2007, pp. 2253–2256.

[10] C. Busso, S. Lee, and S. Narayanan, “Analysis of emotionally salient aspects of fundamental frequency for emotion detection,” *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 17, no. 4, pp. 582–596, 2009.

[11] B. Schuller and G. Rigoll, “Timing levels in segment-based speech emotion recognition,” in *INTERSPEECH*, 2006.

[12] J. H. Jeon, R. Xia, and Y. Liu, “Sentence level emotion recognition based on decisions from subsentence segments,” in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, 2011, pp. 4940–4943.

[13] K. Audhkhasi and S. Narayanan, “Emotion classification from speech using evaluator reliability-weighted combination of ranked lists,” in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, 2011, pp. 4956–4959.

[14] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, “A database of german emotional speech,” in *INTERSPEECH*, 2005, pp. 1517–1520.

[15] S. Steidl, “Automatic classification of emotion-related user states in spontaneous children¡¦s speech,” in *Logos Verlag*, 2009.

[16] B. Schuller, S. Steidl, and A. Batliner, “The interspeech 2009 emotion challenge,” in *INTERSPEECH*, 2009, pp. 312–315.

[17] Y. Li and A. Ngom, “Sparse representation approaches for the classification of high-dimensional biological data,” *BMC Systems Biology (in press, 2013)*, 2013.

[18] P. C. Loizou, *Speech Enhancement: Theory and Practice.* CRC, 2007.

[19] N. Grimault, S. P. Bacon, and C. Micheyl, “Auditory stream segregation on the basis of amplitude-modulation rate,” *The Journal of the Acoustical Society of America*, vol. 111, no. 3, pp. 1340–1348, 2002.

[20] R. P. Carlyon, B. C. J. Moore, and C. Micheyl, “The effect of modulation rate on the detection of frequency modulation and mistuning of complex tones,” *The Journal of the Acoustical Society of America*, vol. 108, no. 1, pp. 304–315, 2000.

[21] A. S. Bregman, *Auditory scene analysis: The perceptual organization of sound.* MIT, 1990.

[22] J. Wright, A. Yang, A. Ganesh, S. Sastry, and Y. Ma, “Robust face recognition via sparse representation,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 31, no. 2, pp. 210–227, 2009.

[23] S.-J. Kim, K. Koh, M. Lustig, S. Boyd, and D. Gorinevsky, “An interior-point method for large-scale l1-regularized least squares,” *Selected Topics in Signal Processing, IEEE Journal of*, vol. 1, no. 4, pp. 606–617, 2007.

[24] L. Y. Yeh and T. S. Chi, “Spectro-temporal modulations for robust speech emotion recognition,” in *INTERSPEECH*, 2010, pp. 789–792.

[25] F. Eyben, M. Wöllmer, and B. Schuller, “Opensmile: the munich versatile and fast open-source audio feature extractor,” in *Proceedings of the international conference on Multimedia.* ACM, 2010, pp. 1459–1462.

[26] C.-C. Chang and C.-J. Lin, “LIBSVM: A library for support vector machines,” *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011, software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.

[27] A. Varga and H. J. 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 Communication*, vol. 12, no. 3, pp. 247–251, 1993.

[28] G. R. A. Yannis Panagakis, Constantine Kotropoulos, “Music genre classification via sparse representations of auditory temporal modulations,” in *EUSIPCO*, 2009.