國立交通大學

電信工程研究所

碩士論文

論情緒辨識架構中時頻調變特徵參數的強健性

On Robustness of Spectro-Temporal Modulation Features in an Emotion Recognition Framework

研 究 生：許晉誠　　Student: Chin-Cheng Hsu

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

中華民國一百年六月

論情緒辨識架構中時頻調變特徵參數的強健性

On Robustness of Spectro-Temporal Modulation Features in an Emotion Recognition Framework

研究生：許晉誠　　Student: Chin-Cheng Hsu

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 2011
Hsin-Chu, Taiwan, Republic of China

中華民國一百年六月

論情緒辨識架構中時頻調變特徵參數的強健性

學生：許晉誠　　　　　　　　指導教授：冀泰石 博士

國立交通大學電信工程研究所

感知訊號處理實驗室

摘要

雜訊無論在情緒辨識或其他任何應用中，都是相當困擾的問題。當前最常見的做法是採取匹配訓練（matched condition）來對抗雜訊；相反的，本文考慮不匹配條件、只訓練單一分類器來對抗各種情況下的雜訊。實驗結果顯示：就算在最嚴格的不匹配條件下，本文採用的時頻調變特徵參數組也有極為穩健的表現。文中亦討論該時頻調變特徵參數的特性以及它如何受雜訊干擾所影響。本實驗包含四項變因：兩組資料庫（Berlin Emotional Speech Database、Aibo Emotional Speech）、兩種雜訊（white noise、babble noise）、兩組特徵參數（spectro-temporal modulation features、INTERSPEECH 2009 Emotion Challenge features）、兩種訓練-測試條件（slack or strict mismatched condition）。針對資料失衡的問題，本文則提出結合效度的樣本合成方案來改善。

On Robustness of Spectro-Temporal Modulation Features in an Emotion Recognition Framework

Student: Chin-Cheng Hsu   Advisor: Dr. Tai-Shih Chi

Institute of Communication Engineering

National Chiao-Tung University

Perception Signal Processing Laboratory

Abstract

Noise is an annoying problem either in emotion recognition or in other applications. Previous research has considered matched condition to counter it. This article, on the contrary, considers mismatched condition which trains only one classifier that confronts all kinds of situation. Experiments show that the proposed feature set, which contains spectro-temporal modulation information, is robust, indicating that the mismatched training/testing condition is feasible. This paper also discussed the properties of the proposed features and how noise affected the features. The experiments included four variables: two databases (Aibo Emotion Corpus and Berlin Emotional Speech Database), two types of noise (additive white Gaussian noise and babble noise), two feature sets (spectro-temporal modulation features and INTERSPEECH 2009 Emotion Challenge features), and two conditions (slack and strict mismatched conditions). As for the issue of data imbalance, a synthetic method based on emotion validity was proposed to deal with it.

There’s no big conceptual difference between solving intellectual problems by the brain or by a computer.

Cybernetics, 1948, Norbert Wiener

On the contrary, their experience was just like some one taking from various places hands, feet, a head, and other pieces, very well depicted, it may be, but not for the representation of a single person; since these fragments would not belong to one another at all, a monster rather than a man would be put together from them. Hence in the process of demonstration or "method", as it is called, those who employed eccentrics are found either to have omitted something essential or to have admitted something extraneous and wholly irrelevant.

On the Revolutions of the Heavenly Spheres, 1543, Nicolaus Copernicus

Content

List of Figures

List of Tables

Preface

This thesis is devoted to the Perception Signal Processing Laboratory, which studies and promotes the spectro-temporal modulation features inspired by human auditory perception system. Previous research showed that the spectro-temporal modulation features, also known as rate-scale (RS) features, were robust against several types of noise. My work here has two major goals: first, to verify that the RS features can be used for recognizing realistic emotion and second, to extend the auditory scheme to a broader use against stricter conditions.

The RS features have some fascinating traits. For one thing, it might be an engineering analogy to the auditory cortex in our brain. For another, it transmute sound wave, varying in durations, into a fixed plot which is nearly sufficient for the recognition of affect. I always think that what lies within the RS features is considerably potent. Nevertheless, I do not think the potential is fully exploited.

Machine learning, one of the most eminent fields in contemporary academia, might very well become prevailing. It's a pity that almost none in our lab recognizes its importance. One of the goals I set to my research is to construct a framework neat and tight that demonstrates the conciseness and utility the discipline can achieve.

As to my research, recognizing emotions in speech, however complete or incomplete, the legacy I leave to future generations is not the work itself, but rather the "research programme". Even now, some of my early anticipations have been realized.

I would like to conceive my deepest gratitude to professor Tai-shih Chi for his extraordinary insight and patience. Discussions with Master Andy and Big Tree were very critical as well as helpful. Without the above "Big Three", this research could probably not get so far. Lan-ying's previous work was an outstanding attempt and it was also my early guidance. I thank all my friends and my family for their support. Finally, I want to give my special thanks to Mr. M. Heidegger and Mr. S. R. Wang for their timely inspirations.

Jeremy Hsu

Hsinchu
June, 2011

Chapter 1

Introduction

1.1 Motivation

As one primary form that shows human affective information, emotions are the most intriguing and mysterious subject in both science and society. Discussions about emotions can be traced back to thousands of years ago, but scientific theories and computational analysis thrive in just about a hundred years. Although psychologists still have difficulty answers the underlying ontological questions about emotions, phenomenological research prospers unexpectedly fast and finally establishes its achievements.

One main reason that renders the research of emotion so fascinating is that the research is interdisciplinary. Psychology, image processing, machine learning society, and audio processing are among those disciplines. The whole research about emotion can be divided into several parts: psychology takes the ontology, refining the whole theoretical framework; other disciplines relate their research in the phenomenology, making the analysis and application possible. Interdisciplinary brings diversity as well as specialization.

Most efforts of emotion research are devoted to recognition which comprises two underlying parts: feature extraction and regression/classification. While the computing community and statistics take care of second part, audition and vision research take the first. The speech processing community also plays an important role since speech is a major expressive form of emotion. Theories about temporal sequences, short-term spectral analysis, and prosodic properties become well-known basic knowledge in speech community. As a member of this community, we try to respond to the problems that arise in the recognition of emotion in speech.

1.2 Current Research Interest

The rapidity of the development of emotion-related research is still increasing in recent years. Since Picard popularized the concept of “affective computing” fifteen years ago (Picard, 1997), more and more application fields are gaining interest in the research of affect. Medical caring for infants and the elder, psychiatric analysis, call center services, ticket reservation and retailing systems and human computer interaction designs are among the list. Back to the academia the INTERSPEECH has held several affect-related challenges including the INTERSPEECH 2009 Emotion Challenge, the 2010 Paralinguistic Challenge and the 2011 Speaker State Challenge. Research topics have shifted from a narrow definition "emotion" to a more wide sense "affect" and the process of generalization of the definition of affect is not rest yet.

Earlier attempts on acted emotions either in visual or auditory modality had been successful. The recognition rate on automatic systems got very close to that on human labelers. However, when it comes to natural or spontaneous emotions, the performance dropped drastically. We will come back to this issue later.

Investigation into robustness is a rather new attempt (Schuller et al., 2011), and the scope is rather limited. First, most discussions are still confined to additive noise, ignoring the more realistic form of Lombard effect which is the alternation of the way people speak in apparently noisy environments. Second, present approaches of matched condition are very basic, having some assumptions that are not so realistic. Last, there is no suitable database dedicated to the investigation of robustness.

1.3 Related Work

INTERSPEECH 2009 Emotion Challenge

Organized in 2009 to find out the best feature set and the best classifier, the challenge was also the official start for spontaneous emotion recognition. A spontaneous emotion database, Aibo Emotion Corpus, was first openly available in the competition. The baseline achieved 38.2% UR using 384 features and a support vector machine (SVM) as its classifier. Feature selection and data imbalance were also discussed throughout the challenge. The following chart shows participant teams and the brief information about their work.

The challenge reveals a wide variety of features in use. Most features were cepstral or prosodic (Mel-frequency cepstral coefficients (MFCCs), dynamic features such as pitch contour), and SVM and Gaussian mixture model were the most frequently adopted classifiers.

Table 1.3.1: Classification results from INTERSPEECH 2009 Emotion Challenge.

				Barra-Chicote et al.					
	A	E	N	P	R	Recall	Precision	UR	UP
A	**211**	294	38	7	61	34.53%	25.51%	38.25%	32.81%
E	130	**1166**	127	6	78	77.37%	26.96%	WR	
N	403	2662	**1446**	247	613	26.92%	80.69%	36.68%	
P	16	39	56	**54**	50	25.12%	15.21%	GR	
R	67	164	125	41	**149**	27.29%	15.67%	34.56%	
				Luengo et al.					
	A	E	N	P	R	Recall	Precision	UR	UP
A	**414**	97	46	26	28	67.76%	17.65%	41.38%	31.94%
E	428	**687**	284	45	64	45.56%	35.27%	WR	
N	1302	1089	**2340**	423	223	43.52%	80.66%	43.35%	
P	29	11	78	**88**	9	40.93%	12.77%	GR	
R	172	64	153	107	**50**	9.16%	13.37%	34.71%	
				Lee et al.					
	A	E	N	P	R	Recall	Precision	UR	UP
A	**290**	171	65	63	22	47.46%	21.23%	41.57%	31.02%
E	210	**752**	325	136	85	49.87%	36.02%	WR	
N	748	1094	**2057**	1109	369	38.26%	78.51%	39.87%	
P	23	13	39	**131**	9	60.93%	8.01%	GR	
R	95	58	134	197	**62**	11.36%	11.33%	36.26%	
				Bozkurt et al.					
	A	E	N	P	R	Recall	Precision	UR	UP
A	**319**	191	59	12	30	52.21%	24.24%	40.90%	34.11%
E	217	**964**	256	8	63	63.93%	33.00%	WR	
N	656	1638	**2516**	212	355	46.79%	80.90%	47.83%	
P	19	18	94	**50**	34	23.26%	15.24%	GR	
R	105	110	185	46	**100**	18.32%	17.18%	36.69%	

*Only 4 groups of participants published their confusion matrix.

*UR: unweighted recall rate; WR: accuracy; GR: geometric mean of class-wise recall rate. Discussions about assessment metrics can be found in Chapter 3.

Features

Features are still an ongoing issue in emotion recognition research. Popular vocal features include duration, intensity, pitch, voice quality, intonation contour, spectral features, prosodic features, wavelet features, and non-linguistic vocalizations. However, their true utility is under study. For instance, some questioned if duration is a useful feature (Burkhardt et al., 2009); others negate the use of intonation (Bänziger and Scherer, 2005).

1.4 Previous Work

Our current system framework is based on Yeh's (Yeh, 2010). The system, the materials and even the results seem the same at first glance, but in reality the framework has been way more than modified. Everything, even the research paradigm, changed. She realized a system that utilized LibSVM (Chang and Lin, 2011) as classifier and spectro-temporal analysis as feature set. The system worked fine on Berlin Emotional Speech Database, but it did not perform well on Aibo Emotion Corpus.

Taking her aim at robustness, Yeh attempted to consider mismatched condition for training and testing, meaning that the training is under only one condition (mostly using pure speech) and the trained model has to deal with all kinds of condition (different noise type or signal-to-noise ratio).

Yeh's paradigm gave four possible directions which later became my starting points. First, the new database, Aibo Emotion Corpus, is highly imbalanced; the majority class has samples 24 times more than the minority does. How to counter the imbalance is an issue. Second, kernels other than linear ones might help. Third, feature selection or reduction may be considered more. Fourth, higher statistics of the RS feature may be taken into consideration. All of the four possible improvements are properly done or at least have a decent investigation during my research phase.

1.5 The Research Paradigm

Our research paradigm, or "research programme" (in Imre Lakatos' phrase), is always clear: to develop a robust affect recognizer. Robustness against noise is one start that spectro-temporal modulation features bring. The ultimate goal is to refine the recognition system to be robust against even speaker. During my research, I can only narrow the ambition down to a simpler version: to verify the robustness spectro-temporal modulation features can offer.

To sum up, my research paradigm has the following structure:

1. **Hard core:** To verify that RS features is still robust (against noise) in stricter cases (e.g. when the case is recognizing spontaneous emotion).
2. **Protective Belts:** Factors that prevent RS features from performing full potential should be ruled out
3. **Positive Heuristics**

Robust against noise
Robust performance of both acted and spontaneous emotions

An obvious positive heuristic here is to draw stricter condition to training and testing condition and see if the RS feature set still performs robust. More specifically, the paradigm asks us to do tests on mismatched conditions. Previously, experiments under matched conditions, i.e. training and testing samples that have the same distortion or noise, have been studied and bolstered. However, matched condition is just a start, never the end. Yeh commenced mismatched condition and now it is our time to carry through it.

The only thing I tried to do is to verify that even under a very strict condition, there are still features that are not disqualified. Many features significantly change their characteristics under the effect of noise, energy profile being one of those. Since human beings can recognize emotion, speech content, or other meaningful information in noisy speech, there is no reason machines cannot. This gives us interests to examine RS features which have inspiration origins from human audition.

This thesis is structured as follows. Chapter 2 covers related work and background knowledge. Chapter 3 lists detailed information about the materials and methods. Chapter 4 is for experiments and discussions. Finally, Chapter 5 illustrates the big picture and some specific future work.

Chapter 2

Literature Review, Related Work and Background Knowledge

2.1 Machine Learning

Recognition is essentially a question of classification (or regression), and the whole process can be purely mathematical. Rooted deeply in statistics and aided by computer, automatic classification has become a powerful well-structured instrument. Machine learning which offers such instruments became one of the most prominent research fields in recent years. Some famous classifiers has been employed even in realistic applications; artificial neural networks (ANN), naïve Bayesian classifier, (NBC), Bayesian logistic regression (BLR), relevance vector machine (RVM), just to name a few. This section explains the reason the support vector machine (SVM) (Vapnik, 1995) is adopted and briefs some background knowledge about it.

2.1.1 Kernel methods and sparse kernel machines

In supervised learning, there are three main perspectives to solve a regression or classification problem. Generative models attempt to model the distribution of inputs as well as the outputs, explicitly or implicitly. Discriminative models only model the posterior distribution. Discriminant functions, however, concern nothing about distribution but seek the decision boundaries. SVM, which is a member of the last category of learning, can still be interpreted in the light of probability. This interpretation will be given later.

Kernel method and the dual form

All regression or classification problems have a similar form of solution:

$$y = f(\vec{w}^{T}\varphi(\vec{x}) + b)$$

y: algorithm output (regression target or class label)

t_n: ground truth (regression target or class label).

$\vec{w}$: weighting vector (or equivalently, the normal vector of the decision boundary in classification case).

b: bias term

$\vec{x}$: input vector

$\varphi(\vec{x})$: feature vector

$f(\cdot)$: activation function that transforms regression target into class labels.

Now, let us consider a linear regression case (for it shares the same core as classification) whose parameters (weighting vector) are determined by minimizing a regularized minimum mean squared error (MMSE) criterion where the error function is given by

$$J(\vec{w}) = \frac{1}{2}\sum_{n=1}^{N}\{\vec{w}^T\varphi(\vec{x}_n) - t_n\}^2 + \frac{\lambda}{2}\vec{w}^T\vec{w}, \text{ where } \lambda \geq 0.$$

and $n = 1,2,\ldots,N$ denotes sample indices

The optimal solution (by differentiating $J(\vec{w})$ w.r.t. $\vec{w}$) for $\vec{w}$ takes the form of

$$\vec{w} = -\frac{1}{\lambda}\sum_{n=1}^{N}\{\vec{w}^T\varphi(\vec{x}_n) - t_n\}\varphi(\vec{x}_n) = \sum_{n=1}^{N} a_n\varphi(\vec{x}_n) = \Phi\vec{a},$$

with $\vec{a} = -\frac{1}{\lambda}\sum_{n=1}^{N}\{\vec{w}^T\varphi(\vec{x}_n) - t_n\}$

If we substitute $\vec{w}$ with $\Phi\vec{a}$ and define the kernel matrix as

$K = \Phi^T\Phi$, where$\Phi_{nk} = \varphi_k(\vec{x}_n)$ is the design matrix, we have

$$J(\vec{a}) = \frac{1}{2}\vec{a}^T KK\vec{a} - \vec{a}^T K\vec{t} + \frac{1}{2}\vec{t}^T\vec{t} + \frac{\lambda}{2}\vec{a}^T K\vec{a}$$

Setting the gradient of $J(\vec{a})$ with respect to $\vec{a}$ to 0, we obtain

$$\frac{dJ}{d\vec{a}} = KK\vec{a} - K\vec{t} + \lambda K\vec{a} = 0 \Rightarrow \vec{a} = (K + \lambda I_N)^{-1}\vec{t}$$

Substituing a back into $y = \vec{w}^T\varphi(\vec{x})$, we get $y(\vec{x}) = k(\vec{x})^T(K + \lambda I_N)^{-1}\vec{a}$

$k\left(\vec{x}_i{}^T\vec{x}_j\right) = \varphi(\vec{x}_i)^T\varphi\left(\vec{x}_j\right)$ is known as the kernel function.

This is the **dual form** of the original problem. In the dual form, the prediction (regression target) can be made solely by the training set (cf. [Bishop, 2006] and [Ng, 2009], for Mercer's theorem and other detail limitation a kernel must obey). We recognize that this form of decision (kernel method) belongs to discriminant function in which probability are not involved at appearance. In the next section, we attempt to link the two.

Probability interpretation

Assume the prior distribution of w obeys an isotropic Gaussian of the form:

$$\vec{w}\sim N(0,\alpha^{-1}I)$$

Given the training samples and basis functions (or equivalently, kernel function), we have the design matrix (or equivalently, the kernel matrix) and then we know that $\vec{y}=\Phi\vec{w}$ is a linear combination of Gaussian vectors and hence is a Gaussian vector.

$$E[\vec{y}]=\Phi E[\vec{w}]=\vec{0}$$

$$\mathrm{cov}(\vec{y})=E[\vec{y}\vec{y}^{T}]=E\left[\Phi\vec{w}\vec{w}^{T}\Phi^{T}\right]=\Phi\left(\frac{1}{\alpha}I\right)\Phi^{T}=\frac{1}{\alpha}K$$

Therefore, $\vec{y}\sim N(\vec{0},\frac{1}{\alpha}K)$. The kernel matrix is the dominating factor of the covariance matrix. This bridges the gap between the two paradigms of discriminant function and discriminative model.

Now we further assume that we can model the problem using Gaussian process:

$$t_n=y_n+\epsilon_n,\text{ where }\epsilon_n\sim N(0,\sigma^2)\text{ is prediction error}$$

The joint distribution of the regression target $\vec{t}=(t_1,\ldots,t_N)^T$ conditioned on the values of $\vec{y}=(y_1,\ldots,y_N)^T$ is given by an isotropic Gaussian of the form

$$\vec{t}=\vec{y}+\vec{\epsilon},\text{ where }\epsilon_n\sim N(0,\sigma^2 I)$$

1) $\vec{t}=\vec{y}+\vec{\epsilon}$,

2) $\vec{y}\sim N(\vec{0},\frac{1}{\alpha}K)$

3) $\vec{\epsilon}\sim N\left(\vec{0},\sigma^2 I\right)$

⇨ $\vec{t}\sim N(0,C)$ where $C=\sigma^2 I+\frac{1}{\alpha}K$. This concludes the training phase.

In the testing phase, the regression target, say t_{N+1}, has correlation with the targets in the training phase. Incorporating training samples $\vec{t}$ with testing sample t_{N+1} into $\vec{t}_{N+1}$, according to Bayes' theorem for Gaussian variable, we know

$$\vec{t}_{N+1}\sim N\left(\vec{0},C_{N+1}\right),\text{ where }C_{N+1}=\begin{bmatrix}C_N & k\\ \vec{k}^{T} & c\end{bmatrix},$$

$$\text{where }\vec{k}=k(\vec{x}_n,\vec{x}_{N+1}),n=1,2,\ldots N\text{ and }c=k(\vec{x}_{N+1},\vec{x}_{N+1})$$

Therefore

$$t_{N+1}|\vec{t}\sim N(\vec{k}^{T}C_N^{-1}\vec{t},c-\vec{k}^{T}C_N^{-1}\vec{k})$$

Since C_N is determined by training data and c and $\vec{k}$ are determined by testing sample, the decision is made naturally. Starting from divergent point of views, discriminant function and probabilistic discriminative model finally reach the same end.

Despite of the same purpose and mathematical analysis, why should someone adopting kernel methods and the dual form if in reality the number of samples (N) is larger than dimension (the kernel matrix is $N \times N$ while solving w, we only have to deal with $M \times M$ matrix where M is the dimension of basis function)? This is because in some cases, not all training samples are necessary and thus the kernel matrix becomes sparse. A sparse kernel is equivalent to a reduced $N \times N$ matrix. Sometimes the reduced N can be very small-- even smaller than M. This justifies the use of the dual form. In the next section, we will introduce one type of sparse kernel machines: the support vector machines (SVM).

2.1.2 Support Vector Machines

Attempts to solve binary classification problems were made in early days. Frank Rosenblatt's perceptron is among the early attempts (Rosenblatt, 1962). Perceptrons have several shortcomings. It cannot solve overlapping classes and its decision boundary might not be optimal (Bishop, 2006).

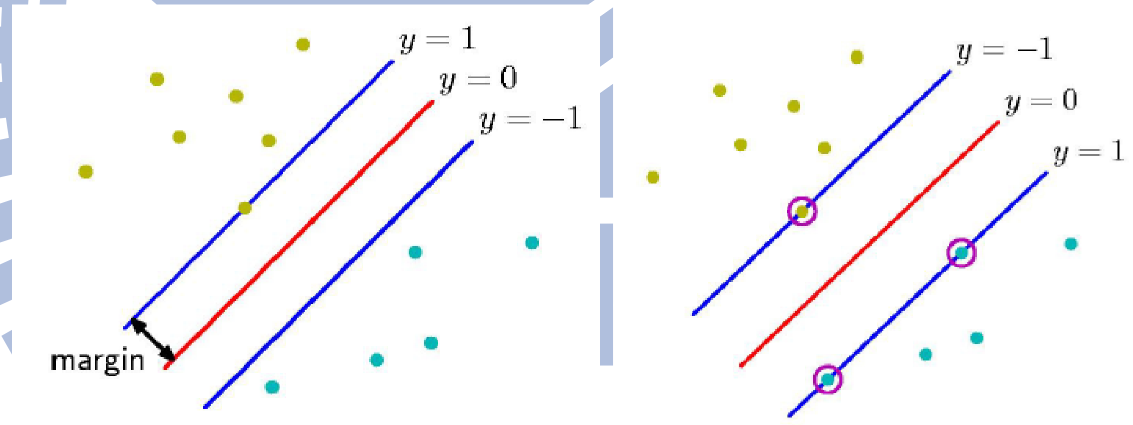


Figure 2.1.1

Left: Correct binary classification without maximizing margin.

Right: Maximum margin classifier.

Source: Lecture notes from Machine Learning. Wang, 2011.

Originally devised for linearly separable binary classification problem, the SVM attempted to maximize the margin between classes. Let the class label be 1 and -1, the distance of any point x to the decision boundary is

$$\frac{t_n y(\vec{x}_n)}{\|\vec{w}\|}, t_n \in \{-1, 1\}$$

Define functional margin as “the shortest distance between the decision boundary and the class”, i.e. functional margin is the distance between the boundary and the sample which is nearest to it:

$$\gamma = \min\left\{\frac{t_n y(\vec{x}_n)}{\|\vec{w}\|}\right\} = \frac{t_n(\vec{w}^T\varphi(\vec{x}_n + b))}{\|\vec{w}\|}$$

Now we try to find a decision boundary that has maximum margin, so we solve

$$\arg\max_{\vec{w},b}\{\gamma\} = \arg\max_{\vec{w},b}\{\frac{t_n y(\vec{x}_n)}{\|\vec{w}\|}\}$$

This optimization problem is too complex to solve, therefore we set the functional margin to unity

Set $\gamma = t_n(\vec{w}^T\varphi(\vec{x}_n + b)) = 1$

$\Rightarrow\ t_n(\vec{w}^T\varphi(\vec{x}_n) + b) \geq 1, n = 1, \dots, N$

The problem becomes

$$\arg\min_{\vec{w},b}\frac{1}{2}\|\vec{w}\|^2\text{, subject to}$$

$$t_n(\vec{w}^T\varphi(\vec{x}_n) + b) \geq 1$$

which is a (solvable) quadratic programming problem.

To solve this problem, we introduce Lagrange multipliers $a_n > 0$ such that

$$L(\mathbf{w}, b, \mathbf{a}) = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{n=1}^{N} a_n \left\{t_n(\mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}_n) + b) - 1\right\}$$

where $\vec{a} = (a_1, a_2, \dots, a_n)^T$

Taking derivatives, we have

$$\frac{dL}{d\vec{w}} = 0\ \Rightarrow\ w = \sum_{n=1}^{N} a_n t_n \varphi(\vec{x}_n)$$

$$\frac{dL}{db} = 0\ \Rightarrow\ 0 = \sum_{n=1}^{N} a_n t_n$$

Substituting $\vec{w}$, we have

$$\tilde{L}(\vec{a}) = \sum_{n=1}^{N} a_n - \frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N} a_n a_m t_n t_m k(\vec{x}_n, \vec{x}_m)$$

bject to Karush-Kuhn-Tucker condition (if it has a solution):

$$a_n \geq 0$$

$$t_n y(\vec{x}_n) - 1 \geq 0$$

$$a_n\{t_n y(\vec{x}_n) - 1\} = 0$$

To predict testing label,

$$y(\vec{x}) = \sum_{n=1}^{N} a_n t_n k(\vec{x}, \vec{x}_n) + b$$

Based on the kernel method we introduced previously, SVM can make prediction based on empirical input data (training samples). Applying certain kernel functions, SVM can also solve non-linearly separable problems.

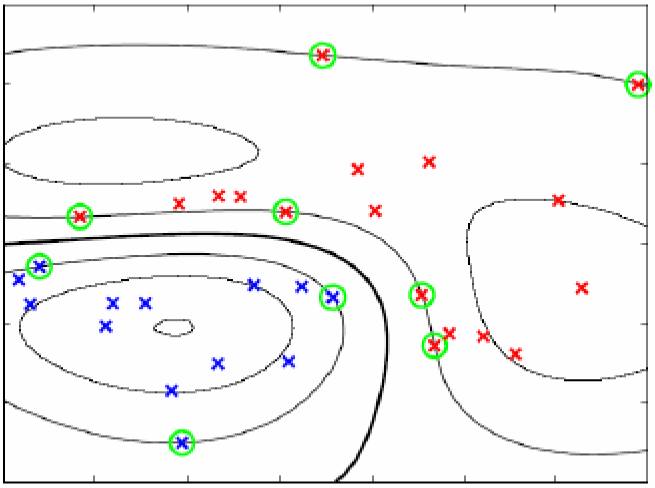


Figure 2.1.2 Non-linearly separable problem using a radial basis function (rbf) kernel. Source: Lecture notes from Machine Learning. Wang, 2011.

Extension to overlapping classes (Non-separable problems)

Since the class distributions are overlapping, the technique mentioned in previous section cannot be directly applied. Consequently, we introduce a slack variable $\xi_n \geq 0$ that allows sample points to be on the wrong side.

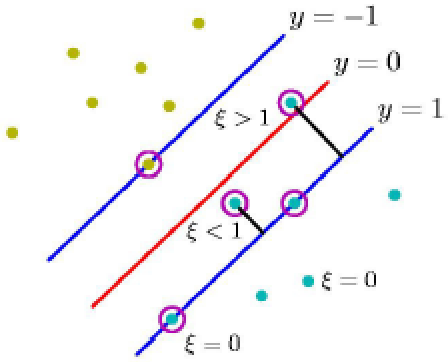


Figure 2.1.3 A simple example explaining how the slack variables work.

Source: Lecture notes from Machine Learning. Wang, 2011.

