A UNIFYING THEORY FOR RANK-BASED MULTIPLE CLASSIFIER SYSTEMS, WITH APPLICATIONS IN SPEAKER IDENTIFICATION AND SPEECH RECOGNITION

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
THE MIDDLE EAST TECHNICAL UNIVERSITY

BY

AFŞAR SARANLİ

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR

THE DEGREE OF

DOCTOR OF PHILOSOPHY

IN

THE DEPARTMENT OF

ELECTRICAL AND ELECTRONICS ENGINEERING

JANUARY 2000

Approval of the Graduate School of Natural and Applied Sciences.

Prof. Dr. Tayfur Öztürk
Director

I certify that this thesis satisfies all the requirements as a thesis for the degree of Doctor of Philosophy.

Prof. Dr. Fatih Canatan
Head of Department

This is to certify that we have read this thesis and that in our opinion it is fully adequate, in scope and quality, as a thesis for the degree of Doctor of Philosophy.

Prof. Dr. Mübeccel
Demirekler
Supervisor

Examining Committee Members

Prof. Dr. Mübeccel Demirekler

Prof. Dr. Yalçın Tanık

Prof. Dr. Kemal Leblebicioğlu

Assoc. Prof. Dr. Buyurman Baykal

Assoc. Prof. Dr. Salim Kayhan

ABSTRACT

A UNIFYING THEORY FOR RANK-BASED MULTIPLE CLASSIFIER SYSTEMS, WITH APPLICATIONS IN SPEAKER IDENTIFICATION AND SPEECH RECOGNITION

Saranlı, Afşar

Ph.D., Department of Electrical and Electronics Engineering

Supervisor: Prof. Dr. Mübeccel Demirekler

January 2000, 154 pages

This thesis presents a theoretical investigation of the rank-based multiple classifier decision combination problem and develop a unified framework to understand a variety of such systems.

The rank-based combination is formulated as a discrete optimization problem with the total probability of correct decision as the objective function to be maximized. This formulation introduces a set of classifier observation statistics to be estimated by observing the classifiers operate on a cross-validation test set. The resulting binary programming problem is shown to have a simple global

optimum solution but requiring prohibitive number of observation statistics. To reduce this dimensionality, a method based on observation space partitioning is developed. By this formalism the number of observation statistics can be reduced to levels feasible to estimate from the available cross-validation test data. Specific partitionings can be defined when reasonable assumptions or prior knowledge about the classifiers are incorporated into the problem. Also, certain specific partitionings effectively lead to the Highest Rank, Borda Count and Logistic Regression methods from the literature and establish the links between Type 1 and Type 2 systems.

The concepts of independence and complementariness of combined rank-based classifiers are investigated using basic concepts from Information Theory and measures on independence and complementariness are developed. The Dominance condition is developed as an indicator of performance improvement through combination. Independence of classifiers is shown to have no direct role in classifier complementariness.

Finally the potential of the theory and practical issues in implementation are comparatively illustrated by applying the theory and the existing methods in two real-life pattern recognition problems from speech processing with encouraging results.

Keywords: Statistical Multiple Classifier Systems, Rank-Based Decision Combination Fusion, Classifier Observation Space, Event-Space Partitioning, Pattern Recognition, Independence, Complementariness, Speaker Identification, Speech Recognition

ÖZ

SIRALAMA TEMELLİ ÇOĞUL SINIFLAYICILI SİSTEMLER İÇİN BÜTÜNLEŞTİRİCİ BİR KURAM VE OTOMATİK KONUŞMACI VE KONUŞMA TANIMA UYGULAMALARI

Saranlı, Afşar

Doktora, Elektrik ve Elektronik Mühendisliği Bölümü

Tez Yöneticisi: Prof. Dr. Mübeccel Demirekler

Şubat 2000, 154 sayfa

Bu çalışmada, sıralama temelli çoğul sınıflayıcılı sistemlerde karar tümleştirme probleminin kuramsal bir incelemesi sunulmakta ve bu tür sistemlerin pek çoğunun anlaşılabilmesi için bütünleştirici bir kuram geliştirilmektedir.

Sıralama temelli karar tümleştirme, enyükseklenecek amaç fonksyonunun toplam doğru sınıflama olasılığı olduğu bir kesikli eniyileme problemi olarak formule edilmektedir. Bu formulasyon, sınıflayıcıların bir çapraz-geçerleme veri gurubu üzerinde işlerken gözlemlenmesi ile elde edilmesi gereken bir gurup sınıflayıcı gözlem istatistikleri ortaya koymaktadır. Ortaya çıkan ikili programlama problemi-

nin basit ve küresel bir çözümü olmakla birlikte bunun engelleyici derecede çok gözlem istatistiği gerektirdiği gösterilmektedir. Bu yüksek boyutsallığı daraltabilmek için, gözlem uzayı guruplamaya dayalı bir yöntem geliştirilmektedir. Bu kuramsal yaklaşım altında gözlem istatistiklerinin sayısı, eldeki çapraz-geçerleme veri seti ile kestirilebilecek makul sayılara indirgenebilmektedir. Sınıflayıcılar hakkındaki önceden varolan bilgiler ve makul varsayımlar probleme kaynaştırılarak belli guruplamalar elde edilebilmektedir. Ayrıca, bazı özel guruplamalar, literatürde yer alan En Yüksek Sıra, Borda Sayım ve Logaritmasal Geri Alım yöntemlerini vermekte, ayrıca Tip 1 ve Tip 2 sistemler arasındaki ilişkiyi ortaya koymaktadır.

Çalışmada ayrıca, tümleştirilen sınıflayıcılar arasındaki ilişkisizlik ve tamamlayıcılık kavramları, Bilgi Kuramı'ndan temel prensipler kullanılarak incelenmekte ve bunlar üzerine ölçütler geliştirilmektedir. Tümleştirme ile başarım arttırımı için belirleyici olan Hükmetme Şartı geliştirilmektedir. Sınıflayıcıların ilişkisizliğinin, tamamlayıcılıkta doğrudan bir rol oynamadığı ortaya konmaktadır.

Son olarak, kuramın potansiyeli ve gerçekleme ile ilgili konular, kuramın ve varolan yöntemlerin karşılaştırmalı olarak, konuşma işleme alanından iki gerçekçi probleme uygulanması ile ele alınmakta ve başarılı sonuçlar gözlenmektedir.

Anahtar Kelimeler: İstatistiksel Çoğul Sınıflayıcılı Sistemler, Sıralama-Tabanlı Karar Tümleştirme, Sınıflayıcı Gözlem Uzayı, Olay Uzayı Guruplama, Örüntü Tanıma, İlişkisizlik, Tamamlayıcılık, Konuşmacı Tanıma, Knuşma Tanıma

To my wife Gamze.

ACKNOWLEDGMENTS

I would like to express my sincere gratitude to Professor Mübeccel Demirekler, my advisor and chairman of my dissertation committee. She introduced me to the area of speech processing and pattern recognition and led me to its challenging frontiers. It is her guidance and support that has made this dissertation possible.

I am also grateful to Dr. Buyurman Baykal for his generous sharing with me, his professional and academic experience and his enthusiastic involvement along the course of this research. This has helped much in the shaping of its development.

I would like to thank Professor Yalçın Tanık, member of my dissertation committee, for his critical eye and his challenging questions and comments, which have made this study more complete.

I would also like to thank Professor Kemal Leblebicioğlu, Dr. Tankut Özgen and Assoc. Professor Volkan Atalay for their valuable comments and suggestions along the course of this study. Their encouragement as well as their criticism has contributed to this dissertation.

My Ph.D. work was supported by TÜBİTAK, Turkish Scientific and Technical Research Council, under a scholarship program led by BAYG, The Scientist

Support Group. I am indebted to the people who have supported this study and made it a reality.

Finally, this work could not have been completed without the continuing support of my wife and my family. It is a great joy and the biggest luck to feel this never ending patience and support. I am grateful.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

Introduction

In recent years, it has been observed in pattern recognition literature that there exist many alternative classification and feature extraction methods often using different algorithms, leading to classifier systems with comparable pattern classification performances. However, the performance of each method within such a system is critically dependent on the task domain and controlled operating conditions. Most methods fail when subjected to different tasks or different operating conditions. Therefore, there is no best classifier system for a wide variety of operating conditions even within a single task domain.

Until recently, limitations in computational capacity allowed only one classifier to operate within a system, tuned for specific operating conditions. This severely limits the robustness of such systems. However, the advance in VLSI technology and the accompanied increase in computational capacity now permit the use of a *a set of classifiers* in a cooperative way to build more robust pattern recognition systems.

In recent studies, this trend is reflected in a number of disciplines and it is shown with promising evidence that more than one *complementary classifiers* may be used in parallel as a *multiple classifier system* to tackle the performance and robustness difficulties encountered by a single classifier.

A critical need for such an approach is a *decision combination method* which combines the outputs of suitable individual classifiers in a cooperative way to integrate their strengths while somehow avoiding their weaknesses.

An important group of multiple classifier systems can be called as *Rank-Based Systems*. For this group of systems, individual classifier outputs are the rankings(orderings) of the candidate class labels. This thesis addresses this group and analyses several existing systems with an attempt to develop a unifying theory to understand the behavior of all rank-based multiple classifier systems. The developed theory is then used to analyze the *independence* and *complementariness* between individual classifiers involved in the multiple classifier system as a means of understanding the potential performance improvement possible with combination and selecting a suitable set of classifiers to be used. The introduced theory is then applied in two different pattern recognition problems from speech processing to demonstrate the applicability of the theory to propose better performing rank-based multiple classifier decision combination methods.

I.1 Definitions

A *classifier* is a system which takes a feature or a sequence of features from a pattern and makes a decision about the membership of the pattern in a predefined set of *classes* of such patterns. For example, in a speaker identification task, each

pattern class is a single speaker and membership of a speech pattern to a class means that the pattern was generated by a specific speaker. The decision of the classifier is considered to be correct if the decision is the same as the class known to generate the pattern.