The introduction of slack variables changes the constraint to

$$t_n y(\vec{x}_n) \geq 1 - \xi_n, \qquad n = 1, \dots, N$$

Similar to the previous section, we again introduce additional Lagrange multipliers $\mu_n \geq 0$ to solve the optimization problem

$$L(\vec{w}, b, \vec{a}) = \frac{1}{2}\|\vec{w}\|^2 + C\sum_{n=1}^{1}\xi_n - \sum_{n=1}^{N} a_n\{t_n y(\vec{x}_n) - 1 + \xi_n\} - \sum_{n=1}^{N}\mu_n \xi_n$$

subject to KKT conditions if it has a solution

$a_n \geq 0$

$t_n y(\vec{x}_n) - 1 + \xi_n \geq 0$

$a_n\{t_n y(\vec{x}_n) - 1 + \xi_n\} = 0$

$\mu_n \geq 0$

$\xi_n \geq 0$

$\mu_n \xi_n = 0$

Taking derivatives

$$\frac{dL}{d\vec{w}} = 0 \Rightarrow w = \sum_{n=1}^{N} a_n t_n \varphi(\vec{x}_n)$$

$$\frac{dL}{db} = 0 \Rightarrow 0 = \sum_{n=1}^{N} a_n t_n$$

$$\frac{dL}{d\xi} = 0 \Rightarrow a_n = C - \mu_n$$

And, of course, by substituting **w**, we can obtain the dual representation. The introduction to the mathematics of standard SVM ends here. An alternative form of SVM, known as ν-SVM and introduced by Schölkopf et al., has the equivalent form of maximizing

$\tilde{L}(\vec{a}) = -\frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N} a_n a_m t_n t_m k(\vec{x}_n, \vec{x}_m)$, subjec to

$$0 \leq a_n \leq \frac{1}{N}$$

$$\sum_{n=1}^{N} a_n t_n = 0$$

$$\sum_{n=1}^{N} a_n \geq \nu$$

Schölkopf proved that if

1) The kernel is analytical.
2) The training samples are independent and identically distributed.

then the value ν is

1. An upper bound on the fraction of margin errors
2. An lower bound on the fraction of support vectors

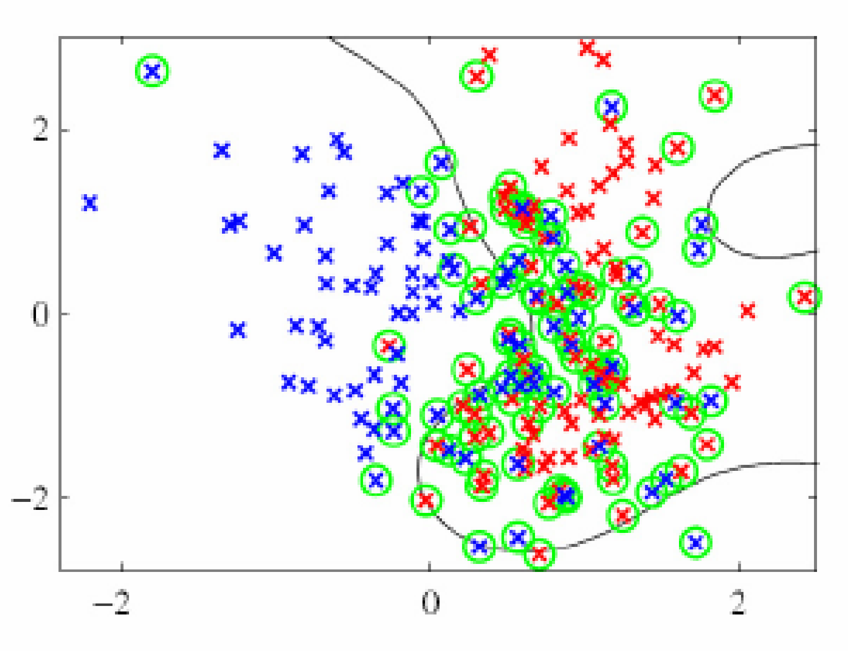


Figure 2.1.4 Illustration of SVM applied to a overlapping 2-dimensional data set. The support vectors are indicated by green circles.
Source: *Machine Learning and Pattern Recognition*. Bishop, 2006.

Philosophy of science in SVM

Vapnik's design of SVM is an instantiation of falsification in the philosophy of science. When it comes to defining the error function, SVM chose a hinge function which exonerates samples very far from the boundary, indicating that only the wrong samples matter. When computing the decision boundary, samples that are nearly wrong or misclassified become support vectors. Even in the theory of Vapnik-Chervonenkis dimension, we can see the falsifiability concept so clear (Vapnik, 2006).

Disadvantages

Originally devised for 2-class separable problems, SVM has some disadvantages when applying to non-separable multi-class problems.

1. It can only give hard decisions (binary outputs) instead of soft ones (probabilistic, numeric ones). In some fields of application (e.g. weather forecast), we prefer a probabilistic prediction rather than a clear-cut outcome.
2. Multi-class problems are theoretically unsolvable by binary classifiers. Commonly adopted schemes, including one-against-one (OAO) and one-against-all (OAA) decomposition of original multi-class problem, leave unresolvable areas where samples cannot be determined (an intuitive explanation is shown in Fig. 2.1.5; for comparison of multi-class SVM methods, cf. (Hsu and Lin, 2002)).
3. Misclassified samples are all supposed to be support vectors, so when classes overlap in the feature space, the amount of support vector increases. That is why highly non-separable problem makes the kernel very non-sparse, increasing training and testing time.
4. It is time-consuming to tune the hyperparameters. For example, for a regression problem that applies rbf kernel, we have to tune C (regularization term), ε(regression tolerance parameter), and γ (rbf parameter).There is no way to know advanced what value the hyperparameters might fall on. Grid search is most commonly suggested method to try parameter, but still, this strategy is not time-saving.

Despite all the above mentioned disadvantages, SVM is still widely adopted in our or other researcher’s experiments for its simplicity.

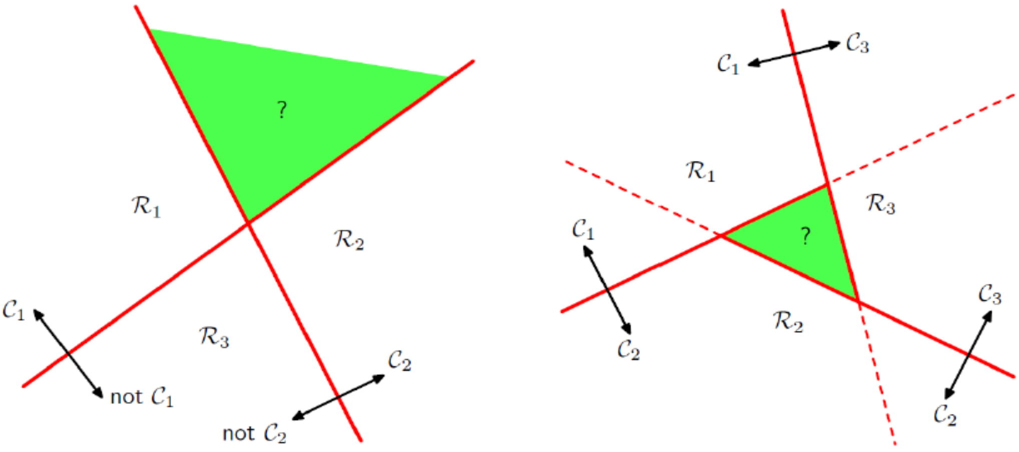


Figure 2.1.5: Binary classifier solving multi-class problem. Unresolvable areas

are shaded green.

Source: *Machine Learning and Pattern Recognition*. Bishop, 2006.

2.1.3 Imbalanced Datasets

Long existing in everyday applications, imbalanced datasets are a major and annoying issue in machine learning. In recent years, the imbalanced learning problem has drawn a significant amount of interest from academia, industry, and government funding agencies (He and Garcia, 2009). The fundamental issue with the imbalanced learning problem is the ability of imbalanced data to significantly compromise the performance of most standard learning algorithms. Most algorithms assume or expect balanced class distributions or equal misclassification costs. Therefore, when presented with complex imbalanced data sets, these algorithms fail to properly represent the distributive characteristics of the data and resultantly provide unfavorable accuracies across the classes of the data. In real-world domains, the imbalanced learning problem represents a recurring problem of high importance with wide-ranging implications, warranting increasing exploration.

There are three major approaches to handle data imbalance. The most intuitive type is sampling methods. Under-sampling the majority, over-sampling the minority, and synthetic sampling are the most popular ones. Cluster-based methods and data clearing methods such as Tomek link are usually applied as auxiliaries. Cost-sensitive methods are another main category dealing with imbalanced datasets. In some domains, cost-sensitive methods are even superior to sampling methods (He and Garcia, 2009). The other method, called kernel-based methods, is to calibrate learning algorithms themselves.

2.2 Auditory Model

The proposed auditory features were extracted from stages of an auditory model, which is based on physiological evidences and consists of early cochlear (ear) and central cortical (A1) modules.

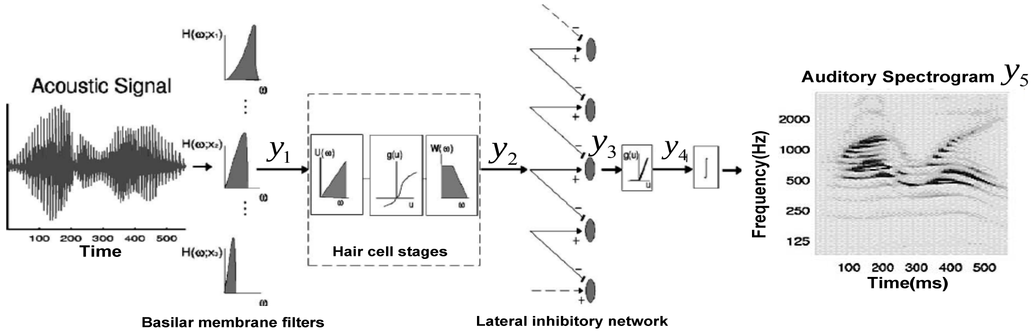


Figure 2.2.1 Detail block diagrams of auditory model (feature extractor).

2.2.1 Cochlear Module

The cochlear module models the functions of the peripheral auditory system. The cochlea behaves like a frequency analyzer. As Fig. 2.2.1 shows, the cochlear module consists of a bank of 128 overlapping asymmetric constant-Q band-pass filters ($Q_{3dB} \approx 4$) that mimic the frequency selectivity of the cochlea. These filters are distributed evenly 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 (a half-wave rectifier followed by a low-pass filter). The non-linear high-gain compression models the saturation of the inner hair cells, which transduce the vibrations of the basilar membrane along the cochlea into intracellular hair cell potentials. The auditory nerve then transmits the hair cell potentials to the cochlear nucleus of the central auditory system. This transmission is simulated by the LIN, which generates a spectral profile by detecting discontinuities along the frequency axis. This is followed by integration over a few milliseconds. This study uses a simplified linear version of this module with a disabled hair cell stage. This approach normalizes all speech signals in advance to avoid the non-linear high-gain compression of the hair cells. As in Fig. 2.2.1, the outputs at different stages of this module can be written as:

$$y_1(t,f) = s(t) *_t h(t,f) \qquad (1)$$

$$y_3(t,f) = \partial_f y_1(t,f) \qquad (2)$$

$$y_4(t,f) = \max(y_3(t,f), 0) \qquad (3)$$

$$y_5(t,f) = y_4(t,f) *_t \mu(t,\tau) \qquad (4)$$

where $s(t)$ is the input speech, $h(t,f)$ is the impulse response of the constant-Q cochlear filter with center frequency f, $*_t$ depicts the convolution in time, ∂_f is the partial derivative along the f axis, the integration window $\mu(t,\tau) = e^{-\frac{t}{\tau}} \cdot u(t)$ with the time constant τ models the current leakage along the neural pathway to the cochlear nucleus (midbrain), and $u(t)$ is the unit step function.

The output $y_5(t, f)$ is an auditory spectrogram that represents neuron activities along the time (t) and log-frequency (f) axis. The auditory spectrogram produced by this simplified linear cochlear module is similar to the magnitude response of a Mel-scaled FFT based spectrogram. The constant-Q criterion of the filter bank shares similar effects of the Mel-scale and the local envelope approximates the magnitude of a FFT based spectrogram. Note that the LIN accounts for the spectral masking effect provided that hair cells behave non-linearly. However, since this study does not consider the hair cell stage, the LIN only effectively sharpens the constant-Q cochlear filters.

2.2.2 Cortical Module and Rate-Scale Representation

The second module models the spectro-temporal selectivity of neurons in the auditory cortex (A1). The auditory spectrogram $y_5(t, f)$ is further analyzed (filtered) by cortical neurons, which are modeled by two-dimensional filters tuned to different spectro-temporal modulation parameters (Chi et al. 2005). The *rate* (or velocity) parameter ω (in Hz) reflects how fast the local spectro-temporal envelope varies along the temporal axis. The *scale* (or density) parameter Ω (in cycle/octave) represents the distribution of the local spectro-temporal envelope along the log-frequency axis. In addition to the rate and the scale, cortical neurons are also sensitive to the sweeping direction of the FM of the sound. This module characterizes directional selectivity using the sign of the rate: negative for upward sweeping direction, and positive for downward sweeping direction.

Therefore, the 4-dimensional output of this cortical module can be formulated as

$$r(t, f, \omega, \Omega) = y_5(t, f) *_{tf} STIR(t, f, \omega, \Omega) \qquad (5)$$

where $STIR(t, f, \omega, \Omega)$ is the joint two-dimensional spectro-temporal impulse response (STIR) of the direction-selective filter tuned to ω and Ω, and $*_{tf}$ is the two-dimensional convolution in the time and log-frequency domains. More detailed formulations and derivations of the $STIR(t, f, \omega, \Omega)$ are available in (Chi et al. 2005). The local energy of the four-dimensional output is then computed as

$$E(t, f, \omega, \Omega) = |r(t, f, \omega, \Omega) + jH[r(t, f, \omega, \Omega)]| \qquad (6)$$

where $H[\cdot]$ is the Hilbert transform along the log-frequency (f) axis. From a functional point of view, cortical neurons perform a joint spectro-temporal multi-resolution analysis (due to various rate-scale combinations) on the input auditory spectrogram. The excitation pattern of cortical neurons associated with a single time-frequency (T-F) unit at (t_j, f_j) of the input auditory spectrogram is referred to as the rate-scale (RS) representation of that particular T-F unit, and is expressed as $E(t, f, \omega, \Omega)$.

The frame-based RS representation of an utterance can be obtained by averaging the RS representations of T-F units over the frequency axis as follows:

$$P(\omega, \Omega, t_j) = \frac{1}{128}\sum_{j=1}^{128} E(t, f, \omega, \Omega) \qquad (7)$$

The bottom panels of Fig. 2.2.2 show the time-varying RS representation $P(\omega, \Omega, t_j)$ of a sample speech around 200 and 550 ms. Each plot of the RS representation clearly shows two attributes: (1) spectro-temporal modulations of envelopes and (2) resolved pitch below 512 Hz. Consider the 550 ms frame as an example. The resolved pitch around 230 Hz produces a strong response around the high rate high scale (pitch related) region. On the other hand, the envelopes of the almost flat harmonic structure shown at 230, 460, and 1150 Hz produces {low rate (due to the flatness, no FM), low scale (2 cycles/periods within 2.32 octave)} strong responses at regions less than 8 Hz and less than 1 cycle/octave. Since flat envelopes do not favor any sweeping directions, the {low rate, low scale} region exhibits symmetric rate responses. Figure 2.2.1 shows that the frame-based $P(\omega, \Omega, t_j)$ encodes the information of the spectral-temporal structures, including but not limited to pitch, harmonicity, formant spacing, and AM and FM of an input sound at each time instant. Some of these structures, such as pitch, AM, and 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) commonly appear in speech emotion recognition researches (Cowie et al. 2001; Mozziconacci 2002; Scherer 2003; New et al. 2003; Ververidis and Kotropoulos 2006; Schuller et al. 2007a; Busso et al. 2009). Therefore, the proposed time-varying RS representation could be a good candidate for speech emotion recognition.

The left and right panels of Fig 2.2.2 show the long-term averaged $P(\omega, \Omega, t_j)$ of clean speech and white noise, respectively. The long-term averaged RS representation of clean speech shown in the Figure 2.2.2 was produced by extracting 30 clean utterances from the NOIZEUS corpus (Loizou 2007). Clearly, the white noise primarily affects the pitch region (> 128 Hz) of speech. In addition to the pitch region, speech possesses high energies in the low-scale low-rate region (< 4 cycle/octave, < 32 Hz), while white noise activates the high-rate high-scale region (>

2 cycle/octave, > 32 Hz) due to differences in the structures of their spectral-temporal envelopes. This indicates that local spectro-temporal speech envelopes are mostly smoother than white noise envelopes along either the time or the frequency axis. These spectro-temporal envelopes critically encode the amplitude modulation and the frequency modulation of the sound, which are vital cues for humans to segregate individual sound streams from a sound mixture (Grimault et al. 2002; Carlyon et al. 2000). This segregation process of human hearing perception is very important to people's daily lives, and is referred to as auditory scene analysis (ASA) (Bregman 1990). Since speech envelope modulation is critical to hearing perception and vastly different from white noise envelope modulation, this study uses the time-varying $P(\omega, \Omega, t_j)$, which decomposes modulations of local envelopes in a multi-resolution fashion, to assess speech emotions under noisy conditions.

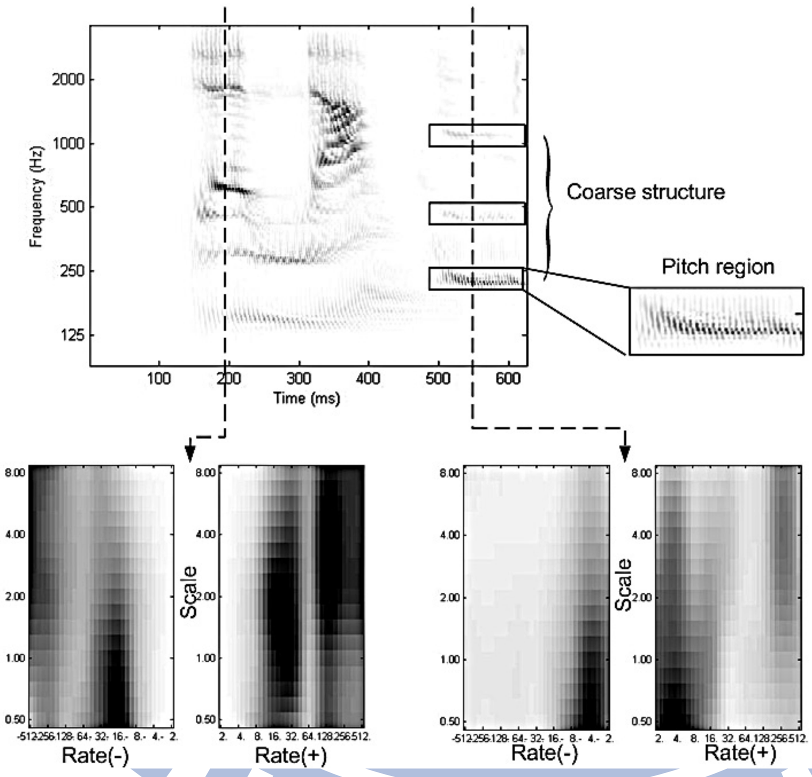


Figure 2.2.2 Rate-scale representation of a speech frame.

2.3 Emotion Psychology

The analysis of emotion has three main perspectives. Discrete theory, having probably the most ancient origins, describes emotions as discrete categories, so it is sometimes called the "Category theory." Dimensional theory, on the contrary, describes emotions as something in N-dimensional space, so it is also known as the "Continuous theory." Besides these two, some psychologists describe emotions as something that consists of smaller components.

2.3.1 Discrete model of Emotion

The discrete theory (a.k.a. categorical theory) of emotion holds that emotions are discrete categories just like how we call them. The number of emotion categories may be finite or infinite, but there is finite number of "basic emotions." Paul Ekman's Big Six (anger, disgust, fear, happiness, sadness, and surprise) has become the most well-known but not undoubted theory about basic emotions. For more complex emotions, mixture theory and compound theory are usually used. Plutchik's circumplex interprets emotions as mixtures formed by four pairs of basic emotion (joy-sadness, trust-disgust, fear-anger, and anticipation-surprise) like colors on a palette. However pretty the circumplex model might be, Reisenzein pointed out that high-class (complex) emotions can happen independently, indicating that the mixture theory might be falsified. Oatley and Johnson-Lairel proposed a compound theory claiming that high-class emotions are compounds of basic emotions. Still, this theory has some problem, including the fact that some basic emotions become reducible and that it cannot explain the inherent similarity between basic emotions (Hewstone et al., 2005). Today, discrete theory is still the most prevailing one that most emotion databases are labeled with categories.

2.3.2 Dimensional model of Emotion

The dimensional or continuous model describes emotions as N-dimensional clusters. Every emotion occupies a specific subspace in the N-dimensional space. Its history can be trace back to Wilhelm Wundt a hundred years ago, and its later follower includes Osgood and other researchers (Hewstone et al., 2005). In this model, **valence** and **activation** are two dimensions that are most mentioned. If there are three dimensions, **potency** (or power) will be added in. Recently, some more dimensions, like **unpredictability**, are introduced based on neurological studies (Fontaine, 2007).

Valence stands for intrinsic pleasure. If we feel good in an emotion, the emotion has positive valence, and it has negative valence if we feel the opposite. Activation represents the degree of psychological activity of the subject. If subjects know that they are in some emotion, the activation is high. Pereira pointed out that emotions with similar activation or potency are usually indistinguishable because they have similar acoustic features (Pereira, 1998).

2.3.3 Component Process Model

Component process model is an appraisal theory for vocal communication based on cognitive psychology (Scherer, 2005). It interprets emotion as a (temporal) series of components, each representing a pattern of variation of acoustic features. Every pattern has their psychological or physiological meaning, and the whole temporal process are all involved with cognition. Basically, the cognition process comprises five internal checks on the events or experiences a subject has been through. These five components are novelty check, intrinsic pleasure check, goal/need significance check, dealing potential check, and normative/spontaneous check. Although the theory might be sound, current applicable databases seldom contains the underlying element—context, thus limiting its applicability.

2.3.4 Brunswik’s lens model

Brunswik’s lens model is a counterpart explanation of the whole affective communication process (emotion recognition system) in psychology. This model has the same three stages as other communication architectures. Initially, emotions are encoded into speech utterances which have acoustic features like duration or intensity. Then the features go through a communication channel which might incorporate speech with noise. Finally, another perception subject receives the acoustic signal, maybe transform it into perception features, and decode the emotion. This model contains three important elements of emotion detection: ecological validity which describes the degree of modulation emotion has on speech, feature extraction which is the transformation from distal cues (acoustic features) to proximal cues (perception features), and recognition (classification).

Table 2.3.1 Synthetic compilation of the review of empirical data on acoustic pattering of basic emotions
Adapted from *Vocal Communication of Emotion* (Scherer, 2003)

	Joy	Anger	Stress	Fear	Boredom	Sad
Intensity	+	+	+	+		-
F0 floor/mean	+	+	+	+		-
F0 variability	+	+			-	-
F0 range	+	+		+ (-)	-	-
Sentence contours		-				-
High frequency energy	(+)	+		+		-
Speech and articulation rate	(+)	+		+	-	-

2.4 Miscellaneous

2.4.1 Cross-validation

Cross-Validation is a statistical method of evaluating and comparing learning algorithms (or models) by dividing data into two subsets: one used for training a model and the other used to validate it. This procedure is widely applied, especially when the database has too few samples so that generalization might be unreliable. Cross-validation procedure offers an estimator for performance, and of course brings some drawbacks.

There are many variant kinds of cross-validation schemes, all of which are variants of a prototypical form of “k-fold cross-validation.” In k-fold cross-validation the data is first partitioned into k equal-sized folds (subsets). Subsequently, k iterations of training and validation are performed such that within each, iteration a different fold of the data is held-out for validation while the remaining k-1 folds are used for learning. See an example in Figure 2.4.1. In addition, data is commonly stratified prior to being split into k folds. Stratification is the process of rearranging the data as to ensure each fold is a good representative of the whole.

Three common variants are also proposed: (See Table 2.4.1 for summary)

1. Hold-Out Validation

   Hold-out validation separates a database into two independent subsets: one for training and the other for testing. Independence between the training and the testing sets assures an accurate performance estimator, but the downside is that the procedure only uses parts of the whole data, so the estimator has large variance.

2. Leave-One-Out Cross-Validation

   Leave-one-out cross-validation (LOOCV) is an extreme case of k-fold cross-validation where k equals the number of samples. In each iteration, one sample is used for testing and all the others are used for training. LOOCV gives an almost unbiased estimator but it has very large variance.

3. Repeated K-Fold Cross-Validation

   Large number of estimates is always welcome for obtaining reliable performance estimation or comparison. In k-fold cross-validation, only k estimates are obtained. A commonly used method to increase the number of estimates is to run k-fold cross-validation multiple times. The data is reshuffled and re-stratified before each round.

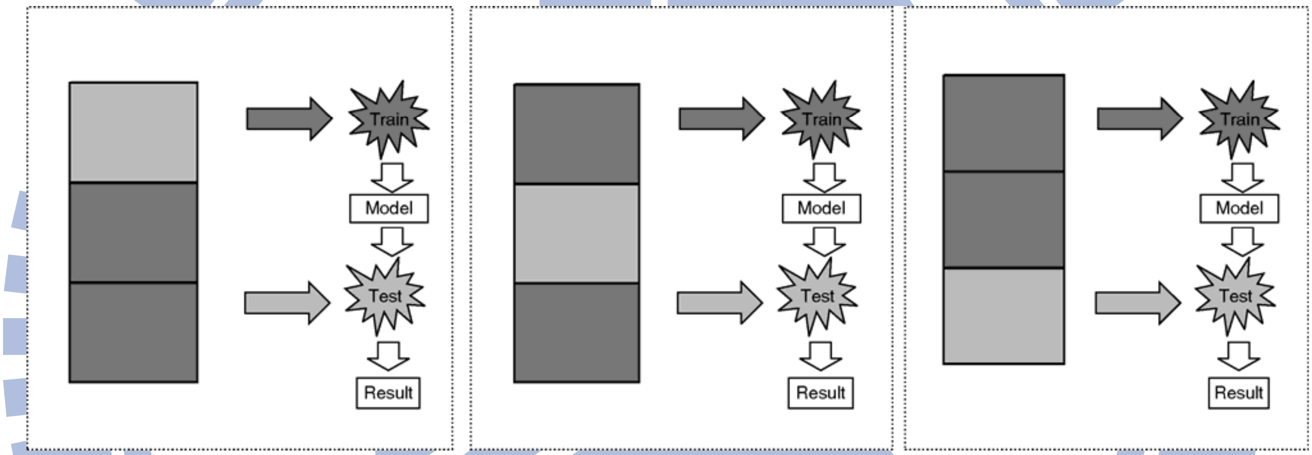


Figure 2.4.1: A graphical illustration of a 3-fold cross-validation.

Source: *Encyclopedia of Database Systems*, Springer, 2009.

Table 2.4.1: Pros and cons of different validation procedures.

Cross-validation method	Pros	Cons
Hold-out validation	Accurate performance estimator	Large variance
K-fold cross-validation	Accurate performance estimator	Underestimated variance
Leave-one-out cross-validation	Unbiased performance estimator	Very large variance
Repeated k-fold cross-validation	Reduced bias	Underestimated variance

In short, cross-validation schemes necessarily meet the problem of trade-off between the size of training and testing sets. Larger number of training samples (indicating smaller number of testing samples) brings smaller bias whereas larger number of testing samples results in smaller variance. Or in a plainer interpretation: too little training is prone to bias (preferring) and too little testing is like gambling (One data point and you're jumping for joy; loss of generalization).

Chapter 3

Methods and Materials

3.1 Database and Assessment Metrics

The databases used here is Berlin Emotional Speech Database (Burkhardt et al., 2005) and Aibo Emotion Corpus (Steidl, 2009). They differ from the types (acted or spontaneous) of emotions and recording condition. Information in detail can be found in Table 1.

3.1.1 Berlin Emotional Speech Database

This database comprises 535 German utterances of about 25 minutes in total. The corpus consists of 7 emotions simulated by 10 speakers. The recordings are studio-quality, and all the sentences have 80% or higher human recognition rate. Previous research showed that state-of-the-art automatic emotion recognition schemes achieve over 80% accuracy in best-case scenario (Kockmann et al., 2011).

3.1.2 Aibo Emotion Corpus

This database comprises 18,216 chunks of about 8.9 hours of spontaneous emotion. It consists of 5 (originally 11) emotions of 51 children, aged from 10-13. The emotions were elicited in an experiment in which children thought the Aibo robot dogs obeying their commands. The recordings are medium quality; some samples have serious microphone clicking noise, others are clipped due to loudness, and the others include coughing. The corpus was used in INTERSPEECH 2009 Emotion Challenge and the best performance reached 41.7% of unweighted recall rate (UR) and the combination of all participants reached 44.0% of UR.

Table 3.1.1: Details about AEC and BES

Database	FAU AIBO Emotion Corpus	Berlin Emotional Speech Database
Language	German	German
Emotion Info.		
Type	Spontaneous	Acted
Class	2: Idle, Negative 5: AENPR	7: ABDFHNS
# of labeler	5	20
Speaker Info.		
Gender	Male: 21 Female: 30	Male: 5 Female: 5
Age	10-13	21-35
Utterance Info		
# of Utterance	18216	535
Sampling rate	16kHz	16kHz
Duration	0.1-24.5 seconds	1.2-9.0 seconds
Validity	0-100%	80-100%
Quality	Normal (some with severe microphone click noise)	Recording studio

*A: Anger. B: Boredom. D: Disgust. E: Emphatic. F: Fear. H: Happy. N: Neutral. P: Positive. S: Sadness. R: the rest.

Challenges in AEC database

Compared to BES which contains acted emotions, Aibo Emotion Corpus (**AEC**) is more challengeable for three reasons other than the fact that it contains spontaneous emotions:

First, the samples have rather low prototypicality (later called “validity” in this article; the usage of prototypicality is left for emotion classes), meaning that even human labelers have difficulty recognizing the emotion. Low prototypicality can be seen as strong **internal noise** that undermines classification performance either for machines or for humans. The macro effect is high intra-class variance and that the problem becomes highly non-separable.

Second, among the 5 classes, the **Rest** is an undefined class; it has no specific traits and even no intra-class similarity. They become an apparent source of disturbance and it makes the problem more difficult according to the following reasons:

(1) Their intrinsic uncertainty makes them similar to other samples from classes.

(2) They distribute in a sparse manner over the whole feature space, appearing little correlation to other inter- or intra-class samples.

(3) The Rest is a class without phenomenal traits; the samples just do not belong to other categories.

(4) The confusion matrix gets bigger, thus raising complexity and error rate.

(5) The assessment metrics might fail to reflect a fair situation because of this internally noisy class.

Third, the 5 classes in AEC are severely imbalanced. The problem is common even in everyday situations but disturbing and when the imbalance is severe, taking accuracy (or weighted recall) as the performance metric is no longer feasible; therefore this paper adopted UR as the main assessment metric. Besides, imbalanced data cause classifiers learn to prefer majority class. Actually, the imbalance is two-fold, because

(1) The numbers of samples of each class are imbalanced

(2) The validity (percentage that human labelers came to consensus) of each class has different distribution. Some classes have very high validity samples while others do not.

Unfortunately, the biggest and the most valid class are the same one—Neutral. Neutral, which has samples 8.9 times more than Positive (the minority), has 83.9% (the highest) validity in average while Positive has only 58.3%. Things deteriorated under this situation because usual applicable techniques can make imbalance worse. While synthetic method creates more samples for minority class, it also creates more unreliable (low validity) samples. Under-sampling methods which cut back on samples in the majority class might cause information loss. Cost methods, on the other hand, have same issues as synthetic methods because they give reliable samples high cost as well as unreliable ones.

In our experiments, data are rebalanced by a mechanism similar to the SMOTE with different costs (Chawla et al., 2002; Akbani et al., 2004). This method utilizes validity as a reference in synthesizing procedure, so it might allegedly avoid some of the problems described above.

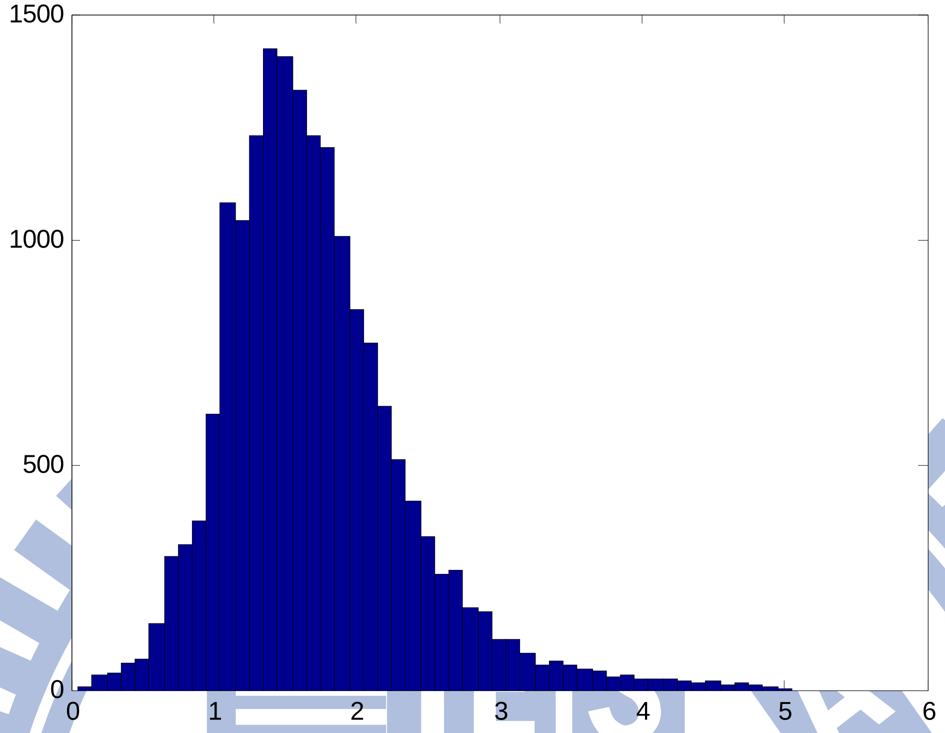


Figure 3.1.1: Histogram of the distribution of duration of the wav file in Aibo Corpus.

*The x-axis represents duration (seconds) and the y-axis represents the number of samples.

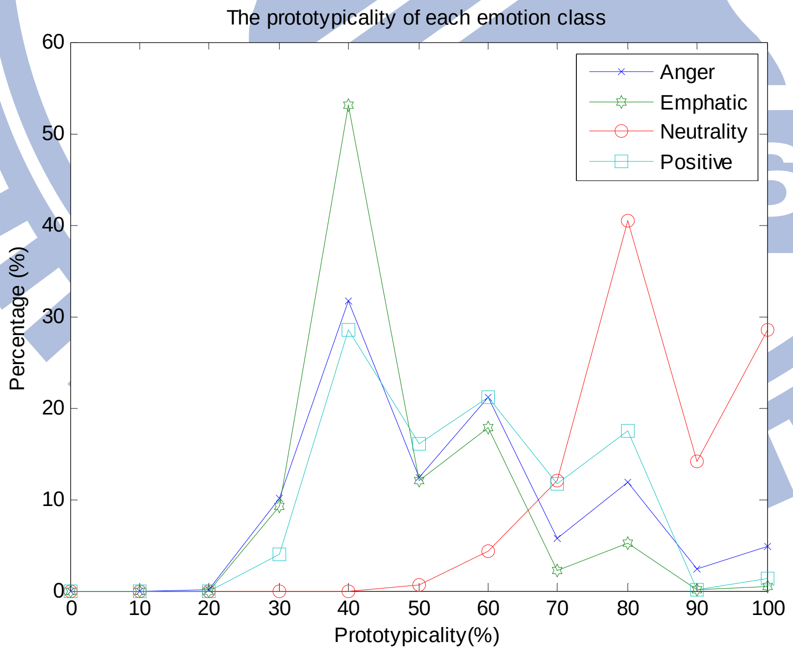


Figure 3.1.2 Distributions of validity of each class.

*Almost all REST samples have zero validity because they do not belong to any other classes.

Table 3.1.2: Duration of wav files in Aibo Corpus

Duration (sec)	Number of files
<1	3183
1-2	11687
2-3	4007
3-4	569
4-5	162
>5	105

3.1.3 Assessment metrics

The assessment metrics, even the terminology of assessment metric, seem to have no consensus yet. In the case of imbalanced datasets, commonly used metrics might even fail to give truthful information. Although in binary classification tasks, the receiver operating characteristic (ROC) or its equivalent form, the precision-recall curve (PRC), is viewed as the most informative metric, its use remains limited (cf. Davis and Goadrich, 2006). The area under the curve (AUC) of ROC or PRC has the same limitation. For hard-type classifiers which output discrete label only (e.g. SVM), none of the above mentioned metrics can be used (because they need soft decision or probability estimate), let alone to be used in multi-class problems ().

Accuracy, or the weighted recall rate (WR), tends to be biased in the case of imbalanced data. Sometimes the geometric or harmonic mean of WR and UR (also known as G-mean and F-measure) is used to represent one overall performance. In this article, performance is evaluated by UR with WR as auxiliary metric. Some common assessment metrics and their definition are listed below:

$$\mathrm{recall}_i = \frac{\sum_{n=1}^{N} I\{t_n = i \cap y_n = i\}}{\sum_{n=1}^{N} I\{t_n = i\}}$$

$$\mathrm{precision}_i = \frac{\sum_{n=1}^{N} I\{t_n = i \cap y_n = i\}}{\sum_{n=1}^{N} I\{y_n = i\}}$$

$$\text{Arithmetic mean: } A \triangleq \frac{1}{K}\sum_{i=1}^{K} Z_i$$

$$\text{Geometric mean: } G \triangleq \sqrt[K]{\prod_{i=1}^{K} Z_i}$$

$$\text{Harmonic mean: } \mathrm{F} = \frac{\mathrm{K}}{\sum_{\mathrm{i}=1}^{\mathrm{K}} \frac{1}{\mathrm{Z_i}}}$$

$$\text{Weighted mean: } \mathrm{W} = \sum_{\mathrm{i}=1}^{\mathrm{K}} \mathrm{weight_i} \times \mathrm{Z_i}$$

$$\mathrm{I}\{\cdot\}\text{: indicator function. } \mathrm{I} = \begin{cases} 1, \text{if the statement is true} \\ 0, \text{if the statement is false} \end{cases}$$

$\mathrm{y_n}$: prediction lable of the nth test sample (classification output)

$\mathrm{t_n}$: class lable of the nth test sample (truth/answer)

i: class index, $\mathrm{i} = 1, \ldots, \mathrm{K}$

n: sample index, $\mathrm{n} = 1, \ldots, \mathrm{N}$

$\mathrm{Z_i}$'s represent recall or precision or both.

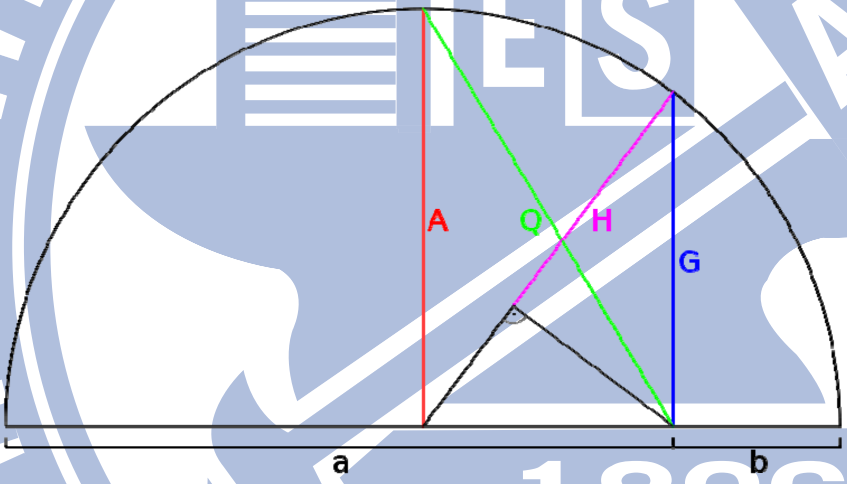


Figure 3.1.3 Geometric illustration of arithmetic, geometric, and harmonic mean.

*Q denotes the quadratic mean (a.k.a. root mean square)

Synonyms

Unweighted recall and unweighted average of class-wise recall rate mean the same thing. Weighted recall or weighted average of class-wise recall rate is also known as **accuracy** where the weight is the fraction of the class. Useful combinations include the A, G, and F of recall rate, precision, or both (only applied to binary classification). The geometric mean is sometimes called the G-mean and the harmonic mean is sometimes called F-measure.

Pros and cons

Weighted recall (WR) makes unfair evaluation of performance when encountering highly imbalanced datasets. In the case of Aibo Corpus, classifying all test samples to Neutral gets 65.12% WR while this classification is useless as well as meaningless because it is not better even than guessing. Unweighted recall (UR) gives a fairer view, but still, it does not reflect the specific performance the classifier has on a specific class. For example, we have a classifier whose recall rate for a 5-class problem are 100%, 50%, 50%, 50%, and 0%, then the UR is 50%. However, we notice that the classifier cannot recognize any sample of the 5th class. On the other hand, under this circumstance, the geometric mean (GR) is 0%, pointing out that one of the classes failed to be recognized. Our conclusion is that there is no one single metric that reflects overall performance, so we had better rely on multiple metrics.

3.2 Features

Features are always the most fundamental issue in emotion or affect research. Adequate feature set consolidate the foundation of recognition. Large scale searching for features was one of the major attempts in last decade. For now, some features have been identified as basic elements. This section briefs two feature sets used in our experiments.

3.2.1 Rate-scale features

The process described in Chapter 2 resulted in four-dimensional (time, frequency, rate, scale) information. First, the information was downsized to 3-dimensional by taking average along frequency. Then we took temporal mean and standard deviation along time dimension. Finally, there were only 2-D information, namely rate and scale, left. The range of rate is from 2^1 to 2^9 and -2^1 to -2^9, and the range of scale is from 2^{-1} to 2^3. Therefore, the total number of features is 2×(# of rates)×(# of scales) ×(statistics) = 2×9×5×2 = 180. Since every utterance is transformed into this rate-scale (RS) representation, conceptually an utterance of any length is represented by the 180 features, namely a 180-entry vector. This provides us a way to visualize an utterance.

Now it is time to explain the physical meaning of every RS feature. Speech manifests itself in both high rate and low rate region. The low rate part carries information about speech rate whereas the high rate part carries the information about pitch. Scale on the other hand, carries information about formants. Different types of noise give rise to the impact on different RS regions. Babble noise inflicts more alternation on low rate region while white noise does (See Figure 3.2.1).

3.2.2 INTERSPEECH 384 features

The 384 features used in 2009 INTERSPEECH Emotion Challenge is the main comparison target in the following experiments. This feature set comprises 16 low-level descriptors and their first-order derivatives and their 12 functionals. The 16 low-level features 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 with HTK-based computation. Each of these 16 features include the delta (time derivative) coefficients. Next, 12 functionals are derived for each low-level and its delta feature on a chunk basis: mean, standard deviation, kurtosis, skewness, minimum and maximum value, relative position, range, and two linear regression coefficients with their mean square errors (MSEs);. Thus, the final feature set contains 384 (16×2×12) attributes and is referred to as the **i384** feature set in the following simulations.

The feature extraction module is an open source toolkit named OpenSMILE (Eyben et al. 2009). Details of this feature set can be found in Table 2. This set of features includes the most common ones associated with prosody, spectral shape, voice quality, and their derivatives. The i384 features contains the most commonly used features in emotion-related research and this is why it is chosen to be our main comparison target.

Table 2. Features used in INTERSPEECH 2009 Emotion Challenge

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

3.3 System Framework and the Paradigms

3.3.1 System overview

As shown in Figure 3.31, a speech sample went through three stages in order: the preprocessing, feature extraction, and classifier stage. In the preprocessing stage, a speech sample was first down-sampled to 8 kHz to meet some requirements in the auditory module. Then the voice activity detection (VAD) module marked voice regions and computed the speech power to which loudness was adjusted to -20dB according. Noise with different types and SNR was added into the speech sample. In the second stage, the auditory module extracted the RS features. In the final stage, clean speech samples were used to train one single SVM classifier and noisy samples were used to test it.

Figure 3.3.1: System Block Diagram

3.3.2 Consistency between Training and Testing

The consistency between training and testing conditions represents the underlying paradigm an experiment adopts. Performance of feature sets differs from one another under different conditions. Here we introduce some consistency criterions that are used in our or other researchers work.

Matched condition

Matched condition between training and testing implies that testing conditions such as a specific noise type or SNR are known advanced in training phase. Under this paradigm, researchers collect or synthesize training and testing samples that have the same condition, say same SNR, for learning machines. Having same conditions implies same (or approximately same) probability distribution of samples, so the learning machine usually has better performance. This paradigm also allows different feature sets when dealing with different conditions, e.g. feature set A for 10dB condition and set B for -5dB noisy condition.

Widely adopted in previous work, matched conditions are popular when it comes to the discussion of noise or interference (cf. Schuller et al., 2006; You et al., 2006). It is the easiest way to counter the effect of noise or interference and the results are usually good enough. However, this paradigm sometimes goes ad hoc and loses generality, and it also requires the learning machine to be very powerful.

Here is an example: We need to detect a signal under a specific type of noise in q dB condition where $q \in \{\infty, 10, 0, -10\}$. We are able to access noise samples, so without analyzing the characteristics of the noise, we train the learning machine with synthetic noisy signal under different SNR conditions. After comparing different sets of features, we finally determine that different feature sets be used in different SNR conditions (Table 3.3.1 exhibits more details).

Table 3.3.1: An example illustrating the degree of freedom matched condition has.

Starting resources before training

- Pure signal samples
- Pure noise samples

Training phase

- Synthetic noisy samples of $\left\{\begin{matrix}\infty\\10\\0\\-10\end{matrix}\right\}$ dB SNR

Training configuration:

SNR condition (dB)	Feature set	Learning Machine
∞	Feature A	Machine A
10	Feature B	Machine B
0	Feature C	Machine C
-10	Feature D	Machine D

*Different features set might require different learning machines (algorithms, theoretical analysis, or strategies).

Testing phase

- Actual/Synthetic noisy samples of $\left\{\begin{matrix}\infty\\10\\0\\-10\end{matrix}\right\}$ dB SNR

Pros and cons

As mentioned earlier, matched condition is an easy way to take into consideration the issues that will be met later in testing phase. This paradigm has the following assumptions:

Assumption 0 (consistency assumption):

The conditions of training phase are the same as that of testing phase. This is equivalent to having training and testing sample from the same distributions.

Assumption 1 (foresight assumption):

The learning machine knows that it will have to face a **specific** condition different from current empirical resources (e.g. pure signal samples in the previous example).

Assumption 2 (access assumption):

Direct access to the interference or noise is available (noise samples are obtainable).

*Learning in consistency assumption requires both foresight and access assumptions; otherwise, the learning might be meaningless or it might be not better than guessing.

Although performance under matched condition is good enough in current literature, it actually relies on some strong assumptions. Removing one of the assumptions above disqualifies the application of matched conditions.

Mismatched conditions

Contrary to matched condition, mismatched condition implies mismatch between training and testing conditions. This requires stronger ability of generalization or more consideration in the training phase. Mismatched conditions can be further divided into mainly two categories, so we separate the discussions.

Slack mismatched conditions

Slack mismatched condition utilizes more prior knowledge that cannot be acquired from empirical data. This paradigm retains the foresight assumption, so it is insufficient to run a matched condition experiment. In this case, the learning machine expects the discrepancy between training and testing data, so it has to get prepared for it. Here, preparation can of several forms. It can be *prior knowledge* from which the machine knows how noise or interference affects feature sets; it can also be strategy to recognize the problem (e.g. noise type and SNR) before classification (so that it can process noisy samples beforehand). For example, if we have no access to noise samples or noisy signal samples but the statistics or properties of noise is known, we can take care of it in the training phase.

There is one thing different in Yeh's and my paradigms though we both adopted mismatched conditions. Yeh adopted an even looser form of slack mismatched condition, called **clarity condition** here, that allows utilization of information from testing data. For example, in the experiments of Berlin database, she normalized the whole database in advance by subtracting sample mean and dividing sample standard deviation (we will come to the discussion about the issues of normalization later). This might seem minor at first glance, but it implies shared information between training and testing sets; or to be more precise, the machine uses information from testing data in the training process. In such a case, irrationality is not undoubtable.

Strict mismatched conditions

Strict mismatched condition is a paradigm that insists on using no knowledge other than given empirical data. As a result, the learning machine requires robust features to deal with the changes of conditions. For example, if the distribution of a feature is insensitive to the change of conditions, the feature will perform at about the same level.

Database

Current available databases of emotion usually have no realistic noisy samples (speech in specific SNR conditions). Although some databases (e.g. SUSAS) contain some noisy samples, it is far from sufficient. This is why most researchers do with synthetic noisy samples in training and testing phases.

Table 3.3.2: Comparison of assumptions of matched and mismatched conditions.

Condition	Assumption 0 (Consistency)	Assumption 1 (Foresight)	Assumption 2 (Access)
Matched	★	★	★
Slack Mismatched	╳	★	△
Strict Mismatched	╳	△	△

*★: necessary. △: not important. ╳: assumption not hold.

*Note that consistency assumption is a sufficient condition of foresight condition and access condition and it is not vice versa. The union of foresight and access assumption is a necessary but not sufficient condition of consistency. Consistency between training and testing is one of the many strategies for leaning machines.

Figure 3.3.2: Visualized illustration of the difference between matched and mismatched conditions.

3.3.3 Normalization Schemes

Numerical normalization for classifier is usually a necessary procedure for machine learning algorithms. Very large or very small scale of numbers cause problems especially when there is no closed-form solution. In addition, normalization might bring a sparser data, which shorten classification time. There are many normalization schemes for classification (cf. Friedrichs and Igel, 2005). While linear transformations are most frequently used, nonlinear warpings may also be employed in some cases. Here we discuss two types of linear normalization schemes in our experiments. Note that the following problems might occur only under our paradigm because we are dealing with mismatched conditions.

Nrm01

LibSVM suggests its user normalize their training data to [0, 1] because it makes data sparser. This scheme, denoted by **nrm01**, has risks when the training set has outliers that are too large in absolute value. When the dimension is high, the risk raises. Actually, under ordinary situation, this is not a problem; it becomes serious when the testing set mismatches the training set, i.e. they obey different probability distributions.

Nrmuv

The other normalization scheme used in our experiments is by subtracting sample mean and then normalizing sample standard deviation. This scheme, denoted by **nrmuv**, achieves about the same performance as that of nrm01 but it usually takes a bit longer for training and testing.

Pros and cons

Both of the methods do not consider the correlation between features; i.e. they both assume that features are independent or that the dimensions each feature spans are orthogonal. This assumption has very high risks, because it does not usually hold. Taking RS features for example, values of adjacent rates or scales are usually highly correlated. This negligence partially explains why performance degenerates in slack condition.

A maximum value or a minimum value is a biased estimator with high variance. Compared to sample mean or sample variance, normalization by extreme values can be risky because they have more uncertainty. For example, it might result in different best SVM parameters when we conduct repeated k-fold cross-validation, making best parameters hard to decide.

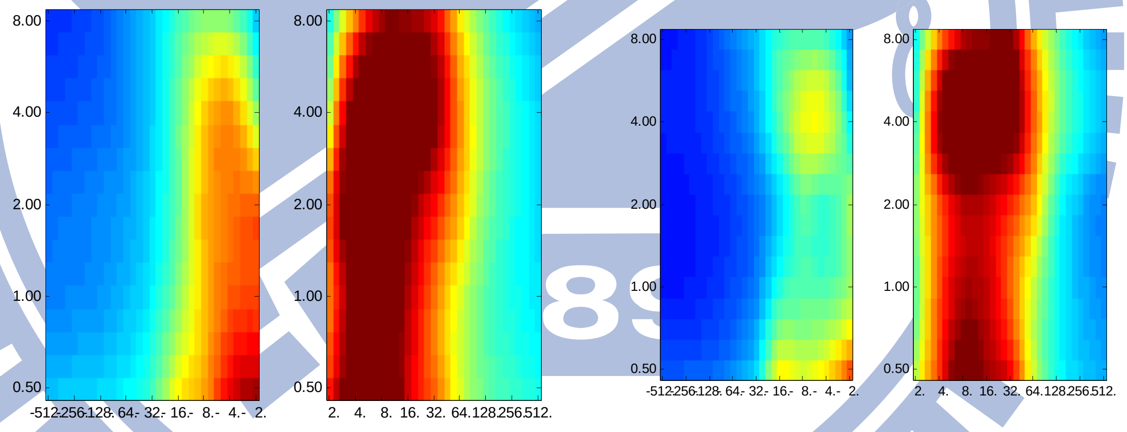
3.3.4 Voice Activity Detection

A voice activity detection (VAD) technique is recommended when the case is Aibo Emotion Corpus, which comprises microphone noise and very long period of silence. Note that the problem is not even manifest in Berlin Database because it was recorded in a studio without noise and with proper process of silence duration. As far as RS feature set is concerned, the two problems are almost a nightmare. If most of a wave file is silence, the characteristics of speech will lose its importance. If microphone noise is very serious, then the high-rate regions will be compromised. Therefore, VAD is required in analyzing Aibo Corpus.

First of all, an energy-based VAD does not serve our purpose because it cannot exclude microphone noise. There are two types of microphone noise. One is clipping, which is caused by microphone setting the value to zero when it detects very high energy. The other is clicking, which happened when the participant hit the microphone.

Second, a VAD based on RS features might be feasible because RS features of clicking and speech are very different in the feature space. Figure 3.3.2 shows no signs of pitch when the case is clicking noise; this gives a clue to separate speech from noise. However, it still needs more consideration and simplification. For one thing, extracting RS features takes too much time compared to energy-based methods; for another, it does need so many features (180 dimensions) to detect voice activity. The lack of response in pitch-related areas implicates that pitch is a good separator of speech, clicking, clipping, and, of course, silence. Therefore, we finally adopt a pitch-based VAD scheme.

(a)

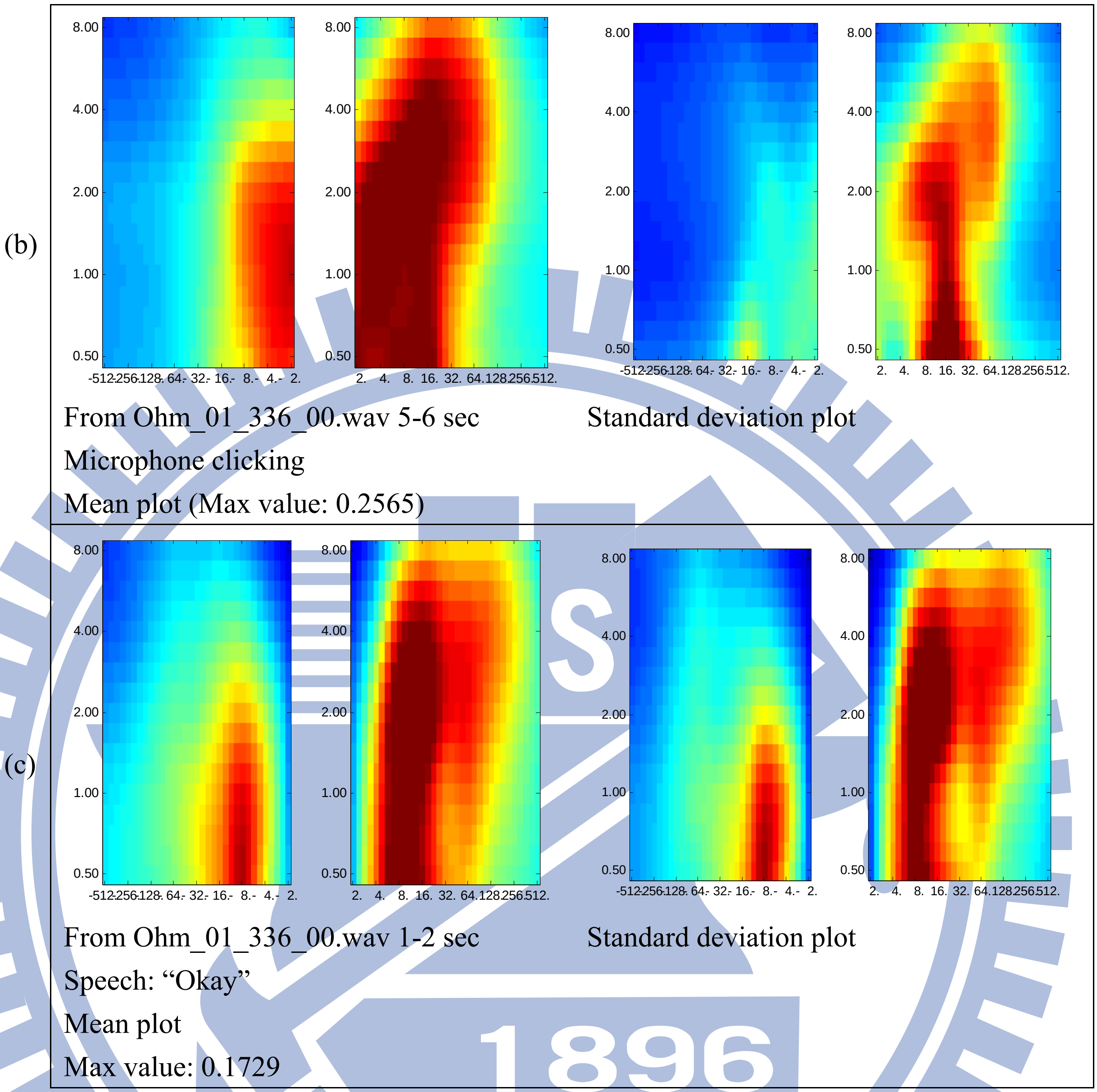


From Ohm_01_336_00.wav 5.5-7 sec

Microphone clicking

Mean plot (Max value: 0.4194)

Standard deviation plot



From Ohm_01_336_00.wav 5-6 sec
Microphone clicking
Mean plot (Max value: 0.2565)

Standard deviation plot

From Ohm_01_336_00.wav 1-2 sec
Speech: "Okay"
Mean plot
Max value: 0.1729

Standard deviation plot

Figure 3.3.2 Samples showing the difference and similarity between a speech and a clicking sound. The microphone clicking in (a) does not look like speech at all because it shows no signs of pitch and the pattern of standard deviation has nothing similar to that a speech sample. The microphone clicking in (b) has a quasi-pitch response in RS_{sd} plot, but it also contains peculiar response in low-rate-low-scale regions (triangular shape centered at rate=16).