Each classifier implements a *feature extraction method* which derives from a pattern, a descriptor or feature for that pattern. This is often a numerical vector or a sequence of vectors. As an example, in the case of speech processing, features can be derived from the speech signal by spectral analysis of the sampled speech waveform or as the parameters of a linear mathematical model of the speech production process [1].

During classifier operation, features extracted from a pattern are matched against representative *models* of the features making up each candidate class. We call this matching process *similarity scoring* during which the *similarity* of the unknown pattern to each class model is measured.

The *multiple classifier system* is composed of a set of classifiers followed by a decision combiner. The decisions of individual classifiers are input to a *decision combiner* implementing a particular *decision combination method.* The decision combiner generates a single *combined* class membership decision, which is the overall multiple classifier system output. This decision is expected to be more reliable than any individual classifier decision in the system.

The type of outputs by the individual classifiers may assume different levels of information. The information is said to be *rank-level* if the similarity scoring procedure within each classifier is used to rank the candidate classes with respect to their similarities to the test pattern and use this ranking as the output in-

formation. Systems using rank-level information are called *rank-based multiple classifier systems*. These are the focus of the present research and represent an important class of multiple classifier systems.

I.2 Motivation

With the increasingly common observation of complementary behavior between different classifier systems and the accompanied increase in the popularity of multiple classifier systems in the pattern recognition literature, a solid understanding of their behavior becomes increasingly important and interesting.

Fundamental questions such as "How to find an optimal combination method?", "When is it productive to combine a set of classifiers?", "Which classifiers supply redundant information?", "What is the effect of classifier independence on combination performance?", "Which task domains are suitable for a multiple classifier system?" need to be answered. A unifying theory is needed both to attempt answering these questions as well as to establish the relations between seemingly different decision combination methods.

I.3 Research Goals

The primary goal of this research is to develop a theory for rank-based multiple classifier systems which would unify the existing rank-based decision combination methods and to provide a framework where a family of better performing methods can be proposed.

A secondary goal is to investigate the applicability of rank-based multiple classifier systems in speech recognition and attempt to use the proposed theory

in developing a successful system to solve illustrative pattern recognition problems from speech processing.

Several issues arised during the development of the thesis which necessitated to be addressed: (i) the evaluation of a number of rank-based decision combination methods for the speech processing tasks considered, (ii) the development of a theory which would explain these existing methods while providing an open-end for the development for alternative methods, (iii) the use of the developed theory to explore the links between rank-based systems and multiple classifier systems using other levels of output information from classifiers and finally (iv) the use of the developed theory for the exploration of the concepts of independence and complementariness among the set of classifiers involved and to develop measures on the potential performance improvement achievable through combination.

I.3.1 Assessment of Existing Rank-Based Systems

The thesis especially focuses on three methods from the literature: The Highest Rank, the Borda Count and the Logistic Regression Methods. The first two methods are rules derived from group decision theory, simple but frequently used. The last method is a statistical generalization of the Borda Count method based on linear classifier weighting [2]. These have been successfully used in optical character recognition but we show that their performance is not satisfactory for the two problems considered from speech processing.

I.3.2 Development of a Unifying Theory

In an attempt to understand the behavior of these existing methods, the thesis takes a new approach to the rank-based multiple classifier decision making problem and propose the Partitioned Observation Space (POS) Theory of Rank-Based decision combination. This theory is based on interpreting decision combination as a problem from discrete optimization and focuses on methods of formally reducing the problem dimensionality through appropriate partionings of the classifier behavior observation space. The theory unifies the existing methods discussed as special cases of a common framework, hence providing an open way for the development of alternative methods.

I.3.3 Assessment of Independence and Complementariness

Whether or not the joint operation of a set of classifiers will improve classification performance is an important question and depends on how much the information provided by individual classifiers are complementing each other. The thesis proposes a framework where the POS theory is used in conjunction with the information theory to understand the concepts of classifier independence and complementariness and try to relate the concepts with the overall performance gain.

I.3.4 Alternative Decision Combination Methods

Finally, the thesis proposes two more general statistical rank-based decision combination methods based on the POS theory. The performance of these two methods are assessed in two pattern recognition problems from speech processing,

briefly described in Section I.4.

I.4 Experimental Framework

The performance of the rank-based multiple classifier decision methods from the literature as well as the applicability of the unifying POS theory developed is assessed in two real life problems from speech processing. These two problems are briefly introduced in the following sections.

I.4.1 The Closed-Set Text-Independent Speaker Identification Task

Speaker identification is a well studied pattern recognition task domain [3, 4, 5] and several different feature extraction and modeling/similarity scoring methods have been proposed in the literature with varying performances. The speaker identification task in general can be summarized as follows. The speech collected from *known* or *labelled* (sometimes called *client*) speakers are used to build models characterizing each of these speakers. For this purpose, properly sampled speech signal from each speaker is passed from a preprocessing or feature extraction stage where the speech pattern is compressed into a sequence of descriptive feature vectors. During operation, speech sample from an unknown identity speaker is passed from the same preprocessing stage and the feature vectors are matched against each of the speaker models by a similarity scoring procedure. The label of the model which is the most similar to the given feature vectors identifies the speaker.

Closed-set identification means that the decision of the classifier should be one of the class labels and that a decision such as *undecided* is not possible.

Text-independence on the other hand denotes the fact that there is no constraint on the textual content of the speech utterances used both for model training and recognition stages. Speaker identification still remains as a challenging task for researchers, especially for real life problems where there is strong mismatch between training and testing conditions. Such an example is the identification task over the international telephone channels considered for the present study. Here, each speech utterance is recorded over a different long-distance telephone connection and at a different time.

Three different classifiers, consisting of three different feature extraction methods combined with the vector quantization(VQ) modeling based on the Binary-Splitting LBG algorithm are considered to assess the effectiveness of the multiple classifier methods for this task.

I.4.2 The Turkish BDEV Discrimination Task

A second test bed for the thesis is selected again from speech processing. This is the task of discrimination among four highly confusable words, namely the Turkish names of the four letters "B", "D", "E" and "V". These words are difficult to discriminate because the distinguishing sounds are very short in duration and are embedded in uninformative, strong vocalic parts and background noise. The task has been used by Lang as a test bed for the time-delayed neural network architecture for speech recognition [6].

The BDEV discrimination task considered is very similar in nature to the process depicted in Figure VI.1. There are 4 class models, one for each letter-word. Each model is trained and tested with isolated utterances of the corresponding

letter-word recorded from 5 different speakers. The same individual classifiers are used to assess the decision combination methods on this task.

I.5 The Outline of the Dissertation

The thesis begins in Chapter II by a survey of multiple classifier systems in pattern recognition literature and focuses on the rank-based systems research. The main theoretical results of the thesis are covered on two consecutive chapters. The development of a unified theory for this class of systems is introduced in Chapter III. Chapter IV elaborates on the independence and complementariness properties of the set of classifiers used within a multiple classifier system and develops a number of results concerning their effects on combination performance. Chapter V introduces the individual classifiers used for the experiments. Chapters VI and VII presents and discusses the experimental results on the speaker identification and BDEV discrimination tasks respectively. The thesis is concluded in Chapter VIII by a summary of the key results and discussions on future research directions.

CHAPTER II

Multiple-Classifier Systems

Multiple classifier systems in general have been the focus of interest, especially in the last decade, and an interesting body of research have been reported. This Chapter presentes a review of this body of research.

A logical block diagram of a typical multiple classifier system is given in Figure II.1. Here, the decision outputs of individual classifiers are fed into a decision combiner which provides the system's overall decision. The decision combination method implemented by the decision combiner may either be based on heuristic rules reflecting our prior knowledge and assumptions on the joint classifier behavior or on a more sophisticated statistical behavior model derived from the past observations of the system.

A summary of the underlying motivations for the recent interest in such systems is given first. Then a possible subdivision of multiple classifier systems is presented. The multiple classifier systems literature is reviewed and the methods are related to this subdivision of systems without restricting the survey to rank-

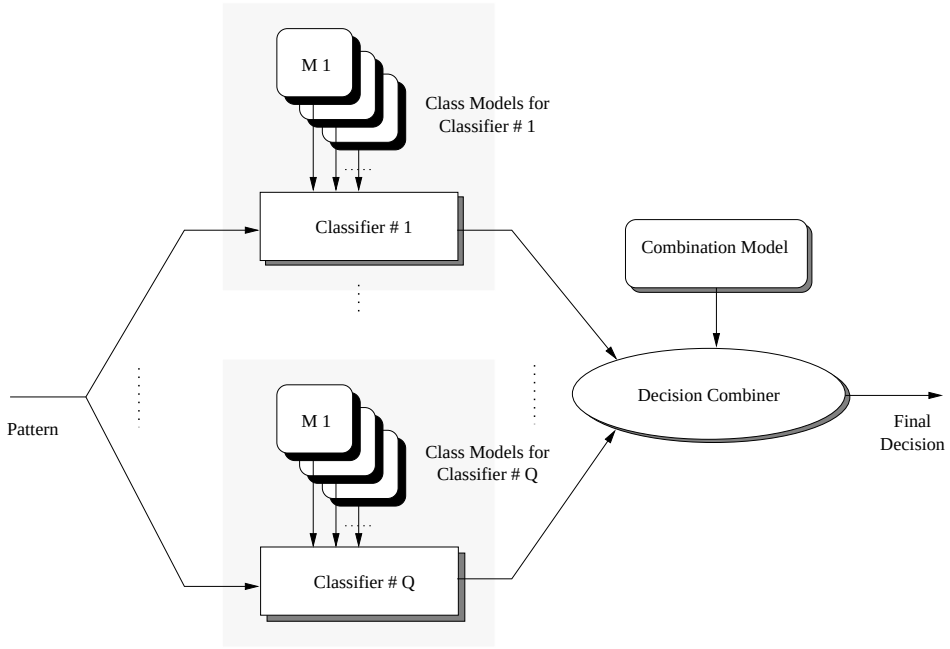

Figure II.1: The multiple classifier system. The individual classifiers operate on the same pattern and provide raw decision outputs. These individual outputs are then used by a decision combiner to arrive at the system's overall decision. The decision combination method implemented by the decision combiner may be based on some heuristic rules or on an optimal statistical model of classifier behavior.

based systems. This helps to understand the place of rank-based system while facilitating the discussion of the thesis on the links between some of these systems. Finally the Chapter focuses on rank-based systems research in particular as the main focus of the present thesis and reviews the existing approaches.

II.1 Multiple Classifier Systems in the Literature

II.1.1 Motivations

In many pattern recognition fields-including among others, speaker identification, speech recognition, optical character recognition and handwritten character recognition-it has recently been observed that there exist various feature ex-

traction and modeling/similarity scoring methods with comparable identification performances. Multiple classifier systems are motivated by the wide availability of such diversity of methods in almost all pattern recognition task domains. Interestingly, none of these methods can be claimed to be the *best one* at all operating conditions. The performance of each one is often maximized within certain controlled operating conditions specific to that method. It has been also often observed that when one method fails, another succeeds, effectively showing a *complementary behavior*.

These observations, supplemented by the recent increase in the digital computation capacity at a declining price, increases the attractiveness of the multiple classifier systems. It now becomes feasible to consider a proper subset of the available classifiers as building blocks for robust real-life classification systems.

The aforementioned motivations are justified by a frenzy of activity in pattern recognition literature following some early contributions [7, 8]. Many attempts have been made to build multiple classifier systems in various pattern recognition fields. These include, among others, machine printed word/character recognition [2], handwritten character recognition [9, 10, 11, 12, 13, 14, 15, 16], speaker recognition, [17, 18, 19, 20], face identification [21, 22], text-to-phoneme translation [23], remote sensing and geographical data processing [24, 25], military target recognition [26] and biomedical image processing [27, 28]. The neural networks community has also been active on this subject [29, 30, 12, 31, 23, 32, 33, 34, 13]. Indeed, the idea of combining multiple sources of evidence for decision making has been a topic of extensive research within the artificial intelligence literature for a long time. Two main approaches which still form the backbone of the methods

found in pattern recognition literature are the Bayesian Approach [35, 36] and the Dempster-Shafer Approach [37, 38]. The diversity of the fields in which the problem has been considered and encouraging results which have been reported show that multiple classifier decision systems are of considerable interest to a large number of pattern recognition fields.

II.1.2 Types of Classifiers

In [9], Xu et.al. propose an interesting taxonomy of multiple classifier decision systems with respect to the level of information available at the outputs of the individual classifiers.

A typical classifier is illustrated in Figure II.2. Here, the feature extraction stage is illustrated as an integral part of the classifier since most classifiers have their own specific feature extraction methods. Often, only the feature extraction method is the distinguishing factor between classifiers. The classifier internally produces two different levels of information from which it finally derives a single class decision as its output. Most of the widely used classifiers in a variety of disciplines with the exception of some others (e.g., a pure syntactic classifier [9]), have this general form of operation. If such intermediate information is made available by individual classifiers, then they may be used as the input to a decision combiner in order to reach a collective decision.

According to [9], Type 1 systems are defined as multiple classifier systems where only the final decision from each classifier is available for combination. Type 3 systems require that all the similarity measures between the unknown pattern and the models be available for combination. Finally Type 2 systems

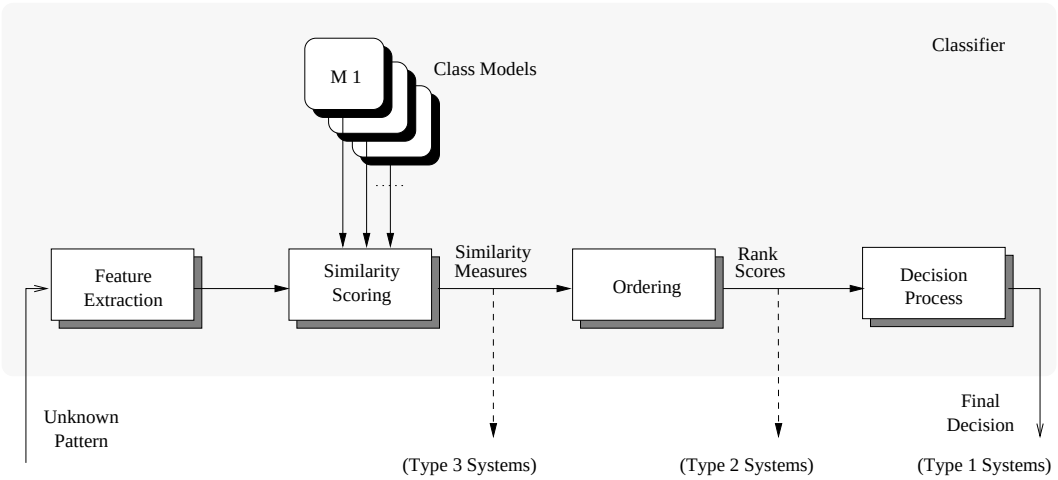

Figure II.2: Operation of a typical classifier. Most classifiers use a feature extraction method to transform the patterns into descriptive features and then use a similarity scoring method to match these features to a set of models for known classes. The similarity scores are ranked from highest to lowest similarity and the class model with the highest similarity score is identified.

constitute an intermediate level where the ranking of the candidate models according to similarity to the unknown input is transformed into discrete scores and used for combination. These different information levels are illustrated in Figure II.2.

As will be discussed and laid out in the present study, the clear distinction between Type 2 and Type 3 systems is perhaps not so important. These type of combination systems can be shown to constitute the two extreme cases of a unifying treatment. Although the thesis preserves the distinction between Type 2 and Type 3 systems, it will be shown that Type 3 systems are in fact a special case of rank-based systems.

II.1.3 Systems Based on Final Decisions: Type 1

Followed by the taxonomy of classifiers, in [9], Xu et.al. briefly discuss Type 3 multiple classifier systems. They analyze the so called *Averaged Bayes Classifier* and some of its versions and show that these can be generalized to other types of classifiers (such as the distance classifiers) as long as the similarity scores can be transformed into probabilities. They concentrate on the study of Type 1 systems, claiming that this group is the most general since all classifiers can at least supply information at this level. In this category, they discuss voting methods and develop a general expression which can describe most voting method variants. They then move into statistical methods and treat the decision combination problem under the Bayesian formalism. However, the development of the combination formulas rely entirely on the independence assumption between individual classifiers. This assumption may sometimes be unrealistic and therefore, its effects in multiple classifier systems are tried to be clarified by the present thesis in Chapter IV. They finally use the Dempster-Shafer theory of evidence [37, 38], which has been first considered in [39] for the same task, and propose an efficient computational model for its use. The experimental results on tasks from unconstrained handwritten numerals show performance gain for all types of combination methods with the best performance being from the statistical combination methods. However, they note the fact that all methods are based on the independence assumption and that the experiments reflect only the gain on a specific task. Furthermore, they ask the question which reflects one motivation for the present thesis, i.e., *whether it is possible or not to generalize the presented*

approaches for dependent classifiers? The unifying theory in Chapter III does not involve an independence assumption and in addition, the effects of such an assumption under the developed theory are investigated in Chapter IV.

In [10], Huang and Suen propose a method of joint statistical modeling of the space of classifier final decisions. This method which is called the *behavior knowledge space method* is shown to outperform the Bayesian and Dempster-Shafer approaches which both rely on the independence assumption. It is argued that the focus of attention may be shifted from designing one best classifier to designing lower performing but complementary behaving classifiers to maximize combined performance and that a representative data set is essential to benefit from the method since the model is statistical. The method presented in [10] is for Type 1 classifiers and can be shown to be a special case of the formulation in the present thesis.

Woods et.al., propose a rather different approach. Instead of operating more than one classifiers in parallel, they propose dynamically selecting which individual classifier's decision should be the collective decision [40]. In this approach called as *dynamic classifier selection with local accuracy*, whenever individual classifiers disagree, a k-nearest neighbors based local accuracy is computed for each classifier and the decision of the one with the highest local accuracy is selected. The local accuracy is computed either as the percentage of correctly classified training patterns(overall local accuracy) or, assuming that a test sample is assigned to class S_j, as the percentage of correctly labelled samples that have been assigned to class S_j (local class accuracy).

The idea of building a system by combining the results of a large number

of very simple classifiers which can do little better than random guessing, is first studied by Kleinberg under the title *stochastic discrimination* [41] and later applied to pattern recognition problems by Ho and Kleinberg [42]. In a later study by Ji and Ma [43], this idea is applied to combine weak perceptron classifiers with a simple majority voting technique with the aim of improving the space-time complexity as opposed to designing and using a single very complex classifier. These suggest that decision combination approaches are also very promising for building complex pattern classifiers from collections of simple, even automatically generated classifiers.

II.1.4 Systems Based on Continuous Similarity Scores: Type 3

Although Xu et.al., claim that they are not general enough, Type 3 systems seems to be of general interest for a variety of pattern recognition applications and especially within the neural networks research community. In such systems, the continuous similarity scores which are often class posterior probabilities are processed by the decision combiner to arrive at a final class decision.

In [44], Wolpert discusses the dynamics of the generalization process and later in [23], proposes a layered neural network architecture scheme called *stacked generalization* with the aim of modeling and correcting the generalization errors made by the individual classifiers. Cross-validation data generation techniques such as the *leave-one-out* method [45] are used to build a second stage neural network which would map the continuous outputs of the first stage networks to the final continuous output. The method is applied to text-to-phoneme translation task. One problem with this approach is that the corrective mapping is done solely

in a data induced manner and does not allow one to integrate prior knowledge and assumptions about classifier behavior into the corrective mapping.