The VAD is implemented by an aperiodicity method called STRAIGHT (Kawahara et al., 2005). Its computation time is about the duration of the original file multiplies three. The VAD by aperiodicity, named VADA in the rest of this article, has three phases. In the first phase, STRAIGHT calculates the spectrogram and its aperiodicity; it then computes the fundamental frequency (F0) according to the aperiodicity of every frame (1 millisecond per frame). In the second phase, if F0 is larger than 600 Hz, the algorithm sets F0 to zero because STRAIGHT usually decides

F0 up to 600 Hz for clicking noise. Then the algorithm identifies active regions longer than 75 ms and eliminates regions shorter than 75 ms (by setting F0 to 0). In the final phase, it encapsulates those speech chunks in the second phase and leave 500ms to the front and end of the encapsulation.

The 75 ms duration minimum was to eliminate microphone clicking and to keep short vowels. Short vowels can have durations as short as 55 ms; for instance, the duration of "e" of "Yezt" in Mont_01_072_00.wav has only 60 ms. But if the duration minimum is too low, it allows too many clicking frames to survive (clicking might sometimes has F0 lower than 600 so that it would not be eliminated in the second phase).

Rewards and risks on applying VAD

Applying VADA helps to identify speech region in wav files and cuts back on the time spent on feature extraction. The original Aibo Corpus contains 8.9 hours of recording but after VADA, only less than 7 hours of speech content left. The main benefit comes from the reduction of the duration of corpus since feature extraction of RS takes huge amount of time (original duration multiplies 24).

Generally speaking, applying VAD is helpful when there are many wav files that contain long duration of silence or equivalently when the fraction of the duration of speech is low. The downside is that the original long-term property is altered after applying VAD. Actual effect is inconclusive in our experiments. Another issue is its robustness. In the presence of serious noise, most VAD's cannot detect anything. In our experiments, we assume that VAD techniques are perfect even at the presence of serious additive noise. Or equivalently, the experiments were conducted on the assumption that clean version all noisy samples were known (this is taken for granted in our or other researchers' experiments because contemporary studies only focus on synthetic noise; in practical use, the clean version of a noisy sample is unknown). This is not an ad hoc assumption because we already declare in Chapter 1 that anything prevents features from showing their potentials should be ruled out; on the other hand, in practice, without VADA, the performance does not change significantly. The assumption only helps us to cut down on the consumption of time.

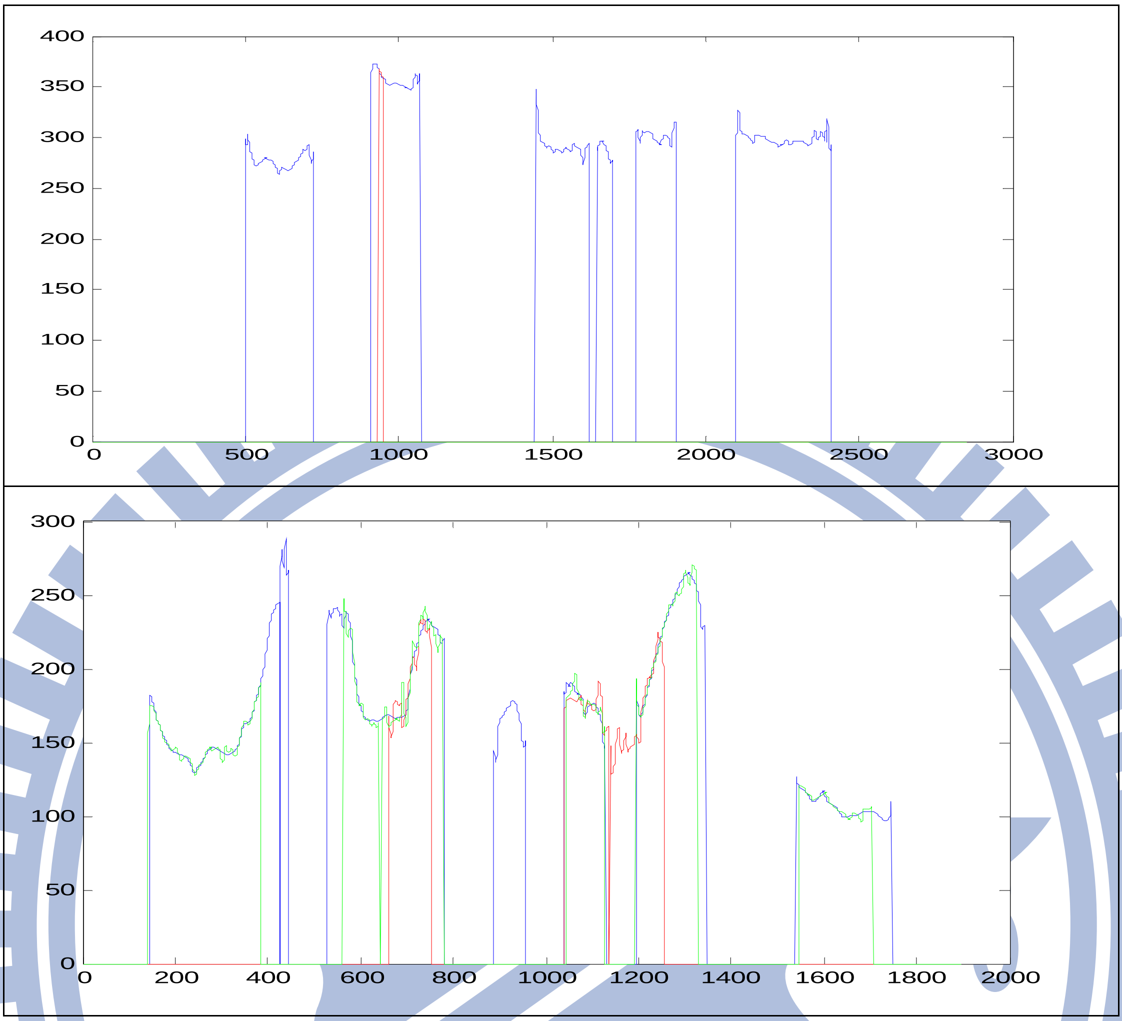


Figure 3.3.3: The result of STRAIGHT applied on (a) Mont_01_034_00.wav and (b) 03a01Fa.wav in clean speech (blue), 0dB white noise (green), and 0dB babble noise (red).

3.4 Rebalancing Imbalanced Dataset

3.4.1 Synthetic Minority Oversampling Technique with Validity

As mentioned earlier, data imbalance is one of the three major issues of Aibo Corpus. A simple over- or under-sampling method is not good enough. A simple over-sampling method increases both training and testing time while we are using a (sparse) kernel-based classifier and it can be simply replaced by a simple cost-sensitive method. A simple under-sampling method, on the other hand, loses too much information. The trouble caused by removing samples is two-fold (as describe in Section 3.1.2) because the most of the majority samples are reliable and most of the minority samples are unreliable. The analysis above leads us to a conclusion: we can adopt a cost-sensitive method or a finer synthetic sampling method.

The Synthetic Minority Over-sampling Technique (SMOTE) partially meets our needs because it still synthesizes too many samples, which is costly to sparse kernel machines. And we noticed that SMOTE synthesizes new samples uniformly with an aim in mind that it does not attempt to change the probability distribution. Nonetheless, in emotion recognition, we have a crucial clue—validity which in other applications is not necessarily given or acquirable. Validity is a measure of how much credibility can be put on a sample. If a sample is unreliable, it does not deserve more derivatives (synthetic samples). Taking validity into consideration, we modified the original SMOTE and name it SMOTEV (SMOTE with validity).

Formulation

The original SMOTE selects a target sample $\vec{x}_i$ and its k-nearest neighbors $\vec{x}_j$'s (kNN) and synthesize k samples on the midway of any pairs $(\vec{x}_i, \vec{x}_j)$ by

$$\vec{x}_{synthesis} = \vec{x}_i + \alpha_{ij}\left(\vec{x}_j - \vec{x}_i\right), \text{where } \alpha \sim U(0,1).$$

It has two shortcomings. First, it needs to compute the distance matrix in order to find kNN's. Second, it may synthesize many unreliable samples. Increasing unreliable samples prolonged training and testing time, and it might make learning more biased (increasing amount of falsification).

To tackle the distance matrix, SMOTEV made another attempt. It first selects samples with 80% or higher validity to form a reference set. Next, all, except the samples with 20% or lower validity, become candidates of target sample. New samples are synthesized on a random position along the line of a reference sample and a target. Taking validity into consideration, we can formulate SMOTEV as

$$\vec{x}_{synthesis} = \vec{x}_i + \beta_{ir}(\vec{x}_r - \vec{x}_i) = (1 - \beta_{ir})\vec{x}_i + \beta_{ir}\vec{x}_r$$

$\beta_{ir} \sim \mathrm{Beta}(1 + v_r, 1 + v_i)$

r: the index of a randomly selected sample in the reference subset.

$\vec{x}_r$: reference sample

$\vec{x}_i$: target sample

v_r: validity of $\vec{x}_r$

v_i: validity of $\vec{x}_i$

The random variable β_{ir} has mode of $\frac{v_r}{v_r+v_i} \in [0, 1]$. The skewed mode value reflects the fact that more credibility is given on the reference sample. Note that SMOTEV, just as SMOTE, cannot apply to nominal or ordinary scales (no technical problems, but the meaning might become nonsense); it only can apply to interval or ratio scales.

The final solution to imbalanced datasets in our experiment was to combine SMOTEV with different cost for each class. Setting different costs saves time and it has similar performance to that of sampling methods. Comparison of five different rebalancing schemes is shown in Table 3.4.1. The SVMs inherently bias toward majority classes since they aim to minimize total error; therefore, training without rebalancing is definitely infeasible.

Aided by validity, SMOTEV conceptually synthesizes new samples that may be more reliable than the original SMOTE. The two synthetic methods have similar performance. However, all synthetic methods unavoidably increase training time. In order to reduce the trouble, setting different cost seems to be the best way. Therefore, we only synthesize part of the data, and setting different costs afterwards.

Table 3.4.1

				No rebalancing					
	A	E	N	P	R	Recall	Precision	UR	UP
A	194	74	333	8	2	31.75%	32.61%	31.10%	35.58%
E	86	337	1069	13	3	22.35%	51.93%	WR	Training
N	252	215	4876	34	0	90.68%	70.60%	65.76%	613.86
P	7	4	181	23	0	10.70%	22.77%	GR	
R	56	19	448	23	0	0.00%	0.00%	0.00%	
				Different Costs					
	A	E	N	P	R	Recall	Precision	UR	UP
A	332	115	54	36	74	54.34%	22.07%	41.32%	31.57%
E	256	693	273	115	171	45.95%	34.32%	WR	Training
N	790	1124	2060	591	812	38.31%	80.94%	40.09%	901.92
P	13	16	39	95	52	44.19%	10.00%	GR	
R	113	71	119	113	130	23.81%	10.49%	39.86%	
				SMOTE					
	A	E	N	P	R	Recall	Precision	UR	UP
A	306	120	53	43	89	50.08%	22.82%	39.22%	31.10%
E	224	699	252	138	195	46.35%	34.78%	WR	Training
N	685	1104	2005	671	912	37.29%	81.08%	38.94%	4720.8
P	13	13	46	88	55	40.93%	8.26%	GR	
R	113	74	117	125	117	21.43%	8.55%	37.68%	
				SMOTEV					
	A	E	N	P	R	Recall	Precision	UR	UP
A	255	93	122	17	124	41.73%	25.81%	35.31%	32.34%
E	149	446	634	24	255	29.58%	38.02%	WR	Training

N	487	575	3108	155	1052	57.80%	74.64%	48.61%	2542.8
P	14	12	92	35	62	16.28%	13.01%	GR	
R	83	47	208	38	170	31.14%	10.22%	32.48%	
SMOTEV with different costs									
	A	E	N	P	R	Recall	Precision	UR	UP
A	294	95	98	25	99	48.12%	22.92%	39.23%	32.41%
E	213	510	511	62	212	33.82%	38.93%	WR	Training
N	666	640	2780	342	949	51.70%	76.82%	46.18%	1252.5
P	11	13	59	73	59	33.95%	12.81%	GR	
R	99	52	171	68	156	28.57%	10.58%	38.23%	

Chapter 4

Experiment Results and Discussions

4.1 Experiment setup

For AEC database, the training set was 9959 clean speech chunks uttered by the children in Ohm school and the test set was 8527 clean speech chunks (or synthetic noisy samples) uttered by the children in Mont school. This is the same settings as that used in the INTERSPEECH 2009 Emotion Challenge. Speaker independence was assumed. For BES database, 10-fold cross-validation (CV) was used due of limited number of samples. Ten subsets were formed according to speakers to assure speaker independence. Note that using independent speakers in training and testing causes about 7% decrease of UR compared to a stratified 10-fold CV scheme (in which the training set contains samples from every speaker that represent the whole dataset better).

Additive white Gaussian noise (AWGN) and additive babble noise (ABN) from NOISEX-92 database (Varga and Steeneken, 1993) were employed in the following experiments. The SNR condition ranges from 15dB to 0dB, 5dB decrease each step.

In order to link the results of the matched condition paradigm, the experiments were conducted in two conditions: in the first part, a slack mismatched condition which allows researchers to use the statistics of test data is adopted. In the second, a strict mismatched condition which allows no knowledge of the test data is adopted. Note that only one classifier is train in both parts. The difference between slack and strict mismatched condition lies in the testing stage. For strict condition, both training and testing samples were normalized using the same scaling factor (computed from the training set) whereas for slack condition, training and testing samples were normalized using different scaling factors according to the noise type and SNR.

4.2 Peripheral Materials

This section tackles several “minor issues” that facilitate or jeopardize the major objective. Some of them affect features or classification results in an extreme sense, even represent a shift of paradigm; others are less important but still more or less boost performance when properly handled.

4.2.1 Parameter Selection by Grid Search

Kernel and SVM parameters

Experiments showed that kernels other than linear do not further improve classification results in our case. This is probably because that the data samples are highly overlapping and that the dimension is high. Actually, in non-separable cases, curvy decision boundary helps only when the distribution of samples from different classes are very different. In our cases, distributions of each RS features are very similar between classes so applying RBF kernels does not improve performance. (A reply to Yeh’s future work in Section 1.3)

SVM parameters have an impact on the shape of decision boundaries and thus influence robustness. Robustness is a question of generalization, so if the boundaries are very twisted, the classifier will have reduced performance of generalization. However, robustness of features should not live on classifier’s competence of generalization. In our experiments, SVM parameters are selected to maximize the performance in training set since our assumption of a strict mismatched condition allows us to use no more information other than training samples.

In our cases of RBF kernels, there are two parameters: parameter C, the regularization cost (or ν in the case of ν-SVM; in our case, ν-SVM has numerical problem determiningν), and parameter γ of RBF kernel. Grid search is applied to find the best parameter combination (C, γ). There are other strategies for parameter selection (cf. Friedrichs and Igel, 2004), but a simple grid search serves our purpose. The procedure of grid search has two phases, coarse search and fine search. Coarse search finds a general trend of how the parameter combination influences performance and fine search zooms in a specific area in order to find a better result (Chang and Lin, 2011). The results of grid search on different databases, feature sets are shown in following Tables. All tests are based on strict mismatched condition and speaker independent setting.

Both SVM and ν-SVM have some problems in parameter settings when encountering highly overlapping problems. If the regularization term C of linear kernel SVM is too large, the algorithm will converge in a very slow fashion; the counterpart of C in ν-SVM, parameter ν, cannot be set too large. In our experiments, ν cannot be larger than 0.2152 if the case is Aibo Corpus, and it cannot be larger than 0.5 if the case is Berlin Database.

Table 4.2.1: Parameter selection in AEC database using full RS feature set.

Database: Aibo Emotion Corpus

Feature set: r180

Normalization to [0, 1]

ν-SVM with linear kernel

ν	UR	WR
0.1	11.0990	10.8998
0.2	23.7597	53.6030
0.21	25.9669	57.9145
0.215	26.5859	56.9820
0.2151	24.1896	59.0893
0.2152	n/a	n/a

Table 4.2.2: Parameter selection in BES database using full RS feature set.

Database: Berlin Emotional Speech Database

Feature set: r180

Normalization to [0, 1]

ν-SVM with linear kernel

ν	UR	WR	GR
0.05	62.9833	65.0467	59.5497
0.1	64.2851	65.9813	61.1472
0.15	**64.7768**	**66.1682**	**61.5982**
0.2	63.9185	65.9813	60.6937
0.25	63.8075	65.9813	60.4542
0.3	63.5807	66.1682	59.8891
0.35	62.5823	65.4206	59.3010
0.4	61.0858	64.6729	57.1871
0.45	59.3344	63.1776	55.2406
0.5	n/a	n/a	n/a

In simpler (separable) cases, ν-SVM is known for its convenience of parameter tuning; in our cases, SVM is a superior option. All assessment metrics exhibit higher performance for the original form of SVM; therefore we carried on experiments with the original form of SVM. When C grows larger, it takes much longer to compute the decision boundaries. This phenomenon only happens in linear kernel because linear kernel only allows linear separation. In this case, there will be a lot of misclassified samples, so the kernel becomes very dense and optimization becomes more difficult (large C indicates smaller allowable distance from the decision boundary.).

The impact parameter C has on performance was simply a trade-off between UR and WR when using i384 features. The best C was around 2^{-5} for i384 and it was 2 for r180 features. The discordance became a problem later when we attempted to fuse the two feature sets. Larger value of the regularization term C did not change classification in Berlin Database, which is in accordance with previous report (Keerthi and Lin, 2003).

Table 4.2.3:
Database: Aibo Emotion Corpus
Feature set: r180
Normalization to [0, 1]
C-SVM with linear kernel

C	UR (%)	WR (%)	Training Time (sec)
1	40.8539	38.0162	117.5781
4	40.4612	**38.3311**	139.2412
16	**41.1159**	38.2342	212.9472
64	40.9177	38.0162	530.5290
256	40.5074	38.2463	1896.6185
1024	39.6739	38.0162	>7200

Table 4.1.4:

C	UR (%)	WR (%)	Training Time (sec)
0.0039	35.1718	36.2965	194.2642
0.0078	36.6687	36.9868	185.4706
0.0156	37.9080	37.6892	184.7538
0.0312	38.6105	37.8467	179.2178
0.0625	39.2341	38.1252	169.9313
0.125	39.6744	37.7377	**148.4252**
0.25	40.2018	37.5802	155.9607

0.5	41.0704	38.0041	179.4198
1	40.8539	38.0162	211.7685
2	**41.2416**	**38.5370**	216.3879
4	40.4612	38.3311	228.1803
6	40.522	38.3553	249.0802
8	40.8762	38.3069	252.9136
12	41.0623	38.2827	236.2706
16	41.1159	38.2342	266.8908
20	40.7435	38.1252	278.7337
24	**41.2673**	**38.3674**	285.9171
28	41.0141	38.1737	292.3508
32	41.0003	38.0041	299.1315

Table 4.2.5

Database: Aibo Emotion Corpus

Feature set: i384

Normalization to [0, 1]

C-SVM with linear kernel

C	UR (%)	WR (%)	Training Time (sec)
0.0039	39.8802	37.5076	243.5482
0.0078	40.4281	38.9851	246.4208
0.0156	**40.9513**	**39.8571**	296.5431
0.0312	**41.0799**	**39.7239**	311.3814
0.0625	40.8276	39.2879	300.0468
0.125	39.6519	38.6218	-
0.25	39.0249	38.6702	-
1	38.3846	39.0335	377.3136
2	37.9609	38.9851	-
3	37.6471	38.9367	-
4	37.7094	39.1183	456.2891
6	37.1659	39.3848	-
8	37.1238	39.4938	-
16	37.2128	39.5785	741.8974
64	36.9992	39.4695	1704.751
256	36.5811	39.6512	-
1024	-	-	???? (>3600*5)

Table 4.2.6:

Database: Berlin Emotional Speech Database

Feature set: r180

Normalization to [0, 1]

C-SVM with linear kernel

C	UR (%)	WR (%)
0.0625	62.9155	65.0467
0.125	64.4259	66.1682
0.25	64.3638	65.7944
0.5	67.0633	68.0374
1	66.1917	67.2897
2	**67.7281**	**68.7850**
4	64.775	66.3551
8	64.5947	66.3551
16	63.6399	65.7944

Table 4.2.7

Database: Berlin Emotional Speech Database

Feature set: i384

Normalization to [0, 1]

C-SVM with linear kernel

C	UR (%)	WR (%)	GR(%)
0.0039	28.793	30.2804	0
0.0078	46.5434	47.8505	0
0.0156	53.9342	54.7664	38.5244
0.0312	64.393	64.8598	63.0003
0.0625	67.833	68.972	67.1822
0.125	**69.2811**	**70.4673**	**68.7423**
0.25	67.7995	69.1589	67.2293
0.5	66.4909	68.2243	65.6157
1	65.3944	67.1028	64.5396
2	65.5069	67.2897	64.627
4	65.5069	67.2897	64.627
8	65.5069	67.2897	64.627
256	65.5069	67.2897	64.627

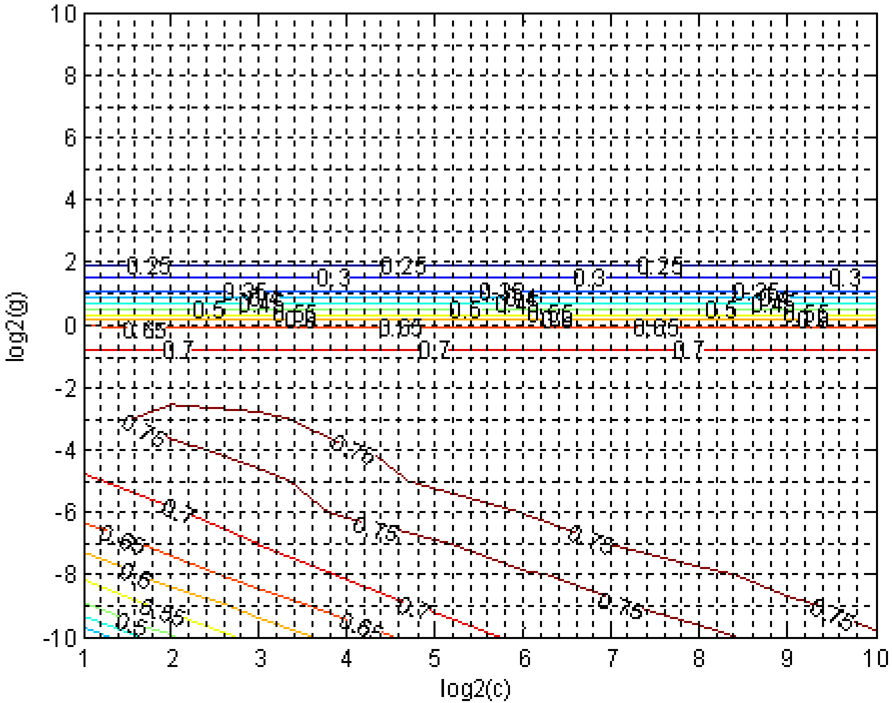
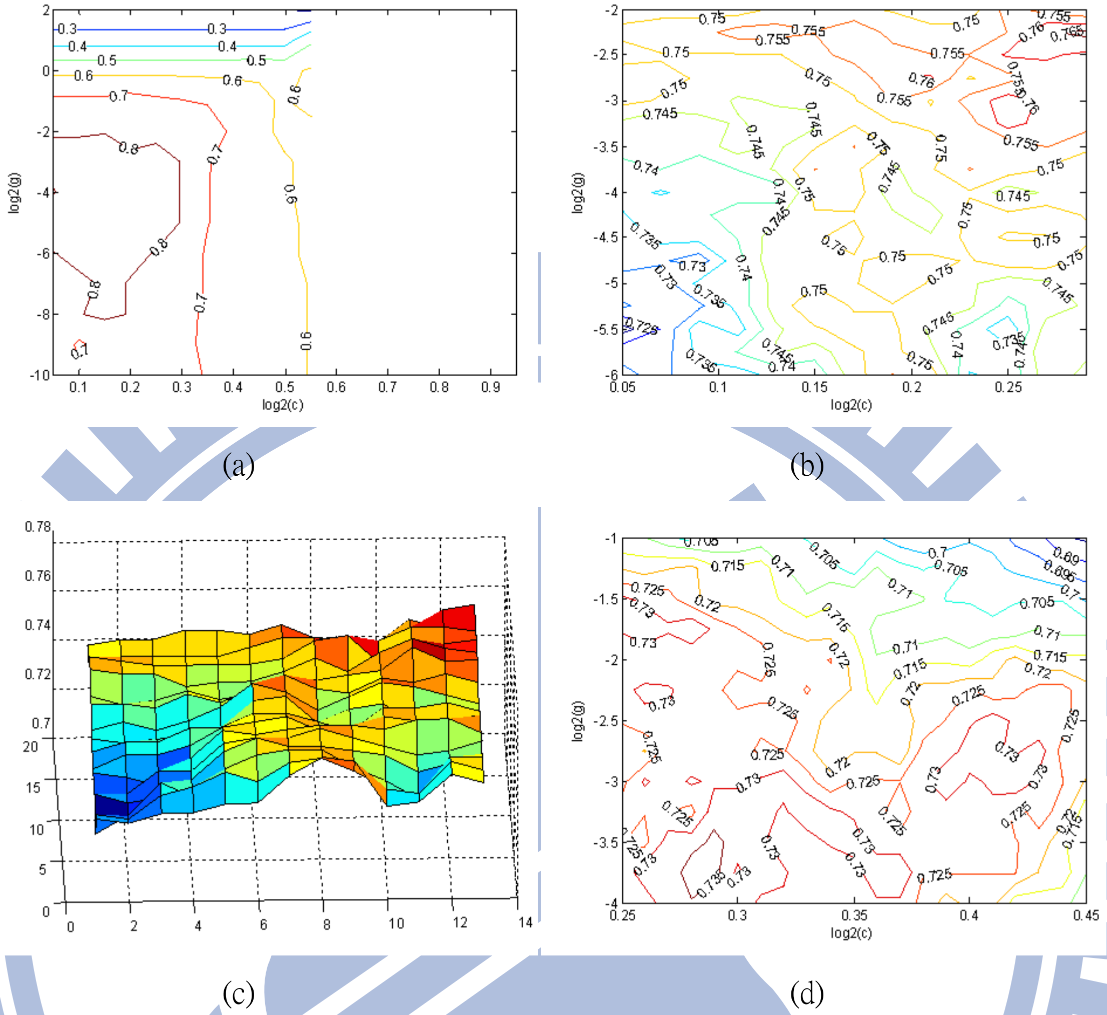


Figure 4.2.1 Grid search for parameter selection in Berlin Database.



(a) (b)

(c) (d)

Figure 4.2.2: Grid search for parameter selection in Berlin Database (ν-SVM).
(a) show the result of coarse-scale search. (c) is a 3D version of (b). (d) is the result of fine-scale search.

4.2.2 Cross-validation Issues in Berlin Database

In Berlin Database, there are only 535 wav files that belong to 7 classes. The number of features is 180 (RS) or 384 (INTERSPEECH). As rule of thumb, the number of training samples is better to be 5 or 10 times to the number of features. If it is not the case, cross-validation (CV) is suggested to be employed for a better estimation of performance.