In [13], Rogova extends the use of Dempster-Shafer theory for the combination of Type 1 classifiers to neural networks applications with Type 3 classifiers. They argue that complementariness among classifiers is an important issue reflected by experiments and combination of the outputs for different feature extraction methods may give better performance than the combination of the outputs for different similarity scoring methods.

Tumer and Ghosh [29, 46] discuss that given infinite training data, neural network outputs approximate the Bayesian decision boundaries by means of approximating the class posterior probabilities to arbitrary precision. Therefore they claim that such classifiers achieve similar performances for a given classification task. However, they also argue that finite data and factors such as different network initializations and noise, cause a neural network classifier to loose its generalization performance by a deviation from the optimal Bayesian decision boundary. They consider *linear* and *order-statistics combiners* where a number of Type 3 network outputs are mapped into a single real valued output either by averaging or by order-statistics and analyze the change in the deviation(error) from the Bayesian decision boundary. They demonstrate that the error variance diminishes during such a combination operation. They also conclude that correlation among the outputs of individual classifiers have a negative effect on the improvement achievable through a combination process. Also, in [47], they use the concepts to estimate the Bayesian error rate from the outputs of a set of classifiers.

Similar considerations form the basis of a previous work by Perrone and Cooper [31]. In their work, a linear combiner with network output averaging is considered within the context of neural networks for regression analysis. The methods introduced are termed as *ensemble methods*. It is argued that the mean squared error in the approximation is reduced by such a combination. They also consider the *generalized ensemble method* as the weighted averaging of individual network outputs and propose an optimal weight selection method. Another interesting discussion is on cross-validation data generation methods which the present thesis make use of for the training of the decision combiner. Neural network output independence is again one major assumption during the development of the work presented.

Another example is the study by Benediktsson and Swain [25] where two groups of methods are discussed, within the framework of multi-source remote sensing and geographical data processing. They are termed as *consensus theoretic methods*. The so called *linear opinion pool* is in fact a linearly weighted version of the averaged Bayes classifier of [9] and the linear combiner of [29]. The *logarithmic opinion pool* is a variant to overcome some of the weaknesses of the linear version but relies again on the independence of the individual classifier. They argue that the selection of the weights for such combiners is an open problem where only ad-hoc methods exist in the literature. Indeed, in their later study in [24], an optimal weight computation scheme is proposed to form the system called the *parallel consensual neural network*.

Kittler et.al. consider in [48] and [15], the *product rule* and the *sum rule* as a means of classifier combination and provide a sensitivity analysis with respect

to estimation errors. Product rule assumes statistical independence among the classifiers while the sum rule assumes that the posterior probabilities do not deviate much from the prior probabilities.

A comparison between the classifier combination rules with probability averaging and probability product with statistical independence assumption is given by Tax et.al., [14]

II.1.5 Rank-Based Systems: Type 2

Although not all classifiers can supply information additional to the final class decisions, as discussed in the previous sections, a considerable number of them can do so. Moreover, using such additional information which is often discarded for the operation of a single classifier is expected to be useful for classifier decision combination [9, 2]. However, in some pattern recognition problems, Type 3 systems with continuous similarity scores as individual classifier outputs may suffer from incompatibility of their similarity scores for combination and this fact may limit their usefulness [2].

Ranking of the classes with respect to an unknown pattern is an information which can be derived from the full similarity scores and does not suffer from such incompatibility problems. It is a useful intermediate level of information in between the full similarity scores and the single class labels. The advantages of using ranking information for classifier combination is discussed in detail by Ho [49] where a number of rank-based multiple classifier systems are discussed within the framework of machine printed character and visual word recognition.

Unfortunately, there are relatively few attempts to analyze and understand

the rank-based systems in the literature. One major contributor is Ho et.al., [49, 2] but their attempt remains mostly experimental and does not attempt an in-depth theoretical analysis of such systems.

In [49] and [2], Ho et.al. discuss a number of methods under two categories, namely *class set reduction* and *class set reordering*. Since class set reduction methods loose the ordering of the classes in the reduced set, they often lead to plural decisions and therefore are not as useful as the reordering methods [1]. Reordering methods preserve the ordering and therefore a final decision can be made on the resulting ranking. This property makes them of a more general interest. By the reordering method, a better ranking of the candidate classes is aimed.

One of the methods of rank-based multiple classifier decision making under this category is the *Highest Rank* method. Another one is the *Borda Count* method, introduced by Black in the context of social decision making [50] and discussed by Ho, within the framework of pattern recognition [49]. With an attempt to generalize the Borda Count method so that the classifiers can be treated according to their contributions to the combination, Ho proposes a *Logistic Regression* method which results in the optimal linear weighting of the classifier rank scores. These three methods are discussed in some detail in Section II.2 and they are also subject to an experimental assessment in Chapters VII and VI.

Al-Ghoneim an Kumar [32] also contributes to this group of multiple classifier

[1] There exist more specialized applications such as in forensic science where it may be interesting to reduce the the number of candidate classes for a recognition task to ease the task of the human operator. For example, in a fingerprint recognition task, to reduce the number of potential criminals from the huge number in the criminal database to a potential few tens may be of tremendous advantage for the human operator doing such a task manually.

systems by proposing a new training algorithm for individual classifiers. With this, individual performances are maximized not only with respect to their final decision but also with respect to their first two higest ranks so that their contribution to the rank-based decision combination method improves.

II.2 Rank-Based Multiple Classifier Systems

In the previous section, an overview of the literature on multiple classifier decision systems is presented and the major contributions are summarized. Also, a brief introduction is made to the rank-based multiple classifier decision combination methods which have been proposed in the literature. In this section, a more detailed description of these rank-based methods will be provided.

Now consider a general classification problem where an unknown pattern x is to be classified as a member of one of the P classes $S_1, S_2, \cdots, S_P$. Moreover, suppose this classification is done in parallel by Q separate classifiers $X_1, X_2, \ldots, X_Q$, relying on different feature extraction methods and possibly by different similarity scoring methods. Each such classifier X_k orders all the P known classes with respect to their similarities to the unknown pattern x and transforms this ordering into a set of P integer scores, known as *rank scores*. This scoring process is illustrated in Figure II.3. In this scoring, the score of a candidate class starts from zero (when the class is at the bottom of the ranking) and increases up to $(P-1)$ as the class approaches the top of the ranking.

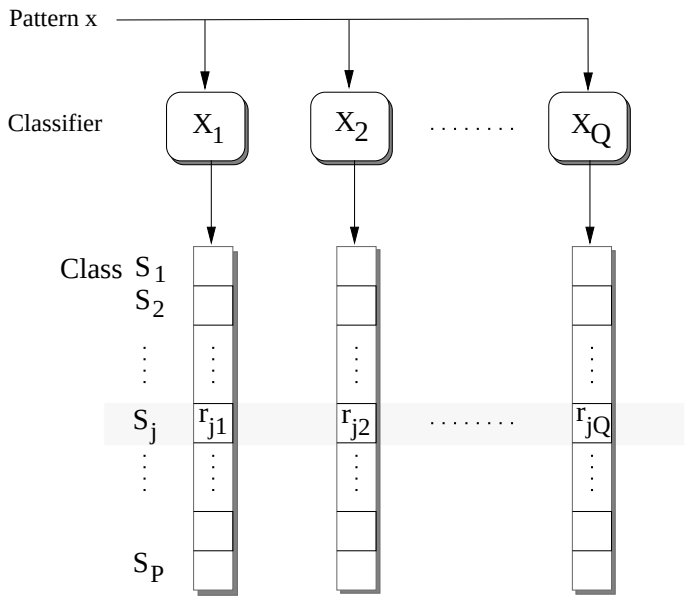

Figure II.3: The rank scoring process by a set of classifiers. Each classifier orders the candidate classes according to the similarity to the unknown pattern. This ordering is then transformed into integer *rank scores* which form a score vector from each classifier.

II.2.1 The Highest Rank Method

In this method, the rank scores which are assigned to a candidate class by all classifiers are transformed into a *max-score* for that class. For each candidate class S_j, $j = 1, 2, \cdots, P$, the scores from all the classifiers $r_{j1}, r_{j2}, \cdots, r_{jQ}$ are considered and the maximum value from this set of scores is assigned as the max-score of this class. These new scores for all candidate classes form the *max-score vector*. The decision is made by selecting the class with the *maximum max-score*. This process is illustrated in Figure II.4.

One problem with this method occurs when the number of candidate classes is small. The computed max-scores may have *score collision*, i.e., more than one class may have the same max-score. In this case, the method cannot provide a unique decision and the collision should be resolved by random selection or by

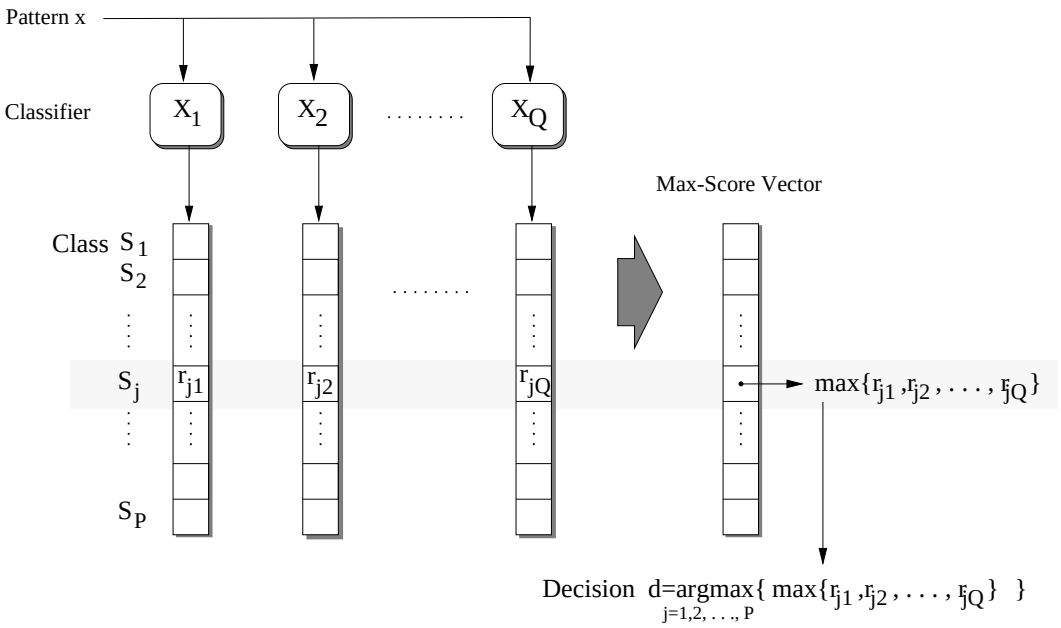

Figure II.4: The Highest Rank rank-based decision combination process. The maximum of the rank scores given to a class by all classifiers is assigned to the entries of a new vector called the *max-score vector*. The class having the maximum max-score entry is the final decision of the system.

some other methods. For the Borda Count method which will be described in the following section, this problem is not so severe.

II.2.2 The Borda Count Method

For this second method of rank-based decision combination, the rank scores from all classifiers for a candidate class are added together to generate a *sum-score* for that class. The sum-scores for all candidate classes form the *sum-score vector*. The decision is made by selecting the class with the *maximum sum-score*. The Borda Count decision combination process is illustrated in Figure II.5.

This method does not suffer from the score collision problem as severely as the Highest Rank method since the number of possible sum scores is much higher. Nevertheless whenever such a collision occurs, there is nothing the method can

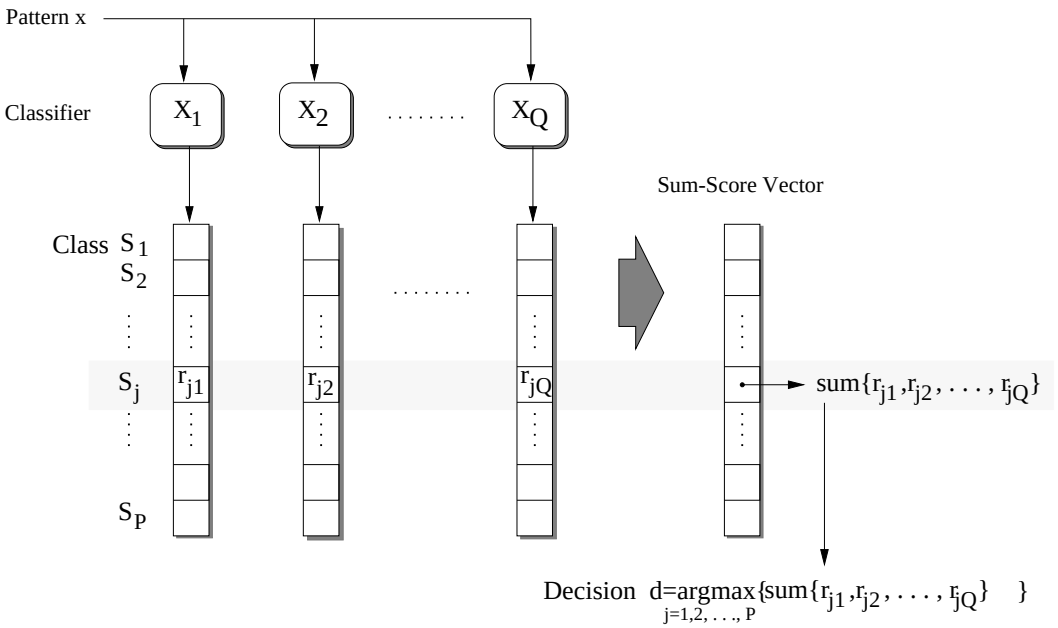

Figure II.5: The Borda Count rank-based decision combination process. The sum of the rank scores given to a class by all classifiers is computed and assigned to the entries of a new vector called the *sum-score vector*. The class having the maximum sum-score entry is the final decision of the system.

further do and the decision should again be done either by random selection or by some other method, among the colliding classes.

Clearly, both of these methods treat all the classifiers involved in the decision combination process equally, without taking into account the differences in their individual performances. It is often observed that these individual performances may be considerably different, rendering such uniform treatment illogical.

II.2.3 The Logistic Regression Method

This method is proposed by Ho [2] in an attempt to compensate for the non-uniform performances of the classifiers. This is a problem present in both the Highest Rank and the Borda Count methods. It is mainly a generalization of the Borda Count method by the introduction of optimal linear weights to the

addition operation of the rank scores involved.

First a binary indicator variable Y_{S_j} is defined. For a given pattern x, $Y_{S_j} = 1$ if S_j is the correct class of that pattern and $Y_{S_j} = 0$ if S_j is the wrong class. Therefore, when an unknown pattern is given, the objective is to predict the value of this binary variable for each class. This suggests that the problem can be formulated as a function approximation task. Assume that $\mathbf{r}_j = [r_{j1} r_{j2} ... r_{jQ}]^T$ represents the vector of rank scores for a particular class S_j generated by the classifiers $X_1, X_2, ..., X_Q$ respectively. Next, the term

$$\pi_j(\mathbf{r}_j) = P(Y_{S_j} = 1 | \mathbf{r}_j), \tag{II.1}$$

is defined as the *probability of the class S_j to be the true class of the pattern x* given the set of rank scores $\mathbf{r}_j$ resulting from the processing of the pattern x. This probability is expected to increase monotonically as the rank scores from individual classifiers assigned to that particular candidate class increase. Due to this expectation, one can define a *logistic function* of the form

$$\pi_j(\mathbf{r}_j) = \frac{\exp(\alpha_j + \beta_{j1} r_{j1} + \beta_{j2} r_{j2} + \cdots \beta_{jQ} r_{jQ})}{1 + \exp(\alpha_j + \beta_{j1} r_{j1} + \beta_{j2} r_{j2} + \cdots \beta_{jQ} r_{jQ})}, \tag{II.2}$$

with an attempt to linearize the parameter optimization problem. By making use of this logarithmic transformation, a linear function of the rank scores can be defined. Using this function, the terms called as the *log-odds* are defined by

$$L_j(\mathbf{r}_j) = \log\left(\frac{\pi_j(\mathbf{r}_j)}{1 - \pi_j(\mathbf{r}_j)}\right) = (\alpha_j + \beta_{j1} r_{j1} + \beta_{j2} r_{j2} + \cdots \beta_{jQ} r_{jQ}), \tag{II.3}$$

where $j = 1, 2, \cdots, P$. In this relation, the P terms at the left hand side can be estimated from the cross-validation data and the parameters $\alpha_j, \beta_{j1}, \cdots, \beta_{jQ}$ should be determined so that the estimated terms can be optimally predicted from the corresponding rank scores.

However, in the original development by Ho, the dependence on the class index j is dropped without convincing justification and the number of optimal prediction problems is reduced to one with the number of parameters reducing to $\alpha, \beta_1, \cdots, \beta_Q$. Thus, Eq. (II.3) becomes

$$L(\mathbf{r}) = \log\left(\frac{\pi(\mathbf{r})}{1-\pi(\mathbf{r})}\right) = (\alpha + \beta_1 r_1 + \beta_2 r_2 + \cdots \beta_Q r_Q). \qquad \text{(II.4)}$$

By using the cross-validation data, empirical values or estimates for the $\pi(\mathbf{r})$ terms are obtained from all sets of classifier rank scores. This data is then transformed into the form of *log-odds* and used in association with the corresponding classifier rank scores to estimate the classifier decision combination linear parameters $\{\alpha, \beta_1, \beta_2, ..., \beta_Q\}$. This rather confusing process of data generation for the optimal weight determination is illustrated in Figure II.6.

This procedure which is a linear regression analysis is called the *logistic regression method* because of the logarithmic transformation involved in order to linearize the optimization problem. Ho claims that the values of the parameters represent the relative significances of the classifiers in the combination process. The optimization criterion is the *minimum classification error for the cross-validation data set.*

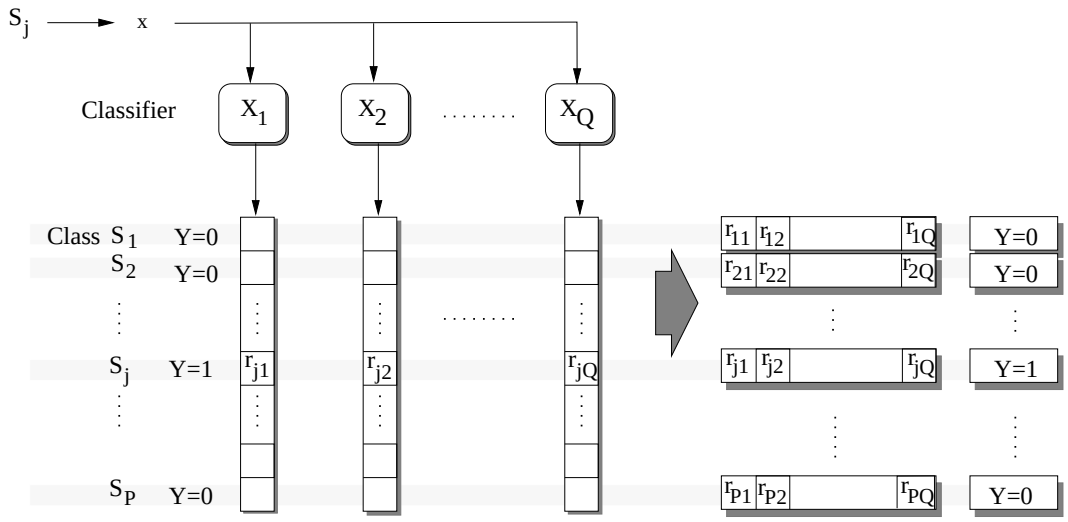

Figure II.6: Data generation for determining the optimal linear weights for the Borda Count method. Each pattern of the cross-validation data (with known identity class label) is processed with all classifiers to generate the rank scores. For each such pattern, P number of $(\mathbf{r}, Y)$ sample pairs are obtained and accumulated to predict the actual $\pi(\mathbf{r})$ values.