There are several schemes of cross-validation. Stratified 10-fold cross-validation schemes are widely adopted to ensure a better (unbiased and minimal variance) estimator. Nevertheless, in our experiments, we intended to apply the same criterions to both Aibo Corpus and Berlin Database. Since a speaker independent training/testing configuration was set in Aibo Corpus, we decided to apply 10-fold cross validation to Berlin Database according to the 10 speakers.

One merit from speaker independent settings is that the performance estimator is fixed. In random 10-fold cross-validation, the estimator is a random variable, which makes comparison between feature sets harder. The downside, of course, is that the performance gets lower than that in a stratified cross-validation. Experimental results in Table 4.2.8 shows about 7% decrease when adopting speaker independent settings.

Table 4.2.8: Comparison between speaker dependent and independent settings

Speaker dependent setting was carried out by random 10-fold CV; the results are shown in the upper chart. Speaker independent setting was carried out by 10-fold CV according to 10 speakers; the results are shown in the lower chart. Both tests were conducted under clarity condition. Note that Anger gets more confusion with Happy, and Neutral becomes less distinctive under speaker independent settings.

	H	A	D	F	N	B	S	Recall	UR
H	33	23	2	10	3	0	0	46.5%	73.2%
A	9	115	0	2	1	0	0	90.6%	WR
D	1	2	25	8	5	3	2	54.3%	75.9%
F	18	4	2	41	4	0	0	59.4%	GR
N	3	0	4	1	65	5	1	82.3%	70.8%
B	1	0	3	0	7	67	3	82.7%	
S	0	0	0	0	1	1	60	96.8%	
	50.8%	79.9%	69.4%	66.1%	75.6%	88.2%	90.9%		

	H	A	D	F	N	B	S	Recall	UR
H	30	27	3	11	0	0	0	42.3%	66.3%
A	20	103	1	3	0	0	0	81.1%	WR
D	3	3	20	11	4	3	2	43.5%	68.8%
F	21	5	3	37	3	0	0	53.6%	GR
N	3	0	4	6	55	9	2	69.6%	63.5%
B	1	0	3	0	8	64	5	79.0%	
S	0	0	1	0	2	0	59	95.2%	
	38.5%	74.6%	57.1%	54.4%	76.4%	84.2%	86.8%		

4.3 Experiments on Robustness

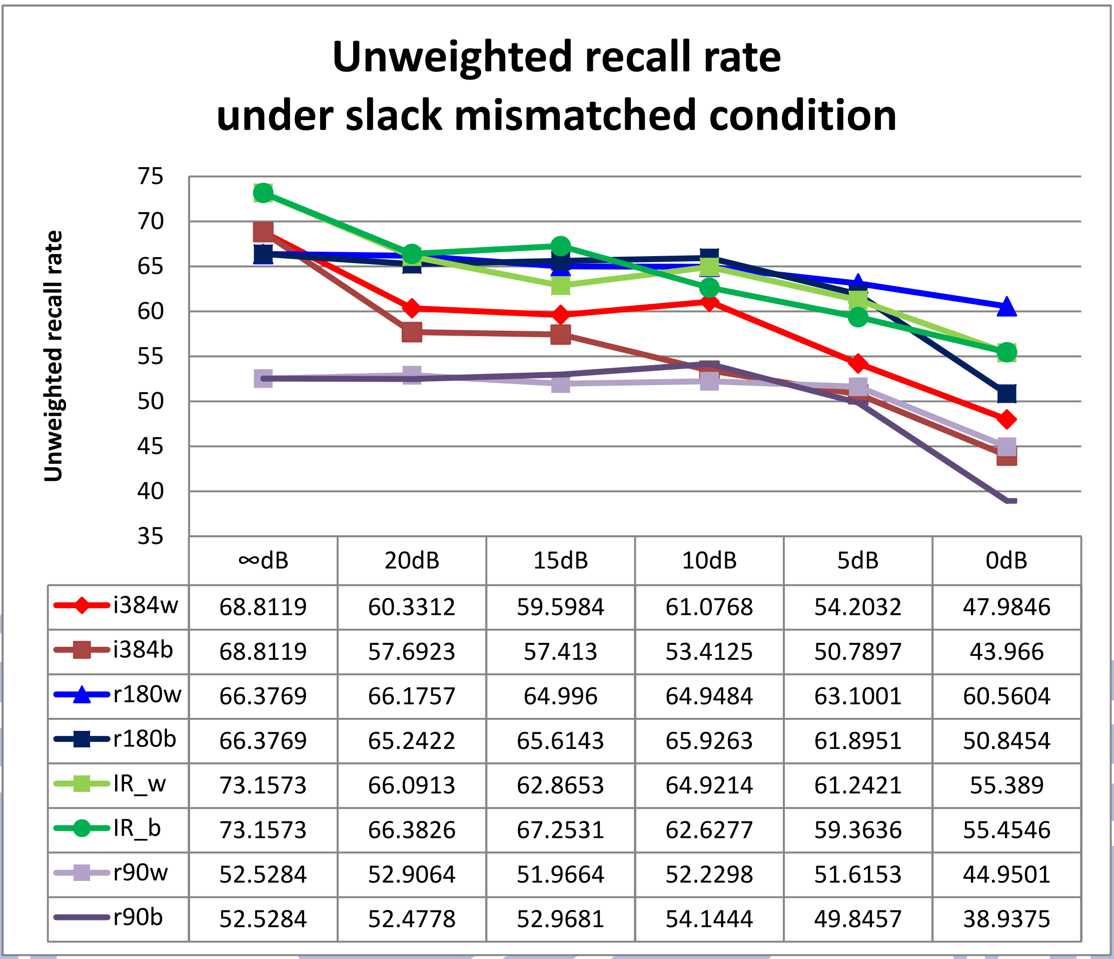
4.3.1 Slack Mismatched Condition

Some abbreviations are listed in this passage. The proposed rate-scale features which contain spectro-temporal modulation information are called the **RS** features. The RS features comprise two subsets: the $\mathbf{RS_{mu}}$ set consists of temporal mean of RS and the $\mathbf{RS_{sd}}$ set consists of temporal standard deviation (SD) of RS. The 384 INTERSPEECH features are denoted as **i384**, and the totality of RS_{mu} and RS_{sd} is denoted as **r180**.

The following experimental results are split into two parts based on slack and strict matched conditions. All other setting are held the same, except for one thing: under slack condition, the hybrid feature set is i384 and r180 but under strict condition it is i384 and RS_{sd}. The reason is that under slack condition, even RSmu is robust, so it can be added into the hybrid set.

The results from slack mismatched condition are shown in Fig. 6 and Fig. 7 and Table 3 and Table 4. In both BES and AEC databases, the unweighted recall rate of RS_{sd} features holds almost the same except for 0 dB condition. On the other hand, the performance of i384 has an apparent trend of decreasing. Although i384 fares about 8% better than r180 does in clean condition, increasing advantage of RS features arises through decreasing SNR. Similar trends also appear under matched condition.

There is one thing that needs notice: a single raise in UR does not mean that the performance is elevated under that SNR condition. It is the trend instead of a single point on the curve that matters.



	∞dB	20dB	15dB	10dB	5dB	0dB
i384w	68.8119	60.3312	59.5984	61.0768	54.2032	47.9846
i384b	68.8119	57.6923	57.413	53.4125	50.7897	43.966
r180w	66.3769	66.1757	64.996	64.9484	63.1001	60.5604
r180b	66.3769	65.2422	65.6143	65.9263	61.8951	50.8454
IR_w	73.1573	66.0913	62.8653	64.9214	61.2421	55.389
IR_b	73.1573	66.3826	67.2531	62.6277	59.3636	55.4546
r90w	52.5284	52.9064	51.9664	52.2298	51.6153	44.9501
r90b	52.5284	52.4778	52.9681	54.1444	49.8457	38.9375

Figure 4.3.1: Curves for UR v. SNR in BES database under slack mismatched condition.

*i384: 384 baseline features used in INTERSPEECH 2009 Emotion Challenge.

*r180: proposed spectro-temporal modulation (rate-scale) features.

*r90: half of the proposed features with only temporal standard deviation (RS_{sd}).

*IR: i384 combining r180 features (564 features in total). w: white noise. b:babble noise.

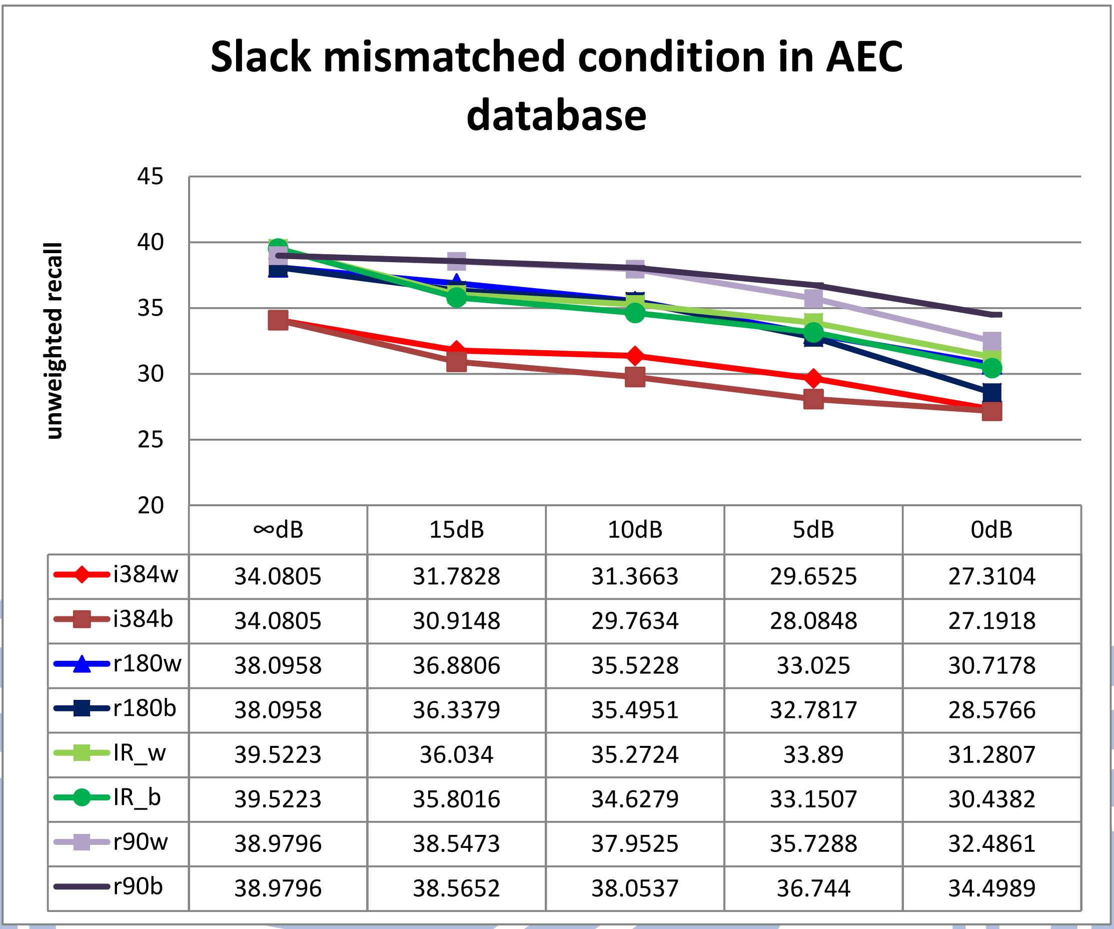
*Training/Testing set: 10-fold cross-validation with speaker independence

Table 4.3.1: The confusion matrices of r180 under ∞, 0dB white noise and 0dB babble noise condition.

	∞ dB						
	H	A	D	F	N	B	S
H	29	24	5	13	0	0	0
A	25	95	3	3	1	0	0
D	2	3	27	9	1	2	2
F	20	3	7	37	2	0	0
N	2	0	8	6	52	9	2
B	1	0	4	0	8	60	8
S	0	0	1	0	1	0	60
UR	66.38%						
WR	67.29%						
GR	64.29%						

	0 dB (white noise)							0 dB (babble noise)						
	H	A	D	F	N	B	S	H	A	D	F	N	B	S
H	26	29	5	10	1	0	0	24	20	5	16	5	1	0
A	20	103	2	1	0	0	1	34	80	5	8	0	0	0
D	2	3	26	9	2	3	1	6	2	23	6	6	1	2
F	20	4	10	31	2	0	2	18	4	12	24	6	0	5
N	2	0	17	6	45	6	3	2	0	15	6	41	6	9
B	1	0	7	1	9	57	6	0	0	4	0	20	43	14
S	0	0	7	0	7	0	48	0	0	4	2	6	7	43
UR	60.56%							50.85%						
WR	62.80%							51.96%						
GR	58.48%							49.30%						

*The columns are classification results and the rows contains true label.



	∞dB	15dB	10dB	5dB	0dB
i384w	34.0805	31.7828	31.3663	29.6525	27.3104
i384b	34.0805	30.9148	29.7634	28.0848	27.1918
r180w	38.0958	36.8806	35.5228	33.025	30.7178
r180b	38.0958	36.3379	35.4951	32.7817	28.5766
IR_w	39.5223	36.034	35.2724	33.89	31.2807
IR_b	39.5223	35.8016	34.6279	33.1507	30.4382
r90w	38.9796	38.5473	37.9525	35.7288	32.4861
r90b	38.9796	38.5652	38.0537	36.744	34.4989

Figure 4.3.2: The performance of four feature sets in two type of noise under slack mismatched condition.

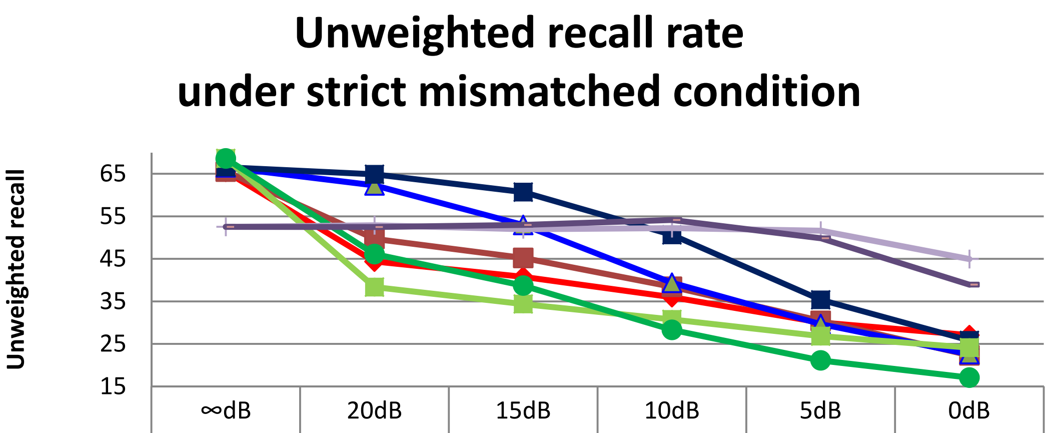
*w: white noise. b: babble noise. r90: RS_{sd};

Table 4.3.2: Confusion matrices of classification result using r90 feature set.

	∞dB					0dB (white noise)					0dB (babble noise)				
	A	E	N	P	R	A	E	N	P	R	A	E	N	P	R
A	284	141	83	29	74	164	151	126	72	98	172	125	148	64	102
E	183	712	409	66	138	236	564	409	115	184	194	481	536	96	201
N	546	1118	2721	338	654	611	993	2475	590	708	709	957	2528	515	668
P	10	20	87	54	44	21	33	82	51	28	23	35	87	40	30
R	91	84	184	72	115	72	84	194	89	107	77	102	191	82	94
UR	38.10%					30.72%					28.58%				
WR	47.06%					40.70%					40.15%				
GR	35.79%					29.27%					26.68%				

4.3.2 Strict Mismatched Condition

Under strict mismatched condition, the performance of both r180 and i384 features decreases in a rather fast fashion. On the contrary, RS_{sd} holds fair performance when noise is not too severe.



	∞dB	20dB	15dB	10dB	5dB	0dB
i384w	65.4998	44.3692	40.7651	35.9994	30.0137	27.0519
i384b	65.4998	49.7023	45.2093	38.4405	30.3997	22.2993
r180w	66.5532	62.2958	52.967	39.377	29.652	22.5866
r180b	66.5532	64.8847	60.6835	50.6261	35.3618	25.8061
IR_w	68.6049	38.3585	34.3785	30.7386	26.8331	24.1201
IR_b	68.6049	46.1104	38.7074	28.3047	21.1159	17.0742
r90w	52.5284	52.9064	51.9664	52.2298	51.6153	44.9501
r90b	52.5284	52.4778	52.9681	54.1444	49.8457	38.9375

Figure 4.3.3: Curves for UR v. SNR in BES database under strict mismatched condition.

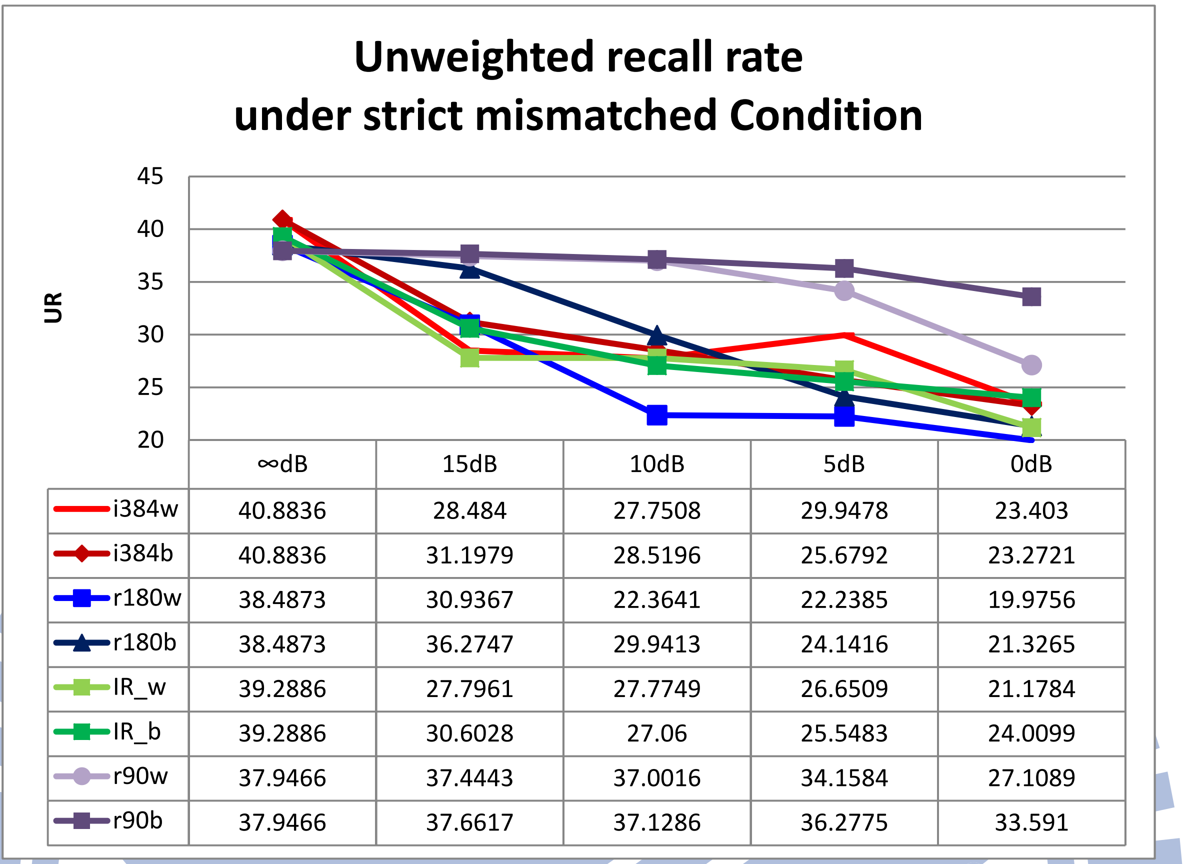
*Training/Testing set: 10-fold cross-validation with speaker independence

*IR: hybrid feature set of i384 and r90.

Table 4.3.3: Confusion matrix using r180 feature set under ∞, 0dB white noise and 0dB babble noise condition in Berlin Database.

∞ dB							
	H	A	D	F	N	B	S
H	19	25	6	19	2	0	0
A	33	86	3	4	1	0	0
D	7	3	20	7	7	1	1
F	22	4	9	22	10	0	2
N	2	0	7	9	43	8	10
B	0	0	2	1	13	60	5
S	0	0	1	2	9	7	43
UR	52.53%						
WR	54.77%						
GR	49.25%						

	0 dB (white noise)							0 dB (babble noise)						
	H	A	D	F	N	B	S	H	A	D	F	N	B	S
H	11	11	8	31	10	0	0	12	19	1	21	14	1	3
A	33	46	8	34	5	1	0	41	68	1	11	4	2	0
D	4	0	13	12	10	1	6	5	2	1	7	20	1	10
F	3	1	8	39	12	0	6	7	2	0	16	20	2	22
N	0	0	4	5	42	0	28	0	0	0	2	31	3	43
B	0	0	2	1	23	32	23	0	0	0	0	9	33	39
S	0	0	0	1	7	1	53	0	0	0	0	2	0	60
UR	44.95%							38.94%						
WR	44.11%							41.31%						
GR	39.90%							25.52%						



	∞dB	15dB	10dB	5dB	0dB
i384w	40.8836	28.484	27.7508	29.9478	23.403
i384b	40.8836	31.1979	28.5196	25.6792	23.2721
r180w	38.4873	30.9367	22.3641	22.2385	19.9756
r180b	38.4873	36.2747	29.9413	24.1416	21.3265
IR_w	39.2886	27.7961	27.7749	26.6509	21.1784
IR_b	39.2886	30.6028	27.06	25.5483	24.0099
r90w	37.9466	37.4443	37.0016	34.1584	27.1089
r90b	37.9466	37.6617	37.1286	36.2775	33.591

Figure 4.3.4: The performance of four feature sets in two type of noise under strict mismatched condition. The database is the Aibo Emotion Corpus.

Table 4.3.4: Confusion matrices of classification result using r90 feature set.

	∞dB					0dB, white noise					0dB, babble noise				
	A	E	N	P	R	A	E	N	P	R	A	E	N	P	R
A	364	91	55	30	71	76	37	55	398	45	277	51	92	132	59
E	400	660	198	95	155	84	221	235	855	113	358	321	408	306	115
N	1356	985	1656	595	785	260	295	998	3399	425	1107	350	2028	1418	474
P	28	17	35	77	58	4	3	17	172	19	23	5	47	110	30
R	168	66	107	97	108	29	16	70	377	54	117	27	141	193	68
UR	37.95%					27.11%					33.59%				
WR	34.70%					18.42%					33.96%				
GR	35.56%					19.30%					29.72%				

* The columns are classification results and the rows contains true label.

4.3.3 Discussion on Emotion

Emotions at opposing extremes of arousal are easier to discriminate according to the dimensional emotion theory. This phenomenon is observable in Table 5 (H versus Others for valence and HADF v. NBS for arousal). The presence of noise inflicts more damage on recognition of closer emotions (emotions in the same emotion family) than that on more unrelated emotions. Again in Table 5, for example, it is more likely that the classifier confuses Neutral with Boredom but it is less likely that the classifier confuses Sadness with Anger (cf. (Schuller et al, 2006) for similar tests). The similarity emotion psychology described above also corresponds to the results of data visualization in Figure 4.3.7.

It is interesting when we take a look on how noise influences RS_{sd} and further influences classification results. As Table 6 and Table 5 show, additive babble noise caused classification to skew toward neutral emotion (N) in both databases. This result is not unimaginable because babble noise contains numerous intelligible pieces of utterance. The superposition of those pieces of utterance makes an emotion neutral speech-like sound which shares similar traits with emotion neutral speech. Therefore, speech samples with high density of babble noise tend to be classified as Neutral.

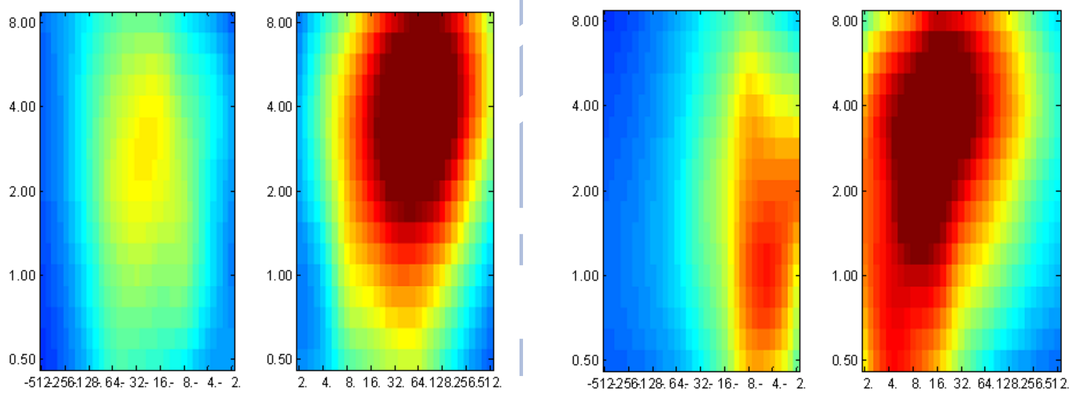
White noise, on the other hand, skews classification toward the emotion category that has more drastic change of pitch-related attributes. In the case of BES, it would be Fear and in AEC, it would be Positive (including Joyful and Motherese). One might wonder why in BES, the result was not skewed toward Happy. This is because in BES, the acoustics of Fear samples is much more significant than that of Happy samples.

4.3.4 Discussion on Robustness

In this paragraph, we discuss how noise affects RS features. Under slack mismatched condition, the effect from noise is partly removed by the normalization and therefore the degradation is greatly reduced; nonetheless, under strict mismatched condition, with the increasing presence of noise pattern, the structure of the temporal mean of RS features (later denoted as RS_{mu}) is gradually destroyed, causing rapid degradation of the UR. This indicates that when there is no available knowledge that can adjust the testing sample, i.e. when it is unable to apply slack mismatched condition, even the RS_{mu} are not robust.

On the other hand, the temporal standard deviation of the RS features (later denoted as RS_{sd}) which fares limited ability of recognition is fairly robust even under strict mismatched condition. The reason the two sets of RS features differ in robustness performance is explained here. The RS features are derived from spectrum. When additive noise comes in, the energy is elevated thus resulting in elevated RS_{mu}. The elevation is not removed under strict mismatched condition, so degradation in performance is inevitable. However, addition in spectrum inflicts minor effects to variance and that is why RS_{sd} is robust.

Noise with the same type usually has the similar RS pattern. Figure 4.3.5 show typical patterns of AWGN and babble noise respectively. Babble nose has stronger response in low rate region in both positive and negative rate half-planes while AWGN affects more on higher rate regions. In this point of view, how noise affects speech is merely a translation in the feature space. (Of course, additive noise does not result in pure translation.) This is why the classification (trained by RS_{mu} and RS_{sd}) appears to give all testing samples the same label in very low SNR under strict mismatched condition.



(a) (b)

Figure 4.3.5: Babble noise and white noise.

Only RS_{mu} is shown here. The x-axis represents “rate” and the y-axis represents “scale”.

Hybrid Features

In Fig. 4.3.1 and Fig. 4.3.2, under slack mismatched condition, combining i384 with r180 features is beneficial to robustness in both AEC and BES databases. Unfortunately, the hybrid feature set did not work well under strict condition. The discrepancy is natural because under slack condition the distribution of testing samples can be regulated whereas under strict condition the normalization results in a biased distribution of testing data. For example, if testing samples are just a translation of original training samples, under slack condition, the translation can be compensated; however, under strict condition, the translation is not mended. In short, under matched or slack mismatched condition, a hybrid set of i384 and RS (either RS_{mu} or RS_{sd}) features helps the totality to gain robustness; under strict condition, applying robust feature sets in classification is more applicable.

An alternative argument that the normalization (under strict mismatched condition) worsened the robustness of i384 features. This argument is half right and it also implies that i384 features are not robust.