II.3 Multiple Classifier Systems in the Speech Literature

Although they are very interesting and challenging task domains of pattern recognition, speech and speaker recognition unfortunately have not been areas where contributions have been made to the multiple classifier decision systems literature. Most of the small number of studies present in the speech literature using multiple classifier systems mainly *make use of* systems developed in the context of other task domains.

In an early study, Soong and Rosenberg [51] consider combining two classifiers, one being based on the cepstrum features and the other based on the delta-cepstrum features of the speech signal for speaker recognition. They use the linear opinion pool for the combination.

Farell and Mammone [17] also consider the linear and logarithmic opinion

pools as tools for classifier combination for speaker identification. They consider two classifiers from two different categories, namely classifiers using *supervised* (neural tree network) and *unsupervised*(vector quantizer) learning. They also argue that only heuristic solutions exist to the weight selection problem involved in these methods.

On the other hand, Bennani and Gallinari [18] consider speaker identification and verification with neural networks and discuss classifier combination as a problem of building hierarchical neural networks. Their so called *modular connectionist system* uses *expert typology detectors*, each a lower level feed forward neural network, which specialize on the discrimination of speakers in a specific typology. The expert decisions are then combined into the overall system decision by within Bayesian probabilistic framework.

In a recent study, Radova and Psutka [20] consider two group of classifiers, one based on *dynamic time warping* (DTW) and the other based on the *cepstral distance measure*. Each individual classifier in a group makes a class decision based on a *specific vowel* among 5 vowel sounds considered. The utterances are segmented to expose the vowel sounds. They consider decision combination by voting methods and the Borda Count method both for combination within individual classifier groups and their full collection and report promising results.

CHAPTER III

A Unifying Theory of Rank-Based Multiple Classifier Systems

III.1 Introduction

When a single classifier is considered, a final class decision (the identified class) is obviously the only desired output. However, for decision combination with multiple classifier outputs, using only a single class label from each classifier output may lead to a loss of valuable information. It should be advantageous to use classifier output forms carrying more information.

The problems of output incompatibility, incomparability and scaling are of concern for using classifier outputs in the measurement form but such problems do not occur when using classifier outputs in the form of candidate class rankings. It is an observed fact that when a pattern is misclassified, the true class of the pattern is often close to the top of the ranking [49, 54]. The resulting ranking gives valuable information about classifier behavior for imperfect classifiers.

Despite this fact and that there have been good theoretical attempts to analyze the properties and behavior of Type 1 and Type 3 decision systems [9, 25, 31, 30, 12, 23], there have been few attempts to develop an understanding of the rank-based combination systems [2, 49, 32].

In this chapter, the rank-based multiple classifier decision combination problem is considered and a unifying theoretical formulation called *Partitioned Observation Space* (POS) *Theory* is developed. During this development, the multiple classifier decision combination is treated as a discrete optimization problem and a controlled smoothing technique is developed to reduce problem dimensionality. The thesis shows that a number of rank-based methods discussed in Section II.2 can be analyzed and formulated as special cases within this unifying framework. Finally, the links with Type 1 systems are established and it is illustrated that in fact, these can also be shown to be special cases of rank-based systems.

III.2 A Binary Integer Programming Approach

III.2.1 Notation and Problem Formulation

The thesis considers a closed-set pattern classification problem [1] where patterns may belong to P different classes $S_j, j = 1, 2, \cdots, P$ which is referred as *candidate classes*. It is assumed that there are Q classifiers $X_q, q = 1, 2, \cdots, Q$ involved in the classification process. Furthermore, x denotes a *pattern*, i.e., the smallest token composed of feature vectors processable inside the classifiers to generate candidate class rankings.

[1] A closed-set pattern identification problem is the case whenever one does not have an option to remain *undecided* . Therefore it is necessary to decide on a unique class label. The opposite case is the *open-set* problem where being undecided is allowed.

Each classifier ranks all the candidate classes according to some internal measure and generates a *rank score* r_{ji} for each such candidate. The rank score for a specific candidate class is defined as *the number of candidate classes placed after it by the classifier in the generated ranking.* With this definition, as the class is placed close to the top of the classifiers' rankings of the candidates, it receives higher rank scores. The source class of an unknown pattern (i.e., the true class generating the pattern x) is represented as an integer valued random variable $\underline{s}_x$ [2] taking index values of an ordered set of class labels $\mathcal{S} = \{S_1, S_2, \cdots, S_P\}$. Hence the fact that a pattern comes from a generating class S_j is denoted by a realization of this random variable as $(\underline{s}_x = j)$. The final decision at the output of the decision combiner is denoted by another integer valued random variable $\underline{d}$ with the same possible values as $\underline{s}_x$. When an unknown pattern x arrives, the pattern is processed by all classifiers. Each classifier ranks all P candidate classes and generates P *rank scores*, namely one for each candidate class. The set of all rank scores generated by the classifiers for all candidate classes form a $P \times Q$ matrix $\mathbf{R}$ which is referred as the *rank score matrix.* Here, each column corresponds to scores by a single classifier while each row corresponds to scores by all classifiers for a single candidate class. By definition, for rank-based multiple classifier decision systems, the combined decision must be done solely using this rank score matrix.

Suppose that the ultimate objective in classifier combination is to achieve the maximum rate of correct classification. Although other objectives can also be defined, this is a meaningful and usual choice for closed-set pattern classifica-

[2] Throughout the thesis, underscore notation denotes a random variable.

tion. If we define a binary valued random variable $\underline{y}$, as an *indicator* of correct classification,

$$\underline{y} = \begin{cases} 1 & \text{if } (\underline{d} = j) \text{ given } (\underline{s}_x = j), \\ 0 & \text{otherwise,} \end{cases} \tag{III.1}$$

then the problem of finding the optimum rank-based multiple classifier decision combination method can be expressed as an *optimization problem* with an objective function being the *total probability of correct classification* as

$$\max P\{\underline{y} = 1\}. \tag{III.2}$$

III.2.2 The Objective Function

The objective function implied by Eq. (III.2) is not useful in this form. It should contain free problem parameters and also statistics reflecting the joint ranking behavior of the classifiers. Let the probability of correct classification be expanded into a sum over the source class and rank score matrix indexes as

$$P\{\underline{y} = 1\} = \sum_{j=1}^{P} \sum_{n=1}^{N} P\{\underline{d} = j, \underline{s}_x = j, \underline{r} = n\}, \tag{III.3}$$

where we have used the fact that $(\underline{y} = 1)$ is equivalent to $(\underline{d} = j)$ once the source class is realized as $(\underline{s}_x = j)$. By using the Bayes rule, the joint probability inside the double summation can be decomposed to give

$$P\{\underline{y} = 1\} = \sum_{j=1}^{P} \sum_{n=1}^{N} P\{\underline{d} = j | \underline{s}_x = j, \underline{r} = n\} P\{\underline{s}_x = j, \underline{r} = n\}. \tag{III.4}$$

By definition, the optimal decision method to be found uses only the available information from individual classifiers, which is the rank score matrix. Therefore, the decision is a deterministic function of $\underline{r}$, as $\underline{d} = f(\underline{r})$. Using the dependence on the rank score matrix only, the first term of Eq. (III.4) can be simplified as

$$P\{\underline{d} = j | \underline{s}_x = j, \underline{r} = n\} = P\{\underline{d} = j | \underline{r} = n\}. \tag{III.5}$$

Hence the objective function expansion in Eq. (III.4) takes the final form

$$P\{\underline{y} = 1\} = \sum_{j=1}^{P} \sum_{n=1}^{N} P\{\underline{d} = j | \underline{r} = n\} P\{\underline{s}_x = j, \underline{r} = n\}. \tag{III.6}$$

The two set of terms inside this expansion should be investigated. The first set of conditional terms $P\{\underline{d} = j | \underline{r} = n\}$ is directly linked with our decision method. When a specific multiple classifier decision method is specified, these terms can be determined. For the case of a deterministic decision method, they are binary valued with possible values 0 and 1. The joint probability terms $P\{\underline{s}_x = j, \underline{r} = n\}$ of the second set on the other hand are independent of the decision method and model the joint output behavior of the classifier ensemble. Since they can be expressed as $P\{\underline{r} = n | \underline{s}_x = j\} P\{\underline{s}_x = j\}$, these joint probabilities can be estimated if the trained classifiers are operated on a sufficient body of cross-validation data consisting of known source class identities coupled with the resulting rank score matrices. As discussed in Chapter VI, such a body of cross-validation data can be obtained by using the leave-one-out method on the training data.

III.2.3 The Optimum Solution and The Optimum Decision Method

In order to develop a methodology for finding the optimum rank-based multiple classifier decision method[3] based on the observation of the classifier ensemble behavior on the cross validation data, we need an unambiguous interpretation of the results obtained so far with respect to finding an optimum decision method. From the expansion in Eq. (III.6), it is seen that the only terms dependent on the decision method are $P\{\underline{d} = j|\underline{r} = n\}$. Any specific decision method manifests itself as a large set of specific 0 and 1 values associated with these probabilities. *Conversely, any specific assignment to this set of probabilities constitutes a specific decision method.* Therefore, if these terms are treated as binary valued free problem parameters b_{jn}, the objective function in Eq. (III.6) becomes

$$P\{\underline{y} = 1\} = \sum_{j=1}^{P} \sum_{n=1}^{N} b_{jn} P\{\underline{s}_x = j, \underline{r} = n\}. \tag{III.7}$$

When this is combined with the optimization criterion, the problem can be expressed as

$$\max_{b_{jn}, j=1,2,\cdots,P; n=1,2,\cdots,N} \left\{ \sum_{j=1}^{P} \sum_{n=1}^{N} b_{jn} P\{\underline{s}_x = j, \underline{r} = n\} \right\}, \tag{III.8}$$

$$\text{Subject to} \quad \sum_{j=1}^{P} b_{jn} = 1 \quad \text{for} \quad n = 1, 2, \cdots, N. \tag{III.9}$$

where the set of constraints arises from the fact that the final output of the decision method should be a single class label. Since all $P\{\underline{s}_x = j, \underline{r} = n\}$ are non-negative, the obvious solution to this optimization problem is given by

[3] Within this framework, an *optimal decision method* based on the outputs of multiple classifiers is equivalent to an *optimal classifier combination method*. Therefore, these two terms can be used interchangeably.

$$b_{jn}^{*} = \begin{cases} 1 & \text{if } j = \underset{k=1,2,\cdots,P}{\operatorname{argmax}} P\{\underline{s}_x = k, \underline{r} = n\}, \\ 0 & \text{otherwise,} \end{cases} \tag{III.10}$$

For $j = 1, 2, \cdots, P$ and $n = 1, 2, \cdots, N$. However, this global maximum may not be unique, specifically when the maximizing j value in Eq. (III.10) is not unique for a fixed $(\underline{r} = n)$. For such cases, more than one set of b_{jn}^{*} values constitute alternative solutions.

Each optimal solution set $\{b_{jn}^{*}\}$ corresponds to a unique global optimum *decision method*. To elaborate, consider the case where an unknown pattern x is processed by an ensemble of classifiers and a rank score matrix $\mathbf{R}$ is generated. This specific rank score matrix is a realization $(\underline{r} = n)$ where the value of n is fixed. The optimal solution guarantees that there is a single variable b_{jn}^{*} which is 1 for fixed $(\underline{r} = n)$ and the global optimum solution is a *decision method* with the decision criteria

$$\underline{d} = \underset{j=1,2,\cdots,P}{\operatorname{argmax}} P\{\underline{s}_x = j, \underline{r} = n\} \tag{III.11}$$

Denoting the global maximum value of the objective function by $P_{\max}$, the classification error rate of the optimal decision method is given by $P_E = (1 - P_{\max})$. The presence of multiple global optimum solutions signify that the overall system is unable to discriminate among certain classes when faced with specific classifier outputs. For closed-set problems, a random choice among these solutions has to be made.

III.2.4 The Curse of Dimensionality

In the previous section, it is shown that *a globally optimum method of combining any number of classifiers, based on their rank scores, can be found if we were able to obtain infinite number of observations of all the classifier rank-level outputs.* However, in practice, only a finite cross-validation data set is available, from which the values of $P\{\underline{s}_x = j, \underline{r} = n\}$ are to be estimated. The reliability of these estimated probabilities are very important for the performance of the optimum decision method.

The number of such probabilities for P classes and Q classifiers is $P(P!)^Q$, which is a prohibitive number for most practical applications. Therefore, although the solution to the problem is simple, it cannot be of practical value if a method of reducing the number of estimated statistics cannot be found. The next section attempts to formulate such a method.

III.3 Dimensionality Reduction by Partitioning

Consider the objective function expansion in Eq. (III.7). The problem domain is composed of two main parts. The first one is the space spanned by the free problem parameters b_{jn} and will be called as the *Problem Parameter Space*. The second one is the space spanned by all the statistics about the joint behavior of the classifiers, i.e., all estimated probabilities $P\{\underline{r} = n, \underline{s}_x = j\}$ and will be called as the *Classifier Observation Space*. *Classifier Observation Statistics* are the elements of this latter space.

This prohibitive cardinality of the classifier observation space is the limiting

factor for the usefulness of the resulting solution since it determines the number of statistics to be estimated from the available data. However, well behaving classifiers do not span the entire observation space. As the performance of the classifiers improve to acceptable levels, the cross-validation samples tend to be highly clustered. Therefore, enough data is accumulated for estimating certain statistics while there is no data available to estimate some very small probabilities. Consider the extreme case of ideal classifiers. Whenever a pattern is supplied, the classifiers will generate very similar rankings with the true source class being at the top of the ranking. For such an extreme behavior, all cross-validation data will be accumulated for a small number of statistics. As the classifiers deviate from ideal behavior, such clusters tend to spread. The implication is that by incorporating the expected or observed behavior of the classifiers for a certain task, we may introduce some assumptions about the possible distribution of the cross-validation data and arrive at considerable compressions of the classifier observation space by means of forming logical groups of statistics. This compression can also be interpreted as a *smoothing* operation on the estimates we are trying to find. Limited cross-validation data necessitates such a smoothing and one may argue that the available data determines the *resolution* with which the classifiers can be observed. However, one should bear in mind that with such a smoothing, the system becomes unable to model the observation data violating the assumptions made and the optimal solution to this smoothed problem is sub-optimal with respect to the original one.

III.3.1 An Observation Event Space

Define an *event space* $\mathcal{F}$ where the realizations of the source class label and rank score matrix indexes are combined into compound events. The most basic events (*event atoms*) in this space are defined as $(\underline{s}_x = j; \underline{r} = n)$. Here the event atom specifies occurrence of the joint event "The source class for the pattern x was S_j and the set of classifiers generated the rank score matrix $\mathbf{R}_n$". This event space is clearly finite and its cardinality is $P \times (P!)^Q$ since we have P source classes and Q classifiers and hence a total of $(P!)^Q$ possible rank score matrices [56].

Now assume that a *mapping* $\mathcal{W}$ partitions this event space into disjoint sets of event atoms. Also assume that the mapping (or partitioning)[4] $\mathcal{W}$ results in $M_{\mathcal{W}}$ such sets, i.e., $M_{\mathcal{W}}$ partitions $W_1, W_2, \cdots, W_{M_{\mathcal{W}}}$ which satisfy the standard properties,

$$\begin{aligned} &\text{(a)} \quad W_i \cap W_j = \emptyset \qquad\qquad \forall i, j \in \{1, 2, \cdots, M_{\mathcal{W}}\}, \\ &\text{(b)} \quad W_1 \cup W_2 \cup \cdots \cup W_{M_{\mathcal{W}}} = \mathcal{F}. \end{aligned} \tag{III.12}$$

Such a partitioning defines a new event space in which each resulting partition defines a *new basic event*. Composed of a set of original event atoms, these are also *compound events* in the original event space $\mathcal{F}$. If such partitions(or events) W_i form an ordered set $G_{\mathcal{W}} = \{W_1, W_2, \cdots, W_{M_{\mathcal{W}}}\}$ which is our new event space, the partitioning $\mathcal{W}$ effectively defines a new random variable,

$$\underline{g}_{\mathcal{W}} : \mathcal{S} \times \mathcal{R} \longmapsto \{1, 2, \cdots, M_{\mathcal{W}}\}. \tag{III.13}$$

[4] Since each such mapping defines a new partitioning of this space, within this context a mapping $\mathcal{W}$ is synonymous with a partitioning $\mathcal{W}$ and the terms *partitioning* and *mapping* will be used interchangeably.

whose values are the index values of the ordered set $G_{\mathcal{W}}$. Here $\mathcal{S}$ is the set of possible source classes and $\mathcal{R}$ is the set of possible rank score matrices.

At this point, a new expansion for the objective function in Eq. (III.2), can be developed. First, the objective function is expanded over the source class labels and rank score matrices to obtain Eq. (III.3). Following Eq. (III.13), the random variable $\underline{g}_{\mathcal{W}}$ can be expressed as a $\underline{g}_{\mathcal{W}} = \mathcal{W}(j,n)$. Therefore, the double sum can also be written by introducing the new random variable $\underline{g}_{\mathcal{W}}$ as

$$P\{\underline{y}=1\} = \sum_{j=1}^{P}\sum_{n=1}^{N} P\{\underline{d}=j, \underline{s}_x = j, \underline{r}=n, \underline{g}_{\mathcal{W}} = \mathcal{W}(j,n)\} \tag{III.14}$$

This is possible since the value of $\underline{g}_{\mathcal{W}}$ is known once the values of $\underline{s}_x$ and $\underline{r}$ are known and no new probabilistic event is introduced. By successively using the Bayes rule, the joint probability inside the double summation can be put into the form

$$\begin{gathered} P\{\underline{d}=j, \underline{s}_x = j, \underline{r}=n, \underline{g}_{\mathcal{W}} = \mathcal{W}(j,n)\} = \\ P\{\underline{d}=j | \underline{s}_x = j, \underline{r}=n, \underline{g}_{\mathcal{W}} = \mathcal{W}(j,n)\} P\{\underline{s}_x = j, \underline{r}=n | \underline{g}_{\mathcal{W}} = \mathcal{W}(j,n)\} \\ \cdot P\{\underline{g}_{\mathcal{W}} = \mathcal{W}(j,n)\}. \end{gathered} \tag{III.15}$$

Again, using the fact that the decision should, by definition, be based on the rank score matrix alone, and inserting Eq. (III.15) into Eq. (III.14) we obtain the final expression for the objective function as

$$P\{\underline{y}=1\}=$$

$$\sum_{j=1}^{P}\sum_{n=1}^{N}P\{\underline{d}=j|\underline{r}=n\}P\{\underline{s}_x=j,\underline{r}=n|\underline{g}_{\mathcal{W}}=\mathcal{W}(j,n)\}P\{\underline{g}_{\mathcal{W}}=\mathcal{W}(j,n)\}. \tag{III.16}$$