Fusion Schemes

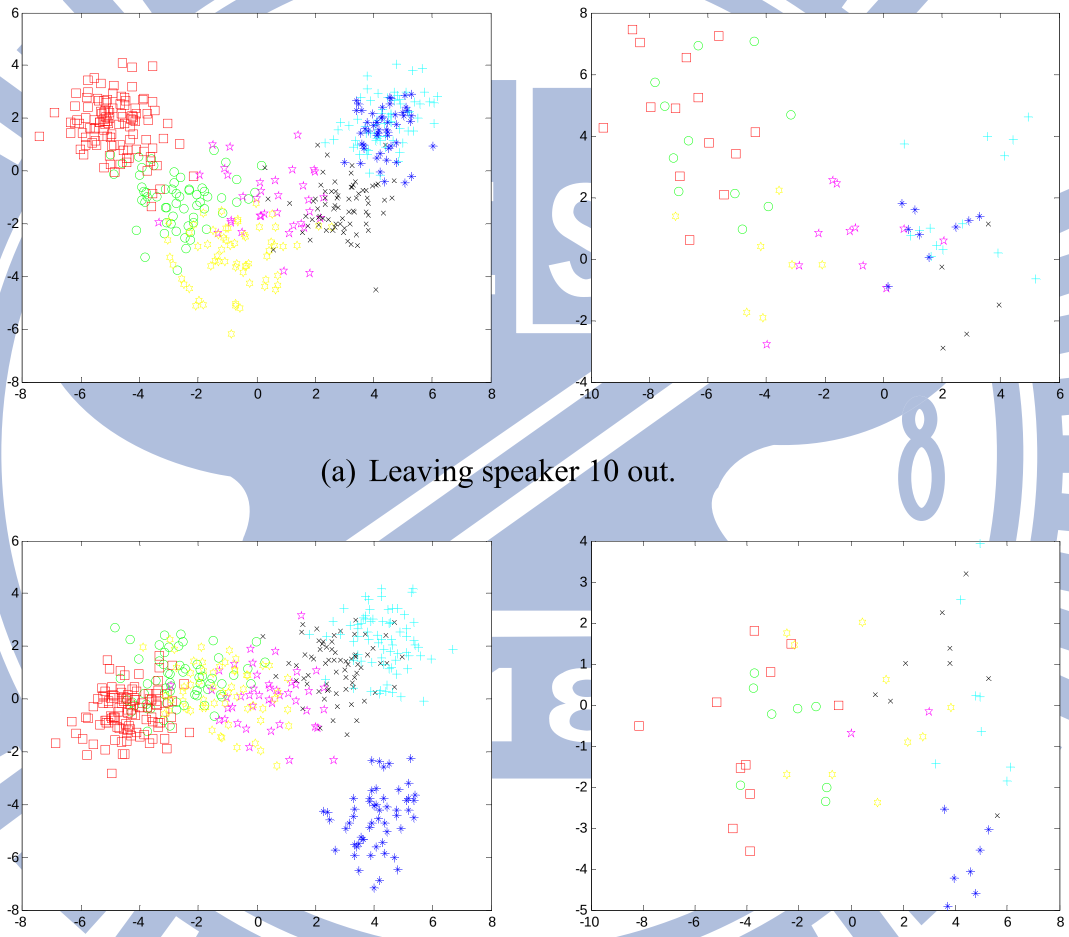
Instead of feature fusion, another approach is to fusion the decision of every learning machine, just like what the Emotion Challenge did. In our preliminary experiments, three classifiers (SVM's) were built up using r180 features, i384 features, and r180+i384 features, respectively. The output then went through a majority mechanism to give a final decision. The following chart shows a slight performance boost (which is a better result than that of any single participants in 2009 Emotion Challenge). Two fusion schemes, committee and expert decision, were adopted and both resulted in better UR. Nonetheless, a robust fusion scheme against noise is still an open and ongoing quest because it requires several robust (and discriminative) sets of features.

Feature Dimension

A brief inspection of feature dimension or feature reduction is presented in the following figures. Results were visualized in two-dimensional plots. Sample distributions after applying linear discriminant analysis (LDA, which is supervised) and principal component analysis (PCA, which is unsupervised) showed similar trends.

The results of LDA (training set) look like a belt of grouped clusters, inlaid with Anger, Joy, Fear, Disgust, Neutral, Boredom, and Sadness, in sequence. The order as well as the constellation also indicates the internal similarity between each class pair. High activation emotions (Anger, Joy, Fear, and Disgust) locate at adjacent places and low activation emotions (Neutral, Boredom, and Sadness) also locate at adjacent places. Despite of LDA's nearly perfect constellations in the training phase, its generalization ability appears to be slightly mediocre, at least not better than PCA. Comparing the right panels of Fig. 4.3.6 and Fig. 4.3.7 (a), we can observe resembling trends such as the fact that Anger and Joy are highly overlapping.

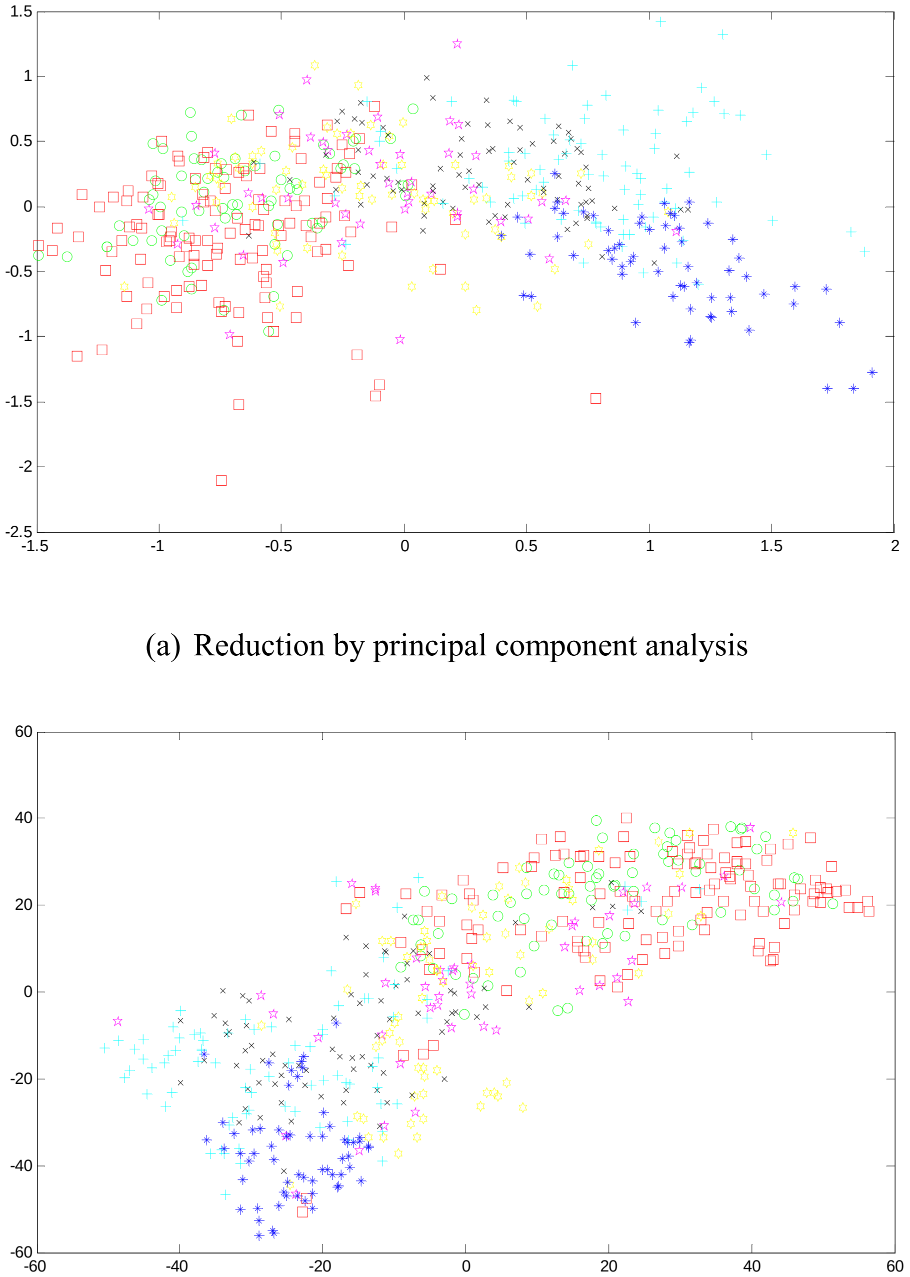
In unsupervised feature reduction scheme (PCA), there is no apparent constellation for samples from the same emotion class. However, high activation emotions and low activation emotions still distribute at opposing locations.



(a) Leaving speaker 10 out.

(b) Leaving speaker 5 out.

Figure 4.3.6: Reducing feature dimension (r180) to two using LDA.



(a) Reduction by principal component analysis

(b) Reduction by t-distributed Stochastic Neighbor Embedding

Figure 4.3.7: Visualization after reducing feature dimension (r180) to two.

*Green circles = Happy; Red squares = Anger; Magenta pentagons = Disgust; Yellow hexagons = Fear; Black crosses = Neutral; Cyan plus signs = Boredom; Blue asteroids = Sadness.

Table 4.3.5: Comparison between combining and the original models

			r180						
	A	E	N	P	R	Recall	Precision	UR	UP
A	392	104	44	24	47	64.16%	18.56%	41.24%	31.67%
E	377	735	209	69	118	48.74%	34.17%	WR	
N	1171	1214	1847	402	743	34.35%	81.94%	38.54%	
P	21	19	54	74	47	34.42%	11.37%	GR	
R	151	79	100	82	134	24.54%	12.30%	39.04%	
			i384						
	A	E	N	P	R	Recall	Precision	UR	UP
A	339	167	39	19	47	55.48%	18.56%	40.94%	32.04%
E	326	861	168	57	96	57.10%	33.58%	WR	
N	1003	1423	1872	404	675	34.81%	83.27%	39.69%	
P	36	18	56	70	35	32.56%	11.11%	GR	
R	123	95	113	80	135	24.73%	13.66%	38.87%	
			r180+i384						
	A	E	N	P	R	Recall	Precision	UR	UP
A	296	188	51	20	56	48.45%	21.48%	38.84%	31.74%
E	262	800	280	30	136	53.05%	31.09%	WR	
N	690	1444	2208	266	769	41.06%	80.79%	42.33%	
P	21	31	51	59	53	27.44%	13.82%	GR	
R	109	110	143	52	132	24.18%	11.52%	37.07%	
			Committee						
	A	E	N	P	R	Recall	Precision	UR	UP
A	353	154	48	18	38	57.77%	21.64%	42.52%	33.27%
E	304	836	234	30	104	55.44%	33.07%	WR	
N	831	1424	2173	281	668	40.41%	82.37%	43.24%	
P	25	18	52	74	46	34.42%	15.74%	GR	
R	118	96	131	67	134	24.54%	13.54%	40.53%	
			Expert						
	A	E	N	P	R	Recall	Precision	UR	UP
A	454	84	20	30	23	74.30%	16.19%	42.73%	32.36%
E	503	719	125	99	62	47.68%	36.35%	WR	
N	1625	1096	1548	635	473	28.79%	85.10%	35.30%	
P	32	15	44	97	27	45.12%	9.96%	GR	
R	190	64	82	113	97	17.77%	14.22%	38.24%	

*The ”Expert” fusion scheme is achieved by this mechanism: r180 decides Anger and Positive and i384 decides the other three emotions.

4.4 Inspection into RS Features

4.4.1 Higher-Order Statistics

Temporal higher-order statistics of RS were a suggested feature set in Yeh's future work, so we made an attempt to investigate it. The 3rd and 4th central moments were tested and the results are shown in the following tables (all tests were conducted under clarity condition and white noise). The definition of the two statistics is given in the following formulae.

Sample skewness: $\gamma \triangleq \dfrac{\frac{1}{T}\sum_{t=1}^{T}(x_t-\bar{x})^3}{(\frac{1}{T}\sum_{t=1}^{T}(x_t-\bar{x})^2)^{\frac{3}{2}}}$

Sample (excess) kurtosis: $\kappa \triangleq \dfrac{\frac{1}{T}\sum_{t=1}^{T}(x_t-\bar{x})^4}{(\frac{1}{T}\sum_{t=1}^{T}(x_t-\bar{x})^2)^2} - 3$

Note that the original definition of skewness and kurtosis is:

Skewness: $\gamma \triangleq \mathrm{E}\left[\left(\dfrac{X-\mu}{\sigma}\right)^3\right]$

(Excess) Kurtosis: $\kappa \triangleq \mathrm{E}\left[\left(\dfrac{X-\mu}{\sigma}\right)^4\right] - 3$

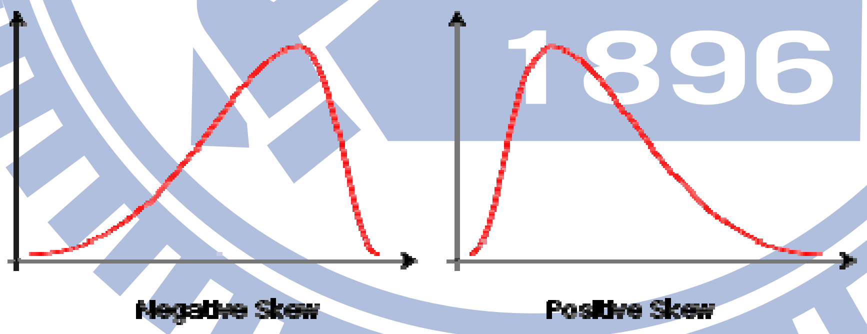


Figure 4.4.1 Illustration of positive and negative skewness.

Experimental results did not show any clues that indicates any forms of improvements. Since skewness and kurtosis are descriptors of the shape of distributions, no improvement indicates that shape is not a major issue or that the distributions are highly overlapped.

Table 4.4.1

Database: Berlin Emotional Speech Database

Random 10-fold cross-validation

Clarity condition

No.	μ	σ	γ	κ	UR (%)	WR (%)
1	O	×	×	×	75.41	76.07
2	×	O	×	×	63.16	65.42
3	×	×	O	×	53.11	55.70
4	×	×	×	O	45.74	48.59
5	O	O	×	×	76.30	75.14
6	O	×	O	×	74.03	75.14
7	O	×	×	O	70.63	71.77
8	×	O	O	×	66.70	69.53
9	×	O	×	O	64.94	66.54
10	×	×	O	O	54.23	56.82
11	O	O	O	×	74.41	75.32
12	O	O	×	O	73.40	74.39
13	O	×	O	O	70.95	71.96
14	×	O	O	O	65.55	67.85
15	O	O	O	O	71.70	73.27

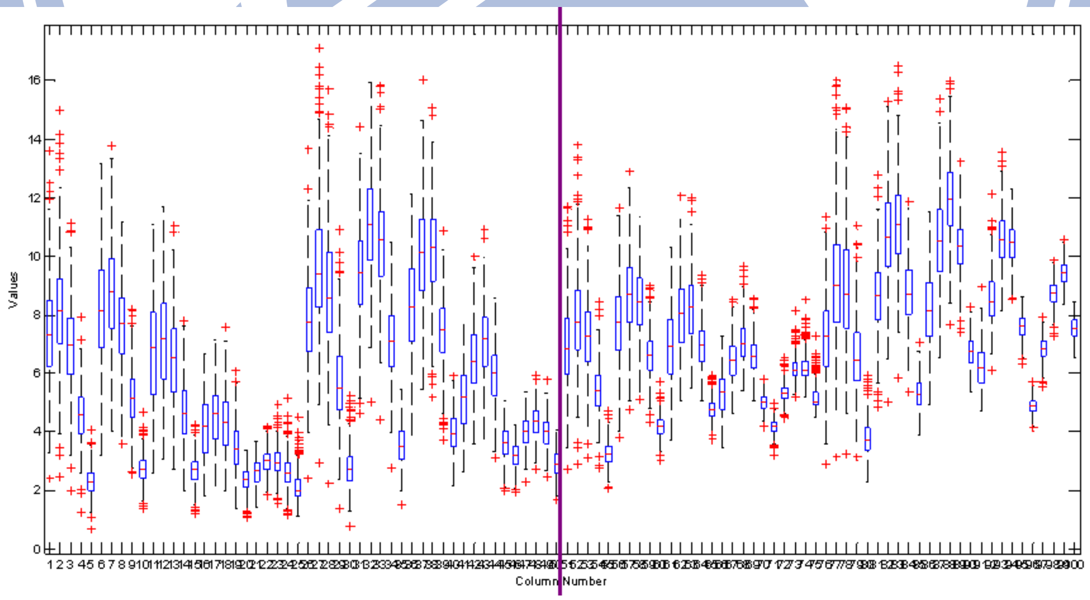


Figure 4.4.2: Box plot of $\mathrm{RS}_{\mathrm{mu}}$ in ∞ and 0 dB SNR condition (white noise).

The vertical purple line separates the two SNR conditions.

Left: ∞ dB SNR; Right: 0 dB SNR. (Same setting for Figure 4.4.2-5)

This figure shows RS_{mu} in the form of vector. Notice that feature distributions in high rate RS_{mu} regions (right hand region of each part) become biased by a positive value (contaminated by noise) in low SNR compared to that high SNR.

*The top, middle, and bottom of a blue bar stand for the upper quartile, median, and lower quartile, respectively.

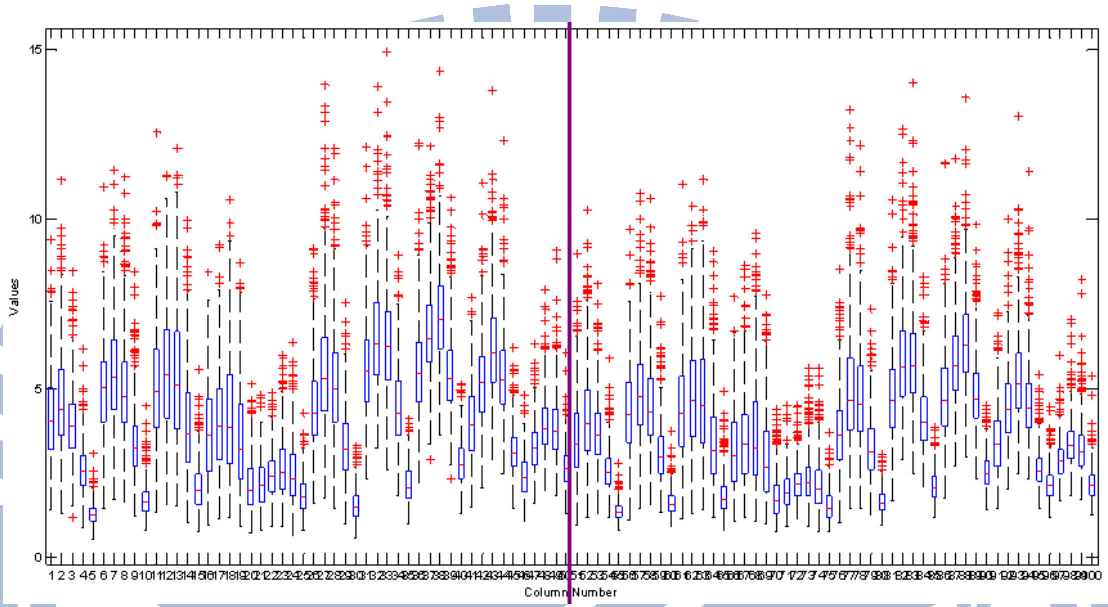


Figure 4.4.3: Box plot of RS_{sd} in ∞ and 0 dB SNR condition (white noise). The distributions remains almost the same in ∞ and 0 dB SNR condition.

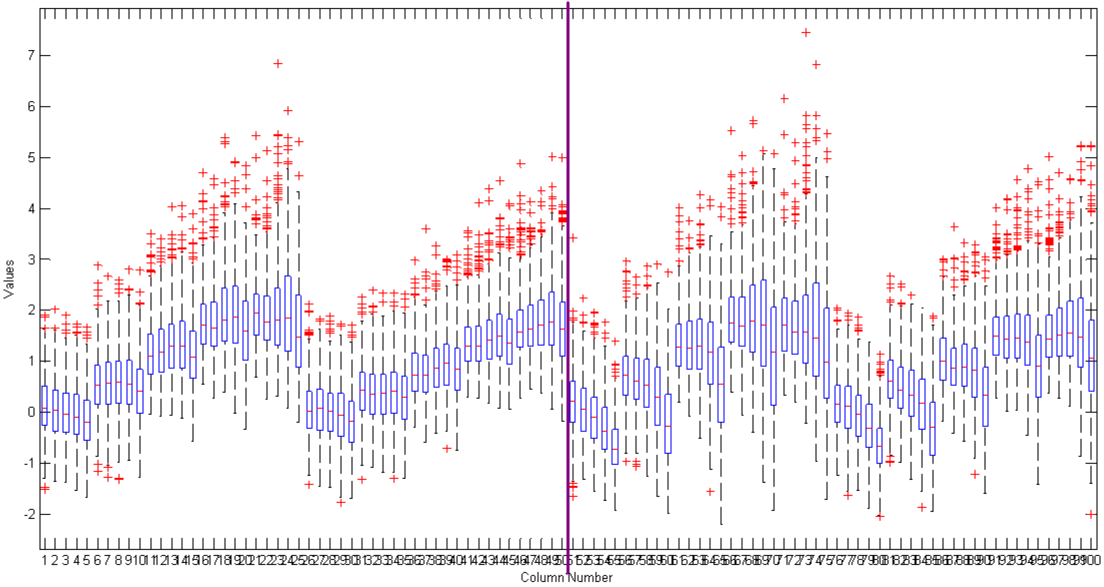


Figure 4.4.4: Box plot of temporal skewness of RS features in ∞ and 0 dB SNR condition (white noise).

An apparent trend of decreasing is observable in some RS regions. The decrease comes from the additive white Gaussian noise (AWGN) because AWGN results in a negative-skewed distribution on the spectrogram (χ^2 distribution). The change of the distribution of skewness directly reveals the fact that skewness is not a robust feature.

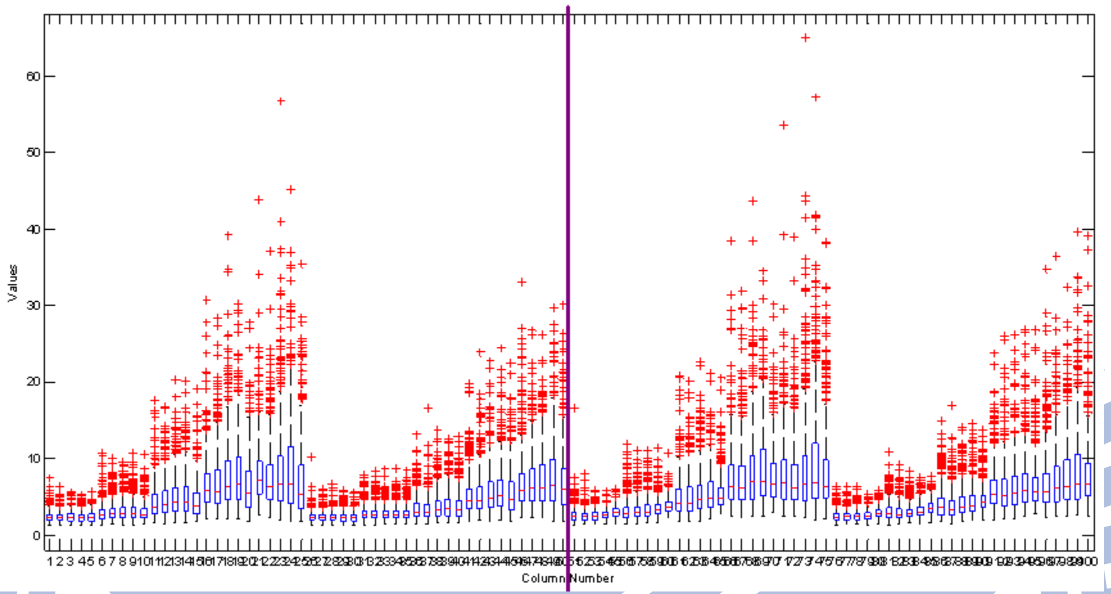


Figure 4.4.5: Box plot of temporal kurtosis of RS features in ∞ and 0 dB SNR condition (white noise).

The change of the distribution of kurtosis is not obvious. Nevertheless, using kurtosis as a feature did not show any sign of recognizibility or robustness.

4.4.2 Subset Features

Intrinsic Dimension and Feature Reduction

The RS features have high correlations with one another, especially with adjacent regions. If a feature set consists of high correlation elements, the feature set may very well be further reduced in number (not necessarily; sometimes highly correlated features still facilitates performance). Experiments showed the possibility to reduce feature dimensionality for both r180 and i384 features. It is likely that the most intrinsic features are not fully exploited in current methods.

Table 4.4.2: Intrinsic dimension estimation for currently used feature sets.

Intrinsic dimension	Estimate	
Estimator / Features	r180	i384
Correlation dimension	4.6457	2.6871
Eigenvalue evaluation	8	6
Maximum likelihood	13.2327	26.1225
Geodesic minimum spanning tree	12.4757	45.3283

In Yeh's previous work, she discussed feature selection or reduction in a quantitative way. Nonetheless, the analysis was inconclusive. Which feature should be included or excluded is still an open question. Her final feature set is a chessboard selection version of the original r180.

In the experiments this paper presents, feature selection was never the major attempt. The RS features certainly need refining, but the critical issue is feature extraction itself. The results of a basic feature reduction based on principal component analysis (PCA) are shown in the following tables and figures. Reducing dimensionality skews classification toward Neutral because Neutral comprises more samples and is easier to describe. The first principal component (pc1) of Aibo Corpus is similar to the first and the second one of Berlin Database (describing responses of ordinary speech). Resembling pairs can be also found in pc2 of Aibo Corpus and pc3 of Berlin Database (describing very low rate region, possibly representing speech rate), and pc6 of Aibo Corpus and pc5 of Berlin Database (very low rate and scale, possibly intonation contour).

As for the question "to what extent can we reduce the dimensionality," our answer is "it depends on to what extent we can tolerate the performance loss." Reducing feature dimensionality by PCA did not show any signs of performance enhancement. Rather, the relation between dimensionality and performance is a trade-off.

Table 4.4.3: Classification results of features with reduced dimensionality

The number M following r represents the first M principal components.

			r180					
	A	E	N	P	R	Recall	Precision	UR
A	385	98	53	16	59	63.01%	18.64%	40.24%
E	367	699	233	68	141	46.35%	33.21%	WR
N	1140	1207	1881	361	788	34.98%	81.43%	38.34%
P	19	19	38	71	68	33.02%	12.03%	GR
R	155	82	105	74	130	23.81%	10.96%	38.11%

r127

	A	E	N	P	R	Recall	Precision	UR
A	382	107	42	29	51	62.52%	18.13%	40.97%
E	368	738	209	67	126	48.94%	33.36%	WR
N	1180	1267	1775	432	723	33.01%	81.09%	37.58%
P	24	20	51	79	41	36.74%	11.63%	GR
R	153	80	112	72	129	23.63%	12.06%	38.78%

r89

	A	E	N	P	R	Recall	Precision	UR
A	371	102	43	32	63	60.72%	17.27%	38.95%
E	370	715	212	73	138	47.41%	33.55%	WR
N	1219	1201	1765	447	745	32.82%	81.83%	36.82%
P	30	25	43	68	49	31.63%	9.65%	GR
R	158	88	94	85	121	22.16%	10.84%	36.66%

r47

	A	E	N	P	R	Recall	Precision	UR
A	335	122	57	21	76	54.83%	15.83%	37.17%
E	350	718	226	69	145	47.61%	32.86%	WR
N	1252	1238	1707	440	740	31.75%	79.58%	35.63%
P	28	23	48	65	51	30.23%	9.53%	GR
R	151	84	107	87	117	21.43%	10.36%	35.15%

r13

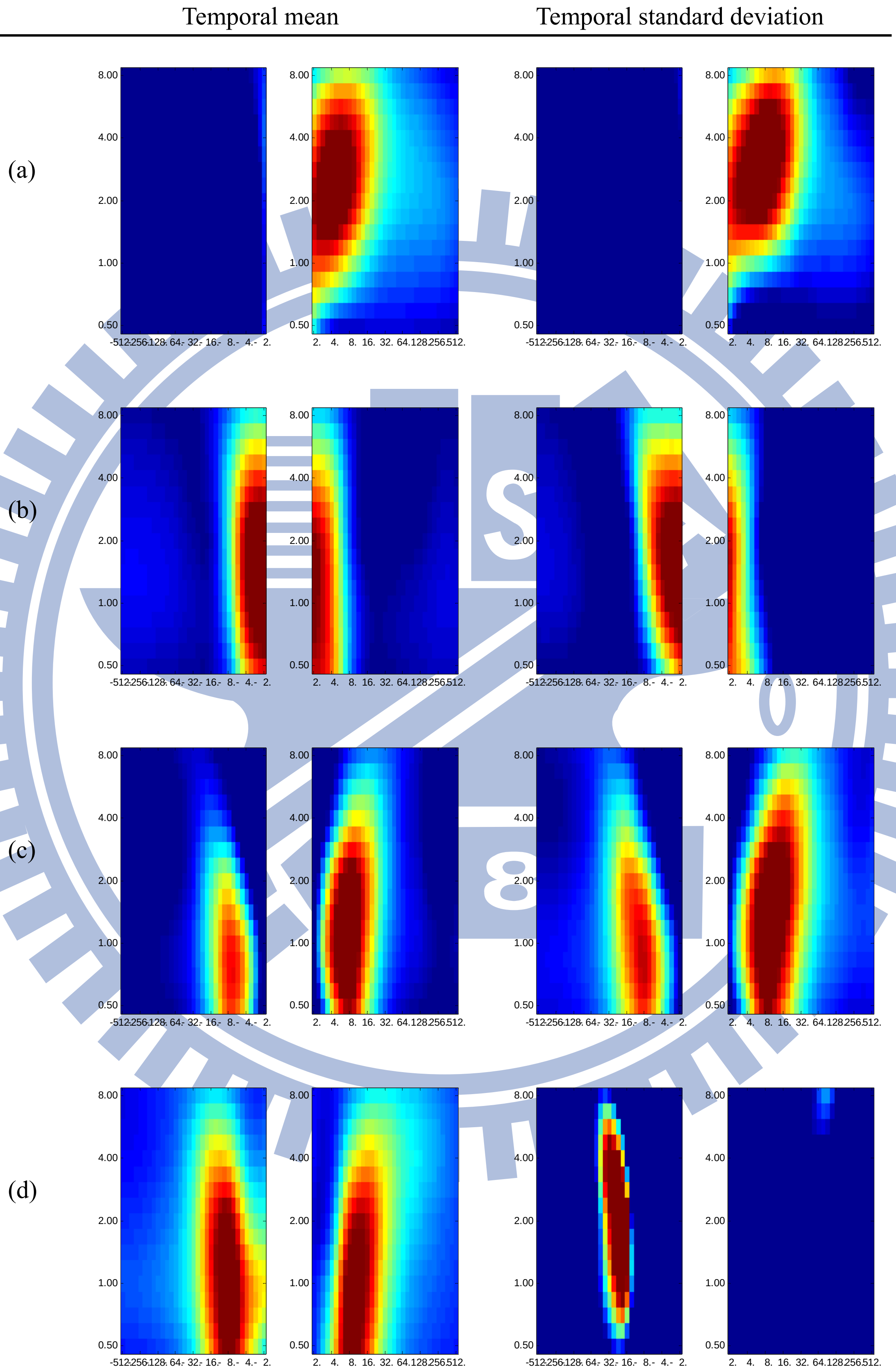
	A	E	N	P	R	Recall	Precision	UR
A	265	116	90	62	78	43.37%	13.95%	34.21%
E	322	589	362	134	101	39.06%	29.07%	WR
N	1120	1196	2008	690	363	37.34%	76.26%	36.43%
P	32	28	49	87	19	40.47%	8.06%	GR
R	160	97	124	106	59	10.81%	9.52%	30.79%

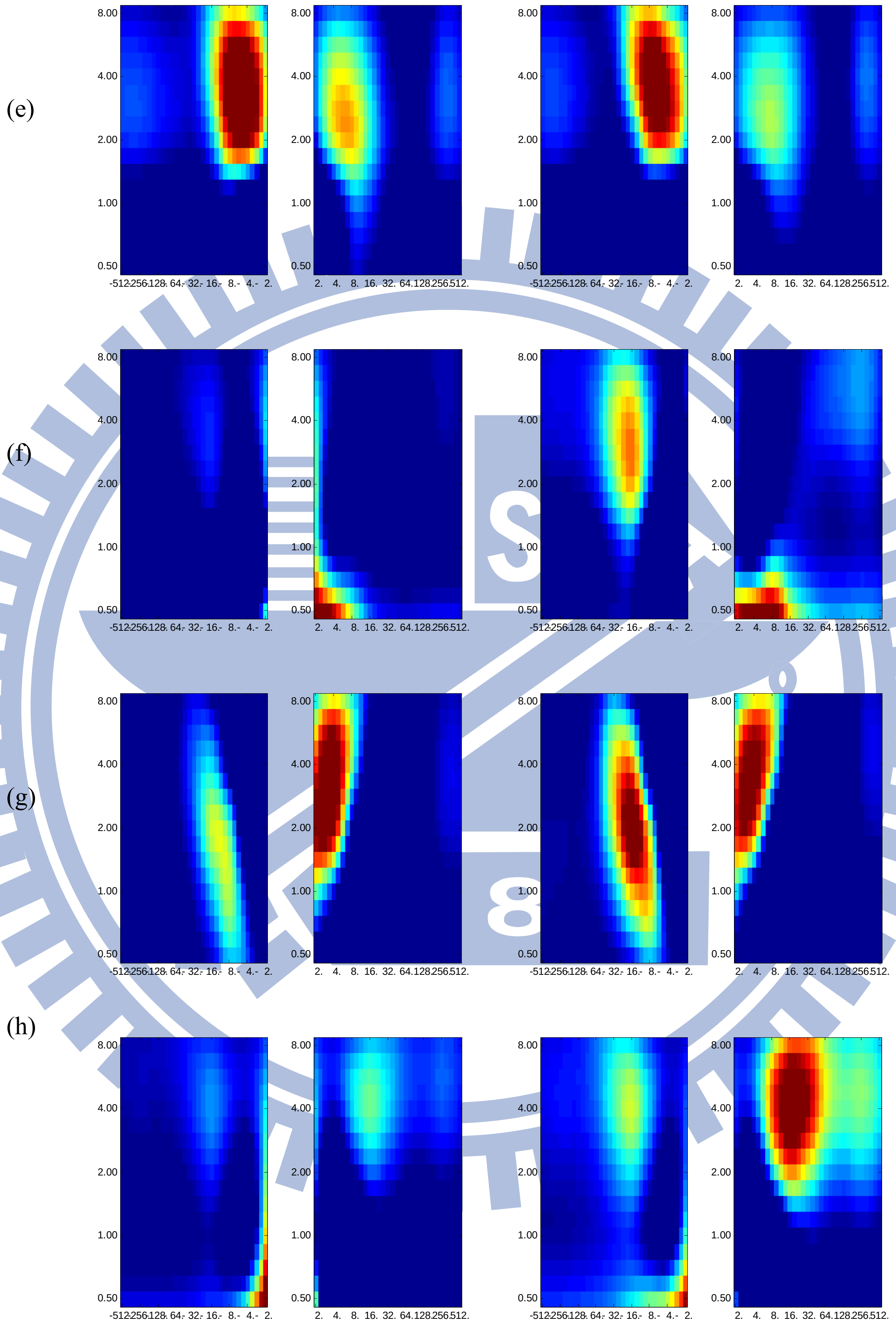
r6

	A	E	N	P	R	Recall	Precision	UR
A	277	136	91	86	21	45.34%	13.58%	31.02%
E	341	587	384	126	70	38.93%	27.51%	WR
N	1219	1265	2009	596	288	37.36%	74.57%	35.95%
P	38	45	67	57	8	26.51%	5.91%	GR
R	165	101	143	99	38	6.96%	8.94%	26.12%

*The dimension reduction toolkit was provided by (van der Maaten, 2009).

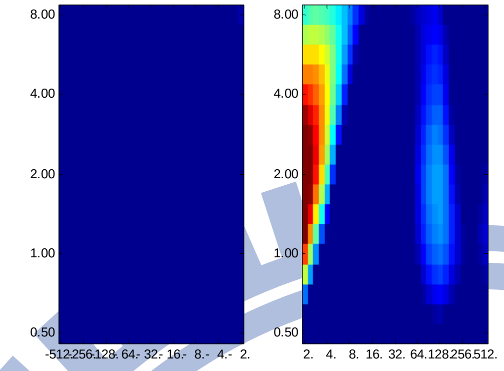
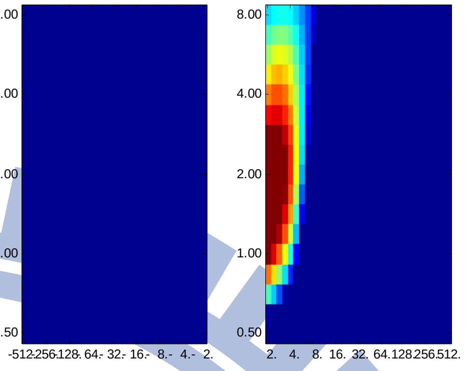
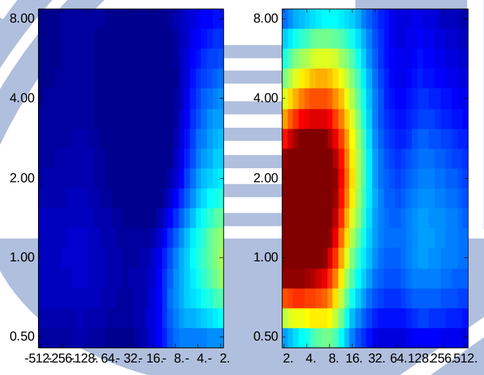
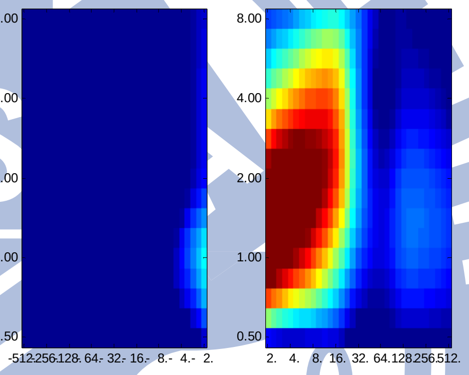
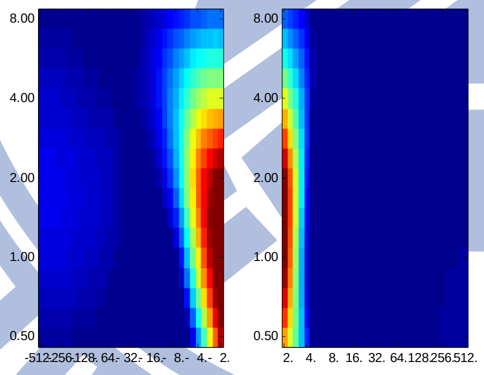
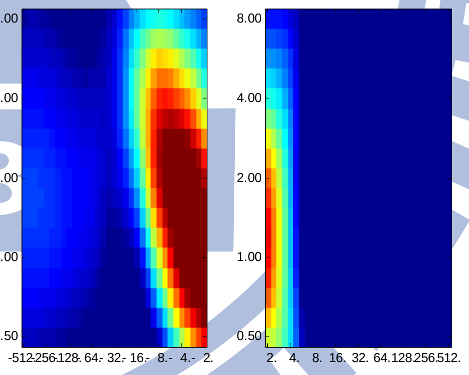
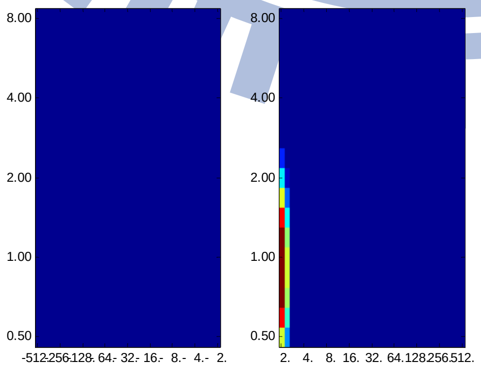
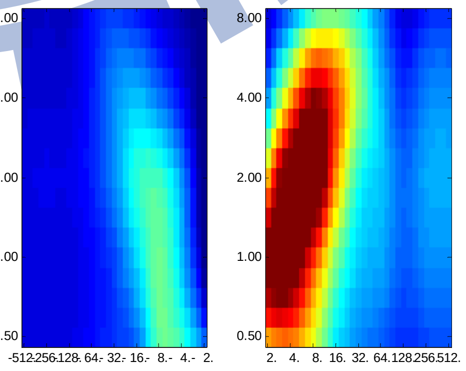
Figure 4.4.6: First eight principal components on RS plot (Aibo Corpus)

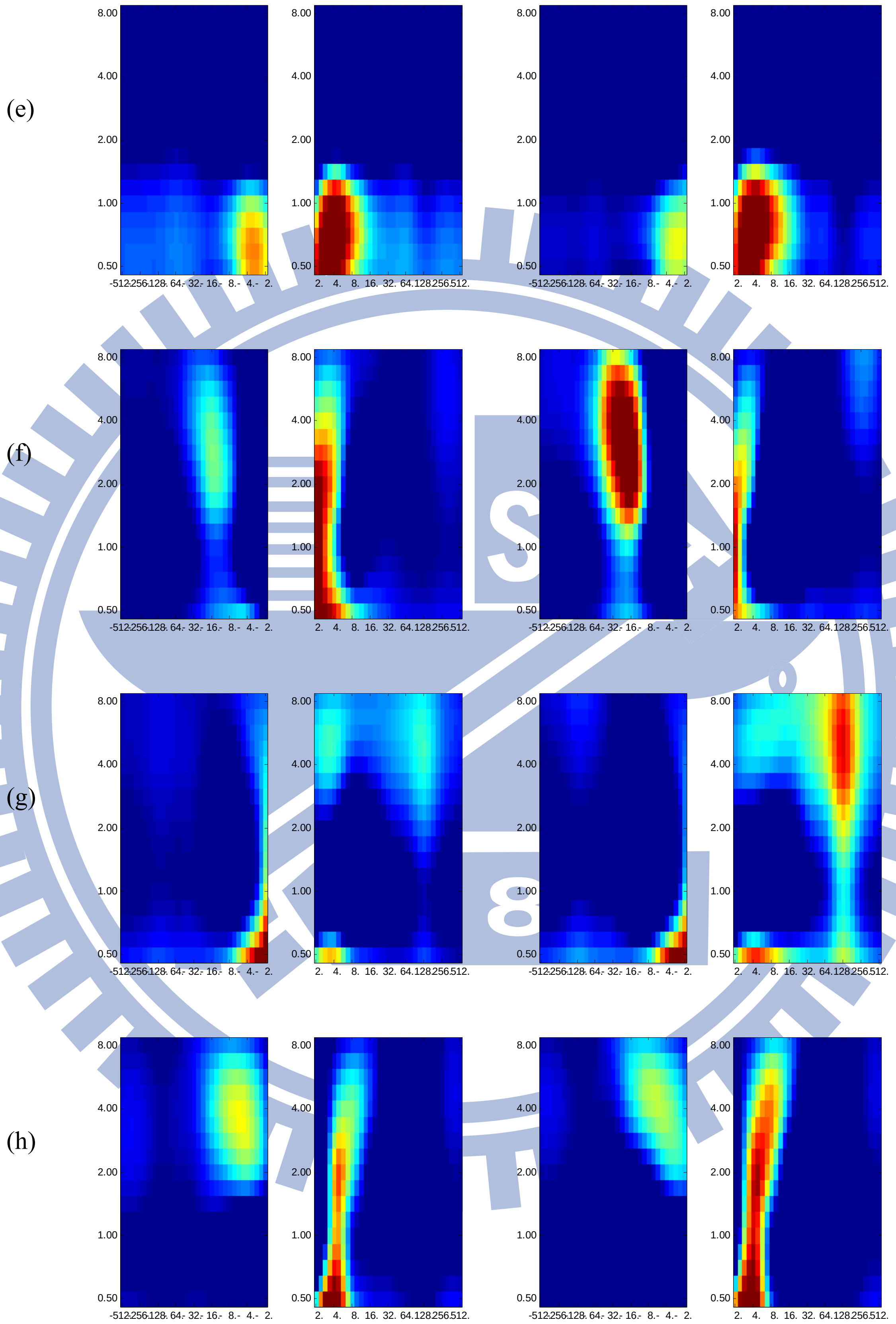




*The meaning of the principal components latter than eighth is not obvious.

Table 4.4.7: First eight principal components on RS plot (Berlin Database)

	Temporal mean	Temporal standard deviation
(a)		
(b)		
(c)		
(d)		


(e)
(f)
(g)
(h)
8.00
4.00
2.00
1.00
0.50
-512-256-128- 64- 32- 16- 8.- 4.- 2.
2. 4. 8. 16. 32. 64.128256512.

Subsets and combining models

This section investigates the performance the subsets of RS features can achieve and whether the combining models improve performance. The RS features were partitioned into several combinations. Details information and experiment settings are shown in following tables. All tests were conducted under strict mismatched condition using Aibo Corpus.

We found that even if the combining model improves performance, the performance was still worse than that of the whole set. These results show indicate that not all sort of combining models improve performance. It is recommended that features with different properties be used in different classifiers (or systems). Classifiers using features that have similar properties are almost useless in combining models. For feature subsets with similar properties, “the whole is more than the sum of its parts.”

Scale partition

Table 4.4.3: Recognition rate of different scale subset

Subset 1: scale = 2^{-1}; Subset 2: scale = 2^{0}; Subset 3: scale = 2^{1}; Subset 4: scale = 2^{2}; Subset 5: scale = 2^{3}

	UR	WR	GR
Subset 1	37.3423	38.0647	34.879
Subset 2	37.2518	33.2809	35.0134
Subset 3	37.7067	34.3951	35.1747
Subset 4	37.7015	36.0906	35.4039
Subset 5	37.5182	35.5577	35.319
Committee	39.5575	37.1459	36.7967

Rate partition

Table 4.4.4: Recognition rate of different rate subset

Subset 1: rate ranges from $2^{1} – 2^{2}$; Subset 2: rate ranges from $2^{3} – 2^{4}$; Subset 3: rate ranges from $2^{5} – 2^{6}$; Subset 4: rate ranges from $2^{7} – 2^{8}$; Subset 5: rate ranges from 2^{9}

	UR	WR	GR
Subset 1	32.957	33.6442	31.7611
Subset 2	34.8385	33.9833	33.3662
Subset 3	36.6005	30.5438	34.4506
Subset 4	34.7996	27.9884	31.7499
Subset 5	30.6498	29.2964	28.0137
Committee	35.7428	32.7477	33.2654

High-rate/low-rate partition

Table 4.4.5: Confusion matrices using high-rate/low-rate features.

Subset 1: rate ranges from 2^1 - 2^5; Subset 2: rate ranges from 2^6 - 2^9; Subset 3: All rates

Low rates

	A	E	N	P	R	Recall	Precision	UR	UP
A	345	120	51	27	68	56.46%	15.58%	37.88%	29.82%
E	389	699	253	53	114	46.35%	31.98%	WR	
N	1289	1269	1679	352	788	31.23%	78.02%	35.39%	
P	31	19	52	67	46	31.16%	12.05%	GR	
R	161	79	117	57	132	24.18%	11.50%	36.13%	

High rates

	A	E	N	P	R	Recall	Precision	UR	UP
A	399	99	35	14	64	65.30%	15.41%	37.77%	30.04%
E	471	667	161	69	140	44.23%	33.02%	WR	
N	1500	1148	1346	401	982	25.03%	81.63%	31.68%	
P	39	27	36	60	53	27.91%	9.74%	GR	
R	180	79	71	72	144	26.37%	10.41%	35.09%	

All rates

	A	E	N	P	R	Recall	Precision	UR	UP
A	392	104	44	24	47	64.16%	18.56%	41.24%	31.67%
E	377	735	209	69	118	48.74%	34.17%	WR	
N	1171	1214	1847	402	743	34.35%	81.94%	38.54%	
P	21	19	54	74	47	34.42%	11.37%	GR	
R	151	79	100	82	134	24.54%	12.30%	39.04%	

*The combining model got 41.12% UR and 38.74% GR. It did not improve even class-wise performance, let alone overall performance.

4.5 Emotion in Perceptual Features

4.5.1 Acted Emotions

Acted emotions are easier to recognize both to human and to machines. However, similarity does exist between some emotions. As mentioned in Section 2.3, anger and joy are prone to be misidentified through vocal communications in spite of the fact that it is highly unlikely that people confuse the two in facial expressions. The RS plot showing RS_{mu} and RS_{sd} features of the seven emotions in Berlin Database are shown in following figures.

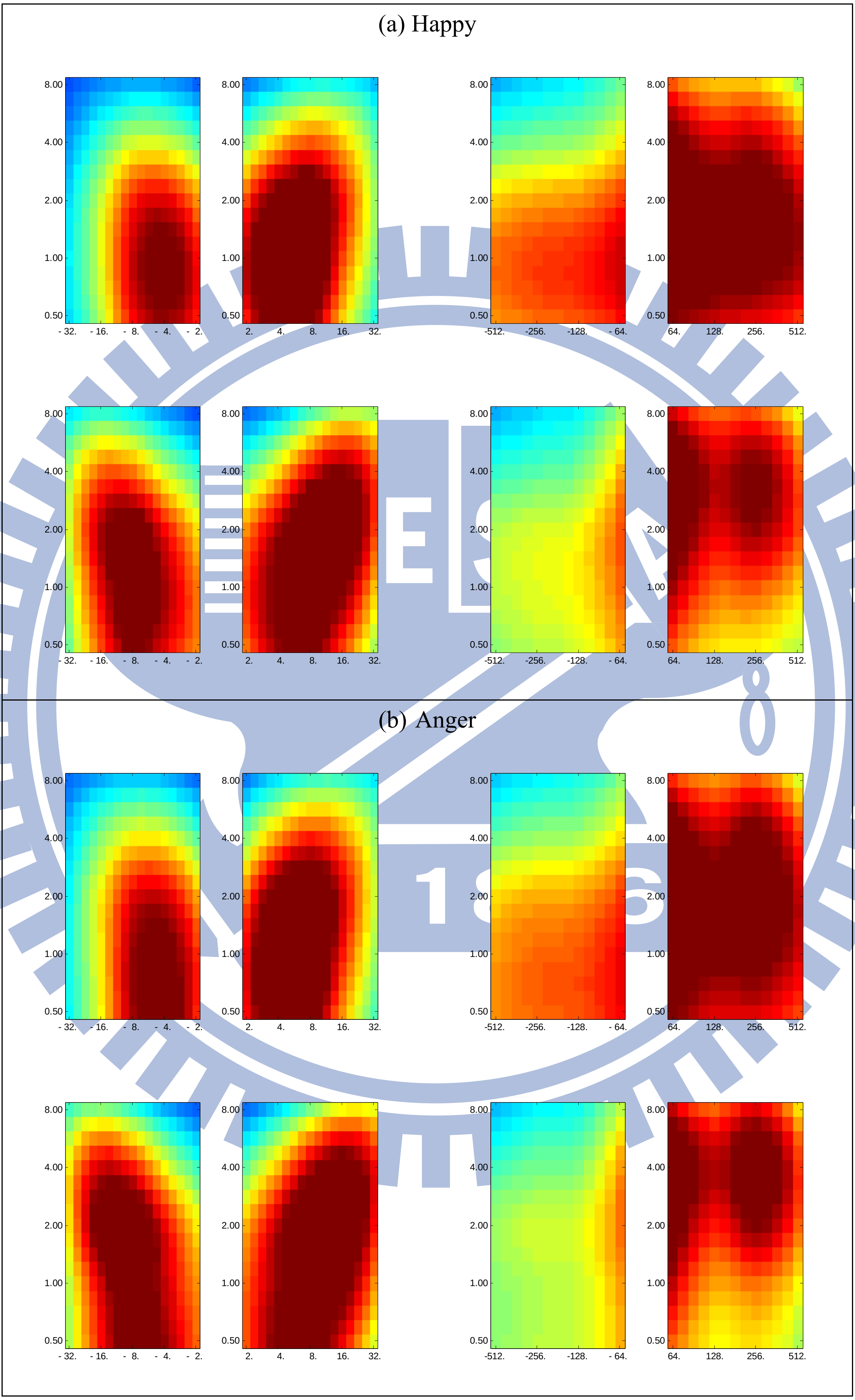
1. Climbing Voice

   High activation emotions, namely Anger and Happy, have strong responses in “climbing voice” (left part of RS plots); low activation ones (Neutral, Boredom, and Sadness), on the other hand, have weak response in climbing voice. This fits intuition and everyday experiences because people tend to speak with plainer intonations when they feel no interest in the conversation or feel nothing particular. When people are sad, they speak slower and with less upward intonation. The imbalance between upward and downward intonation is the most significant indicator of high and low activation emotions.
2. Speaking rate

   Low rate areas are indicators of speaking rate. When the speaking rate is slow, very low rate area (rate<2) will have much response. Speaking rate is a good separator of low activation emotions.
3. Pitch

   Pitch activities usually become much lively when people are activated. Although the RS_{mu} plots showed no apparent differences, the RS_{sd} plots do. The RS_{sd} plots of Happy and Anger showed vivid activities, while those of Neutral and Boredom showed plain. Note that the maximum value of each plot is different.
4. Low scale variation

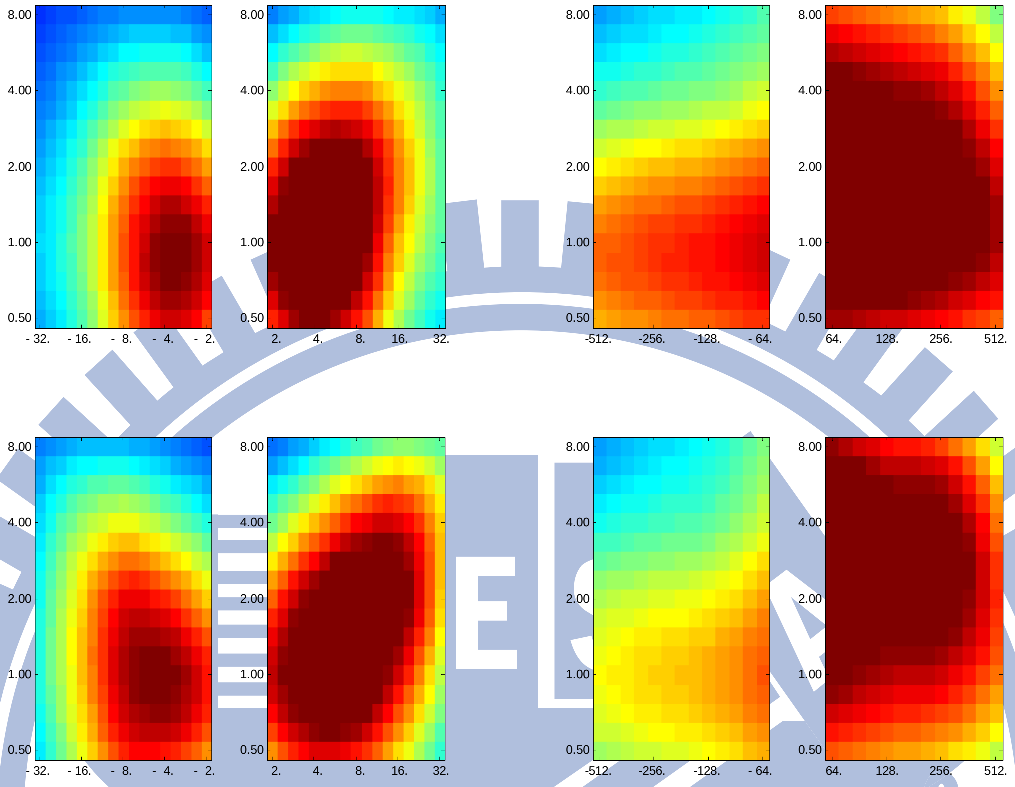
   The variation of low scale regions is a secondary discriminative factor for emotions at similar activation level. Anger and Happy have strong variation (RS_{sd}) in low scale regions while Disgust and Fear do not.



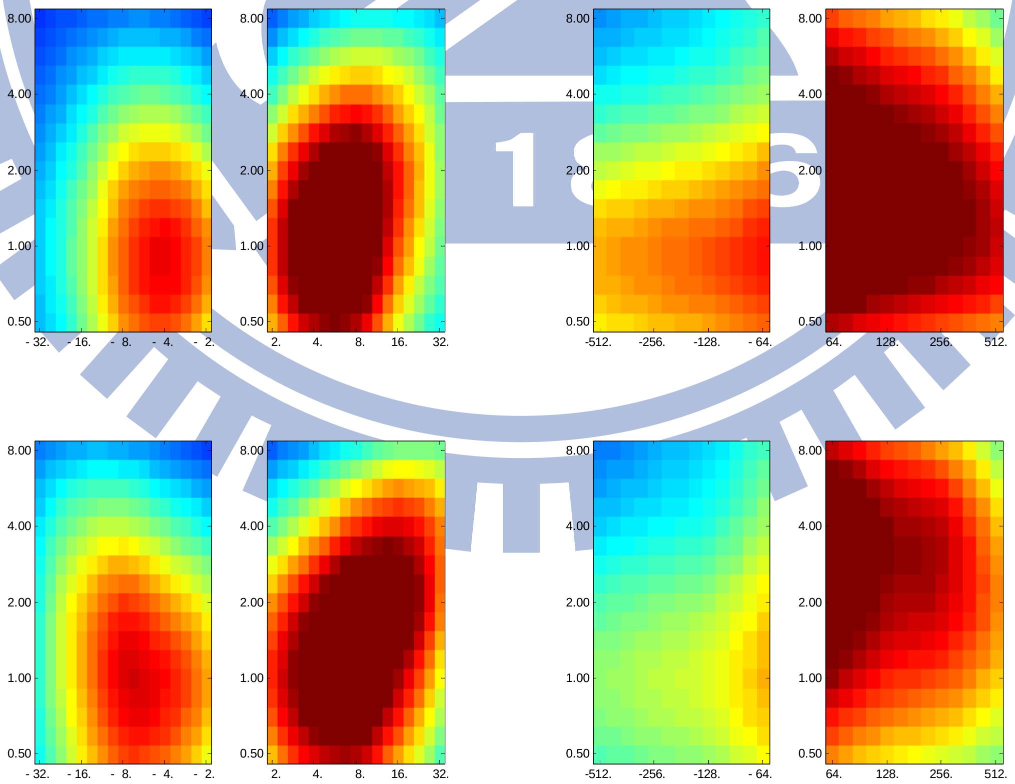
(a) Happy

(b) Anger

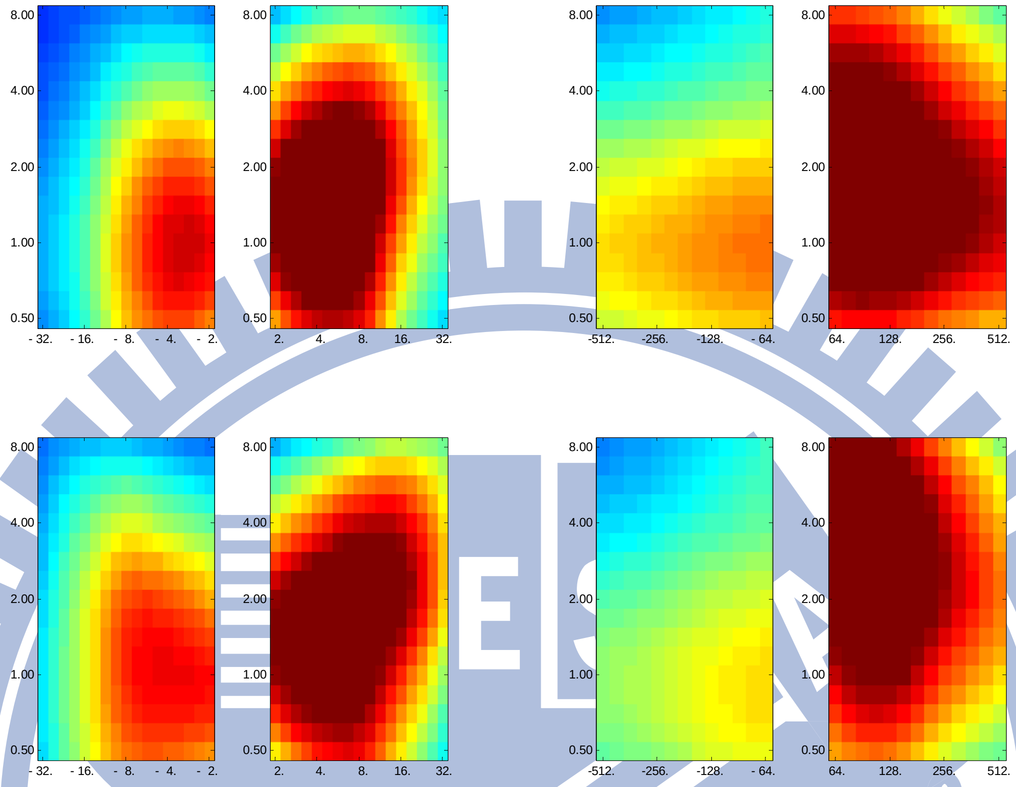
(c) Digust



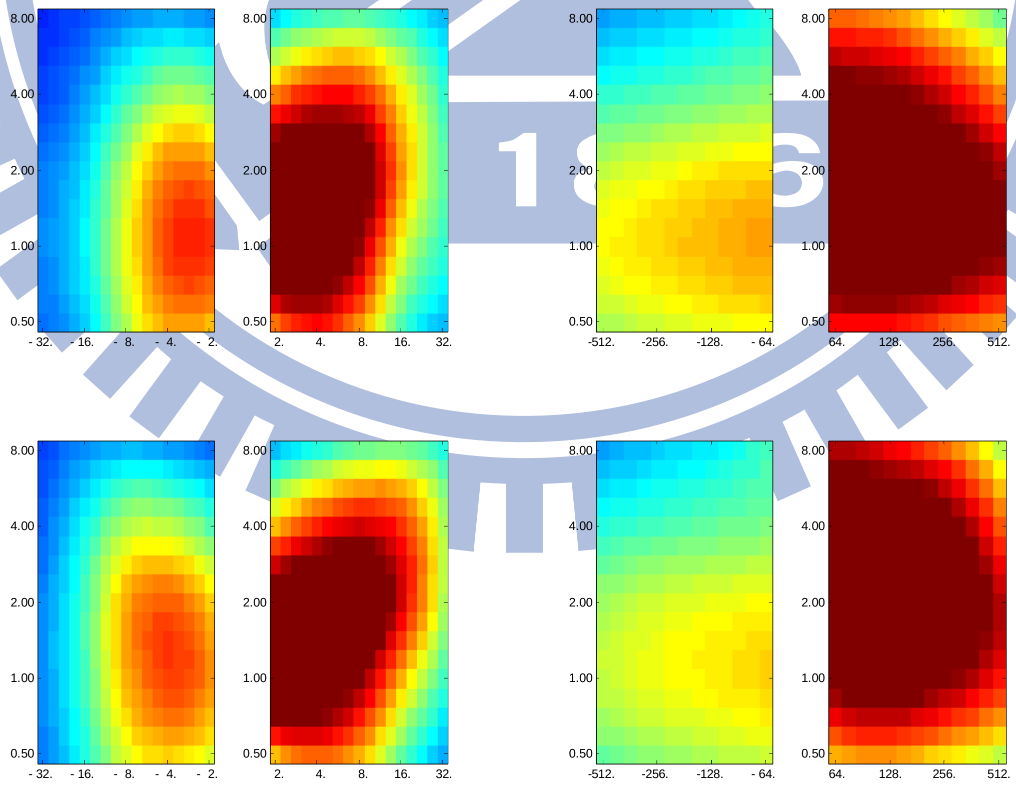
(d) Fear



(e) Neutral



(f) Boredom



(g) Sadness

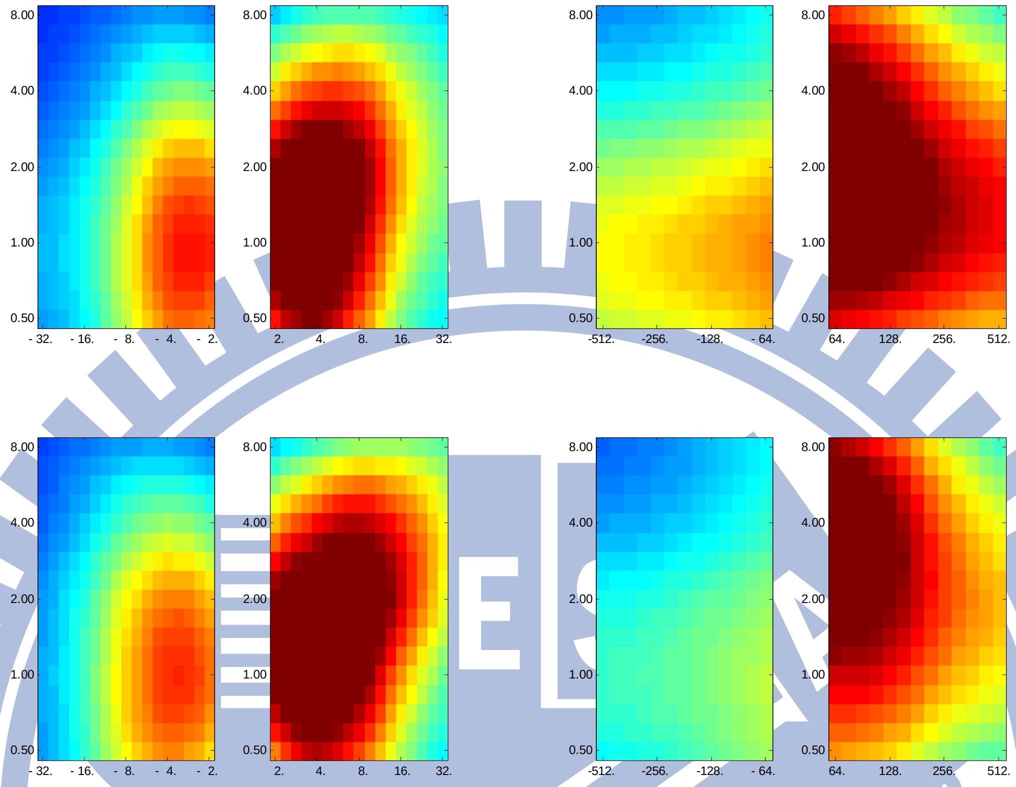


Figure 4.5.1: The RS features of 7 emotions in Berlin Database.

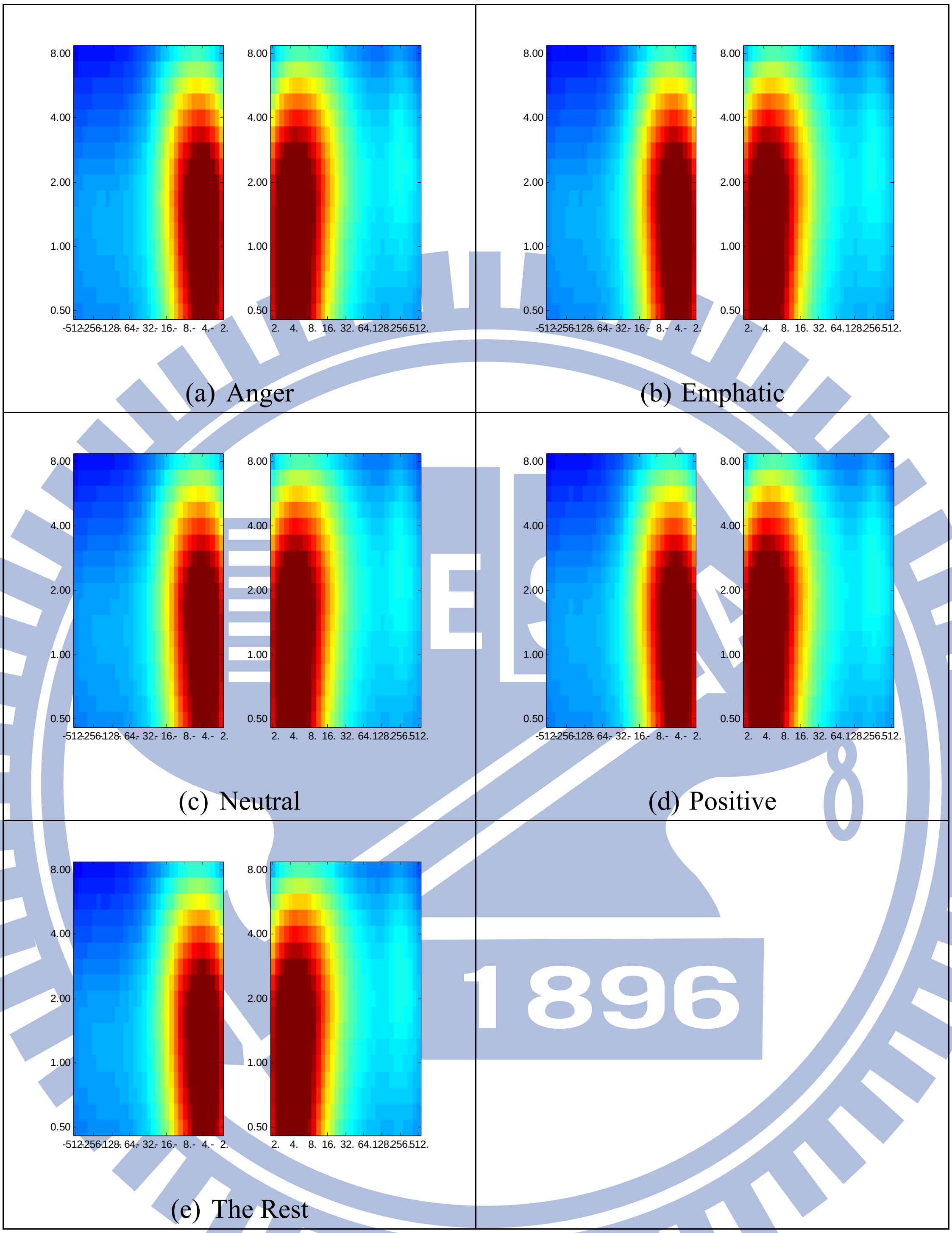
Upper left: RS_{mu} in low rate region; Upper right: RS_{mu} in high rate region.

Lower left: RS_{sd} in low rate region; Lower right: RS_{sd} in high rate region.

The figures were displayed in a zoomed-in fashion in order to highlight the nuance.

4.5.2 Spontaneous Emotions

Spontaneous emotions are intrinsically similar to one another, especially when the emotions are not particularly strong. Another difficulty is that most of the emotions in Aibo Corpus have low validity. In this case, recognizing emotion by analyzing RS plot with bare eyes is barely feasible.



(a) Anger

(b) Emphatic

(c) Neutral

(d) Positive

(e) The Rest

Figure 4.5.2 The RS features of the 5 emotions in Aibo Corpus.

*RSsd plots were not displayed here for they provide equally low recognizibility.

Chapter 5

Future Work and the Big Picture

I always keep in mind that the ultimate goal of this research is to recognize affect in many different situations. Recognizing emotions in speech is just the starting point. Audio cues such as non-speech uttering, visual cues such as facial expression or gestures, text cues, and social context cues are candidates for horizontal expansion. Vertical expansion research including speaker or gender or age normalization continues to be a challenge under current research paradigm in the future.

5.1 Speaker Normalization

Some issues of speaker normalization have to be considered under current paradigm of RS features. The speech samples from speakers of different genders or ages have different "hot areas" on the RS plot. For example, speech samples from male speaker have stronger response in 128-rate region while those from female speaker have strong response in 256-rate region. Response of scales also differs due to speakers. The plight is equivalent to a translation invariant problem. This could be solved if speaker identification or gender recognition is applied prior to emotion recognition.

5.2 Extension to Other Databases

If the learning result can be extended to other databases, the learning is not only successful but also robust to the change of databases. If emotions do have prototypicality, inter-database validation should not be unfeasible. Here are some issues to be overcome:

1. Definition of emotions in different database

   Emotions with the same name in different databases differ from their definition or expressional representation. They may very well share some prototypicality, but they are not the same.

2. Speaker variability

   Databases are constructed by different groups of people, containing different languages, ages, and cultural background.

3. Emotion content

   Different databases include different types of emotions. If the emotions have no union, inter-database trial cannot be practiced.

4. Multimodality and other cues

   As the main modality of emotion expression, vision plays an important role in affect recognition. For example, visual cues of anger and joy are so distinct that they hardly confusion us, but vocal cues of them are sometimes confusing. Besides vision, some database even include blood pressure, heart beat, electroencephalogram (EEG), functional magnetic resonance imaging (fMRI), galvanic skin response (GSR, a.k.a. skin conductance), etc.

5. Interference

   Although contemporarily popular databases seldom contains defect, in the future, databases will be obtained with less control. On the occasion, problems such as missing data, reverberation, packet loss (such as that in voice over Internet protocol), or issues about codices will be met.

As for database, the Danish Emotional Speech Database (Engberg et al., 1997), the Speech under Simulated and Actual Stress Database (SUSAS) (Hansen and Bou-Ghazale, 1997), and the 2002 Emotional Prosody Speech and Transcripts acted database (Hirschberg et al., 2003) are some early audio database; the Vera Am Mittag Corpus (VAM) (Grimm et al., 2008) offers audiovisual data.

5.3 Other Forms of Implementation

5.3.1 Sparer Kernel Machines

In our implementation, the support vector machine (SVM) was the core classifier. However, the kernel was not sparse at all. Most of the samples were support vectors. If the problem itself is highly overlapping or the intrinsic noise is severe, maybe it is wiser to adopt a sparser kernel machine. Relevance vector machine (RVM) is reported to be sparser than SVM and sometimes it even has better performance (Tipping, 2001).

5.3.2 Regression-converted classification

Emotion research differs other classification problems from one aspect: in the process of labeling during the construction of a database, the final label is determined by several expert labelers. Every labeler has their recognition result, and the average result is the validity and the majority label becomes the final ground truth. Now that we have validity of an emotion category, we can implement several regressions and convert their results to a final soft-label classification.

5.3.3 Decomposition

The only thing forbidden under strict condition is the consistency between training and testing conditions because it is not a general solution and has suspicion on being ad hoc. Building up this paradigm of strict condition aims at finding out the most general system or the most robust features to deal with real-life situations.

A graphical model helps to decompose the whole problem. It can take different situations into consideration. This could establish a system having approximately the same performance as that in matched condition.

Incorporating noisy samples into training set is certainly useful for tackling noise, but it also has a risk of contaminating the original clean set. It might cause more intrinsic noise which undermines the recall rate if the noise space is not orthogonal to the signal space. Allowing to include noisy samples into training set is an equivalent justification of applying noise reduction technique in advance.

5.3.4 Boosting

Combining models is a possible way to boost performance. Preliminary experiments have shown that direct addition of a robust feature set did not help the original set to counter the effect of noise. However, combination of several learning machines can possibly boost their total performance as well as robustness. The INTERSPEECH 2009 Emotion Challenge is an evidence for the former; the latter can probably be achieved by similar schemes.

5.3.5 Feature Set

Since the intrinsic dimension estimation showed lower dimensionality for RS features (than that of i384), there may very well be some information that current RS does not extract. Temporal derivatives become a main category of suspicion.

Combination with prosodic or short-term or dynamic features in HTK or other traditions might also have some help. Pitch, on the other hand, is still an issue. Under current extraction scheme, r180 only contains “pitch information” rather than pitch itself. However, before any addition of feature set, robustness should always come in first place.

Whether current RS feature extraction scheme has fully uncovered its potential is a question that requires exploration. In recent reports, features based on mechanisms similar to RS achieve over 80% UR in Berlin Database and over 90% UR when combining with prosodic features (Wu et al, 2010).

5.4 Miscellaneous

5.4.1 Manifold and Dimension Reduction

The issue of manifold directly links the issue of feature reduction. Sometimes the features have very high dimensionality but the samples might distribute on a manifold that has very low intrinsic dimension. If this is the case, then the problem of manifold might be worthy of inspection. A manifold is a topological space that resembles the Euclidean space on a small enough scale. The surface of the Earth is a good example. On the surface of the Earth, as naïve version of our world view, it looks like that the world is flat and we can actually describe any place by two dimensions— longitude and latitude. See Figure 5.4.1 for more examples. Reduction of dimension is not the only merit we can obtain from manifold topology; it also helps to understand the physical meaning of our features.

5.4.2 Discussion about Performance Metric

Classification results are usually shown in the form of confusion matrices. A confusion matrix consists of integer elements because a sample is either misclassified or not. However, in emotion research, even the ground truth is not always certain. We argue that a sample with high uncertainty should not be counted one misclassified sample; rather, it should be counted according to its validity. Therefore, a sample with 60% validity should be counted as 0.6 sample, and the confusion matrix should be a real-valued one as opposed to an integer-valued one.

Another issue about confusion matrix is about data imbalance. In an imbalanced dataset, testing samples are not uniformly distributed through classes. In such a case, precision becomes a biased metric (even useless) because it overestimates the actual precision of the majority and underestimates that of the minority. Whether to normalize the number of each class numerically deserves consideration.

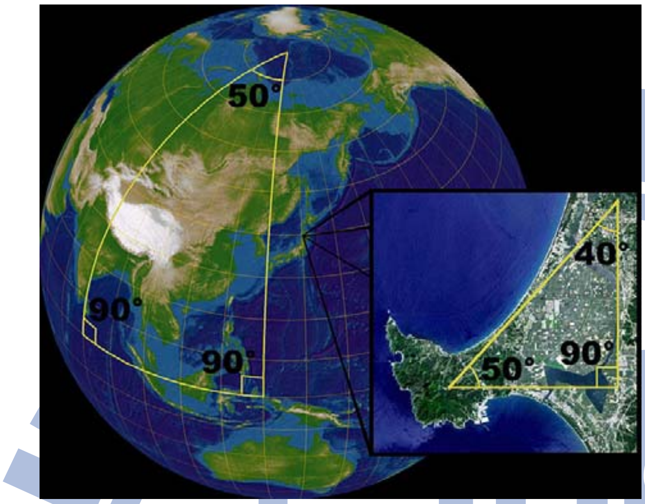


Figure 5.4.1: Surface of the Earth is the best example for manifold.
The sphere (surface of a ball) is a two-dimensional manifold since it can be represented by a collection of two-dimensional maps.

Reference

B

Tanja Bämziger, Klaus R. Scherer, 2005. The Role of Intonation in Emotional Expressions. Speech Communication 46(2005), pp. 252-267.

Bishop, C.M., 2006. Pattern Recognition and Machine Learning.

Bregman AS (1990) Auditory scene analysis: The perceptual organization of sound. MIT press.

Burkhardt, F., Polzehl, T., Stegmann, J., Metze, F., Huber, R., 2009. Detecting Real Life Anger. ICASSP, pp. 4761-4764.

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W., Weiss, B., 2005. A Database of German Emotional, Speech. In: Proc. INTERSPEECH, ISCA, Lisbon, Portugal, 1517-1520.

Busso C, Lee S, Narayanan S (2009) Analysis of emotionally salient aspects of fundamental frequency for emotion detection. IEEE Trans. Audio Speech Language Process., vol. 17, pp 582-596.

C

Carlyon RP, Moore BCJ, Micheyl C (2000) The effect of modulation rate on the detection of frequency modulation and mistuning of complex tones. J. Acoust. Soc. Am., vol. 108, pp 304-315.

Chih-Chung Chang and Chih-Jen Lin, LIBSVM : a library for support vector machines. ACM Transactions on Intelligent Systems and Technology, 2:27:1--27:27, 2011. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm

Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P., 2002, SMOTE: Synthetic Minority Over-sampling Technique, Journal of Artificial Intelligence Research, pp. 321-357.

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

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

Cowie R, Douglas-Cowie E, Tsapatsoulis N, Votsis G, Kollias S, Fellenz W, Taylor JG (2001) Emotion recognition in human-computer interaction. IEEE Signal Processing Magazine, vol 18, pp 32-80.

D

Jesse Davis, Mark Goadrich, 2006. The relationship between Precision-Recall and ROC curves. In: Proc. ICML.

E

Eyben F, Wollmer M, Schuller B, 2009, Speech and music interpretation by large-space extraction. http://sourceforge.net/projects/openSMILE.

F

Johnny R.J. Fontaine, Klaus R. Scherer, Etienne B. Roesch, Phoebe C. Ellsworth, 2007, The World of Emotions Is Not Two-Dimensional, Psychological Science.

Frauke Friedrichs, Christian Igel, 2005, Evolutionary Tuning of Multiple SVM Parameters, Neurocomputing.

G

Isabelle Guyon, André Elisseeff, 2003, An Introduction to Variable and Feature Selection, Journal of Machine Learning Research, pp. 1157-1182.

Grimault N, Bacon SP, Micheyl C (2002) Auditory stream segregation on the basis of amplitude modulation rate. J. Acoust. Soc. Am., vol. 111, pp 1340-1348.

Michael Grimm, Kristian Kroschel, Shrikanth Narayanan, 2008. The Vera Am Mittag German Audio-Visual Emotion Speech Database. In: Proc. ICME, Hannover, Germany, pp. 865-868.

H

He, H., Garcia, E.A., 2009, Learning from Imbalanced Data, IEEE Trans. On Knowledge and Data Engineering.

C.-W. Hsu and C.-J. Lin. A comparison of methods for multi-class support vector machines, IEEE Transactions on Neural Networks, 13(2002), 415-425.

Miles Hewstone, Frank Fincham, Jonathan Foster, 2005. Psychology. Wiley-Blackwell

K

Kawhara, H., Alain de Cheveigné, Banno, H., Takahashi, T., Irino, T., 2005. Nearly Defect-free F0 Trajectory Extraction for Expressive Speech Modifications based on STRAIGHT, Proceedings of INTERSPEECH, pp. 537-540.

Keerthi, S.S., Lin.. C.-J. Asymptotic behaviors of support vector machines with Gaussian kernel. Neural Computation, 15(2003), pp. 1667-1689.

Kockmann, M., Burget L., Cernocky, J., Application of speaker- and language identification state-of-the-art techniques for emotion recognition, Speech Communication, 2011.

L

Loizou PC (2007) Speech Enhancement: Theory and Practice. CRC, New York.

M

Mozziconacci S (2002) Prosody and emotions. In: Proc. Speech Prosody, pp 1-9.

N

Andrew Ng, 2009, Machine Learning, CS229 lecture notes.

New T, Foo S, DeSilva L (2003) Speech emotion recognition using hidden markov models. Speech Comm., vol. 41, pp 603-623.

P

Pereira, C., Watson, C., 1998. Some acoustic characteristics of emotion. In: Proc. ICSLP 98, Sydney, Vol. 3, pp. 927–934.

Rosalind W. Picard, Affective Computing, M.I.T Media Laboratory Perceptual Computing Section Technical Report No. 321

R

Payam Refaeilzadeh, Lei Tang, Huan Liu, *Cross-Validation* in *Encyclopedia of Database Systems*, Springer, 2009.

Frank Rosenblatt, 1962. Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms. Spartan.

S

Scherer, K.R., 2003. Vocal communication of emotion: A review of research paradigm. Speech Communication 40, pp. 227-256.

Schölkopf, N., Smola, A.J., Williamson, R.C., Bartlett, P.L., 2000. *New Support Vector Algorithms*, Neural Computing.

Schuller, B., Arsic, D., Wallhoff, F., Rigoll, G., 2006, Emotion recognition in the noise applying large acoustic feature sets, in Proc. Speech Prosody 2006, Dresden, German, ISCA, p. no pagination.

Schuller, B., Seppi, D., Batliner, A., Maier, A., Steidl, S., 2007, TWOARDS MORE REALITY IN THE RECOGNITION OF EMOTIONAL SPEECH, ICASSP

Schuller, B., Steidl, S., Batliner, A., The INTERSPEESH 2009 Emotion Challenge, Proc. Interspeech, 2009, pp. 312-315.

Schuller, B., Batliner, A., Steidl, S., Seppi, D., 2011, Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge, Speech Communication.

Steidl, S., 2009, Automatic Classification of Emotion-related User States in Spontaneous Children's Speech, Logos Verlag, Berlin.

T

Tipping, M. E., 2001, Sparse Bayesian learning and the relevance vector machine, Journal of Machine Learning Research, pp. 211–244.

V

L.J.P. van der Maaten, E.O. Postma, and H.J. van den Herik. Dimensionality Reduction: A Comparative Review. Tilburg University Technical Report, TiCC-TR 2009-005, 2009.

Varga, A., Steeneken, H.J.M., 1993, 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(3), pp. 247-251.

Vladimir Vapnik, Statistical Learning Theory, 1995

Vladimir Vapnik, Estimation of Dependences Based on Empirical Data, 2006

Ververdis, D., Kotropoulos, C., 2003, A state of the art review on emotional speech databases. In: Proceedings of Interspeech, pp. 2249-2252.

W

Siqing Wu,Tiago H. Falk, Wai-Yip Chan, 2010, Automatic speech emotion recognition using modulation spectral features.

Sheng-Jyh Wang, 2011, Lecture notes of Machine Learning.

Y

Lang-ying Yeh, 2010, Spectro-Temporal Modulations for Robust Speech Emotion Recognition, master thesis

Yeh, L.-Y., Chi, T.S., 2010, Spectro-Temporal Modulations for Robust Speech Emotion Recognition, Interspeech 2010.

Mingyu You, Chun Chen, Jiajun Bu, Jia Liu, Jianhua Tao, 2006. Emotion Recognition from Noisy Speech, ICME.

Z

Zeng, Z, Pantic, M., Roisman, G.I., Huanh, T.S., 2009, A Survey of Affect Recognition Methods: Audio, visual, and Spontaneous Expressions, IEEE Transaction on pattern analysis and machine intelligence.