This time there are three sets of terms inside the expansion. The first set of terms is, as before, the one directly associated with the decision method, yet unknown. The last set of terms on the other hand, is again the statistics about the behavior of the classifiers. However, this time, the observation space is the result of the partitioning $\mathcal{W}$ and the observable events for modeling the classifier behavior are the resulting partitions $W_m, m=1,2,\cdots,M_{\mathcal{W}}$ represented by the possible values of the random variable $\underline{g}_{\mathcal{W}}$. This is a coarser resolution for the classifier observation space where the actual rank score matrices are hidden inside the observable partitions. The middle set of terms, which is new as compared with the previous expansion, is a set of *transition terms* defining the relation between the coarser resolution of the partitions and the finer resolution of the class label, rank score matrix pairs. Since a deliberate decision is made to set the observation resolution to the coarser one, by definition, there is no cross-validation data left to estimate these transition terms. In fact, due to this choice, one is *ignorant* about this finer detail. These terms allow us to formally introduce our ignorance within the Bayesian formalism [35]. as a uniform distribution among the individual elements of any partition W_m; i.e., we have

$$P\{\underline{s}_x = j, \underline{r} = n | \underline{g}_{\mathcal{W}} = m\} = \begin{cases} 0, & \text{if event atom } \{\underline{s}_x = j, \underline{r} = n\} \notin W_m, \\ \frac{1}{|W_m|}, & \text{if event atom } \{\underline{s}_x = j, \underline{r} = n\} \in W_m, \end{cases} \tag{III.17}$$

where $|W_m|$ is the cardinality of the partition W_m.

The previous discussion makes it clear that the first set of terms will again be labelled as the problem parameters to find an optimum decision method. The last terms, on the other hand, will again be estimated from the classifier behavior on the cross-validation data.

With this new expansion, a controlled tool to selectively decrease resolution on the observation and modeling of the classifier ensemble behavior is obtained. By the selection of the partitioning, it is possible to reduce the number of partitions, hence the number of events in the observation space. Specifically, for the above expansion we have $M_{\mathcal{W}}$ statistics to estimate. For limited cross-validation data, a reduction in the number of statistics to estimate, corresponds to an increase in the number of observations about each individual statistic, hence to an increase in the reliability of the estimate. It is well known that the reliability of the estimates is crucial to the generalization performance, hence to the classification performance of the system [23, 45].

Although we have mentioned that the number of observable events can be arbitrarily reduced, the nature of the partitioning is crucial for the usefulness of the resulting solution. The objective should be to maintain the maximum observation resolution which is reasonable for the fixed amount of data available,

and not a finer one. It is also illogical to use a very coarse resolution while enough data for a finer one is available since this over-smoothing, being unable to properly model the collective behavior of the set of classifiers, will limit the usefulness of the approach.

The motivation in compressing the observation space is not only the reduction in the number of statistics to estimate. For some cases, one may have an intuition about the pattern recognition task at hand, or prior knowledge about the dynamics of individual classifiers which may suggest that some higher level of resolution in the observation space may in fact be *irrelevant* or *unreliable*. This leads to the intuitive formulation of suitable partitionings, for which a clear example is presented in Section III.3.3.

In the present chapter, we consider three specific partitionings to unify three methods from the literature and one to establish the links with Type 1 systems. However, a challenging and yet open problem is the selection of a partitioning in an optimal manner based on the actual distribution of the data. E.g., using Genetic Algorithms to determine an optimal partitioning rule may be a promising direction for future research.

III.3.2 A Computational Model to Implement the Optimal Solution

The terms $P\{\underline{g}_{\mathcal{W}} = \mathcal{W}(j,n)\}$ in the new Eq. (III.16) makes it clear that the number of unique statistics is $M_{\mathcal{W}}$, a number necessarily lower than the original $P \times (P!)^Q$. However, the number of terms inside the double summation is still $P \times (P!)^Q$. Even with this huge number of terms inside the expansion, the optimum solution presented in Section III.2.3 may be converted into an algorithmic form so

that only a small number of computations is necessary for making the optimum decision based on the estimated statistics. Considering the optimum solution given in Eq. (III.10), this algorithmic form can be summarized as follows: For each pattern, process the pattern by all classifiers and obtain the rank score matrix. For the fixed index ($\underline{r} = n$), compute exactly P objective function coefficients $P\{\underline{d} = j|\underline{r} = n\}P\{\underline{s}_x = j, \underline{r} = n|\underline{g}_{\mathcal{W}} = \mathcal{W}(j,n)\}P\{\underline{g}_{\mathcal{W}} = \mathcal{W}(j,n)\}$; i.e., one coefficient is computed for each candidate class, using the corresponding estimates and the transition terms determined by the partitioning rule. The final step is to decide on the class with the maximum coefficient. A total of at most P multiplications and a search is involved. A final random choice is necessary for the closed-set problem if there is more than one maximum coefficient for this $\underline{r}$ value. Note that the determination of the transition terms is only possible if the partitioning is based on a rule which can be easily applied when the rank score matrix is given. For an arbitrary partitioning, a look-up table of the original size would be required, which is infeasible.

III.3.3 A Sensible Partitioning: First Two Ranks

The implications of a partitioning choice are discussed in Section III.3.1. One may often have prior insight into the task and classifiers involved before the observation statistics are collected. This may be incorporated into the solution by means of a partitioning of the observation space.

The partitioning we call as the *First Two Ranks* is based on such a general observation about the behavior of the classifiers. Assume that *the resolution below the topmost two ranks (largest two rank scores) is unreliable.* This is a

reasonable assumption, e.g., for distance classifiers, since the separation between class models becomes less significant as the models become more and more distant from the pattern being classified in the feature space. Hence, noise on the feature vectors has a greater chance to disturb the lower ranks. Therefore, we decide not to discriminate among the ranks lower than the second and group them to represent the *last rank*. The resulting new score assignment (hence partitioning) can be expressed as

$$\hat{r}_{jq} = \begin{cases} r_{jq} - P + 3, & \text{if } \; r_{jq} > P - 3, \\ 0, & \text{if } \; r_{jq} \leq P - 3. \end{cases}$$

An illustrative example for $P = 4$ classes and $Q = 2$ classifiers is considered in Table III.1, where there is a total of $M_{\mathcal{W}} = 576$ partitions instead of the original 2304 event atoms. A shortcut notation is used as follows: Each row is the summary of four actual rows which correspond to the source classes S_1, S_2, S_3, S_4 summarized as $S_{1,2,3,4}$ in the table.

III.4 Special Cases by Means of Specific Partitionings

In this section, the theory will be used to unify three existing rank-based decision combination methods discussed in the literature. It will be shown that they indeed correspond to specific partitionings, hence for a given data their optimality can be analyzed by the introduced theory.

Table III.1: Illustration of the First Two Ranks partitioning. The first column is the actual partition contents, the second column is a label for the partition and the last column is the partition random variable. Actual class labels and rank score matrices are used instead of their corresponding random variables for clarity.

$$\left\{\left(S_{1,2,3,4},\begin{bmatrix}3&3\\2&2\\1&1\\0&0\end{bmatrix}\right);\left(S_{1,2,3,4},\begin{bmatrix}3&3\\2&2\\1&0\\0&1\end{bmatrix}\right);\left(S_{1,2,3,4},\begin{bmatrix}3&3\\2&2\\0&1\\1&0\end{bmatrix}\right);\left(S_{1,2,3,4},\begin{bmatrix}3&3\\2&2\\0&0\\1&1\end{bmatrix}\right)\right\}\mapsto\left(S_{1,2,3,4},\begin{bmatrix}2&2\\1&1\\0&0\\0&0\end{bmatrix}\right)\mapsto \underline{g}_{\mathcal{W}}=1,2,3,4$$

$$\left\{\left(S_{1,2,3,4},\begin{bmatrix}3&3\\2&1\\1&2\\0&0\end{bmatrix}\right);\left(S_{1,2,3,4},\begin{bmatrix}3&3\\2&0\\1&2\\0&1\end{bmatrix}\right);\left(S_{1,2,3,4},\begin{bmatrix}3&3\\2&1\\0&2\\1&0\end{bmatrix}\right);\left(S_{1,2,3,4},\begin{bmatrix}3&3\\2&0\\0&2\\1&1\end{bmatrix}\right)\right\}\mapsto\left(S_{1,2,3,4},\begin{bmatrix}2&2\\1&0\\0&1\\0&0\end{bmatrix}\right)\mapsto \underline{g}_{\mathcal{W}}=5,6,7,8$$

$$\vdots \qquad\qquad \vdots$$

$$\left\{\left(S_{1,2,3,4},\begin{bmatrix}1&1\\0&0\\2&2\\3&3\end{bmatrix}\right);\left(S_{1,2,3,4},\begin{bmatrix}1&0\\0&1\\2&2\\3&3\end{bmatrix}\right);\left(S_{1,2,3,4},\begin{bmatrix}0&1\\1&0\\2&2\\3&3\end{bmatrix}\right);\left(S_{1,2,3,4},\begin{bmatrix}0&0\\1&1\\2&2\\3&3\end{bmatrix}\right)\right\}\mapsto\left(S_{1,2,3,4},\begin{bmatrix}0&0\\0&0\\1&1\\2&2\end{bmatrix}\right)\mapsto \underline{g}_{\mathcal{W}}=573,574,575,576$$

III.4.1 The Highest Rank Method

Highest Rank method is discussed in [2] and is a simple technique of rank-based multiple classifier decision. Since this technique does not use any model of the observed classifier behavior it is not optimum for the general case. In this section, the conditions under which the Highest Rank solution coincides with the optimum solution will be established by means of a specific partitioning. The Highest Rank method may be described as follows: "For each source class, select the highest of the rank scores assigned by all classifiers for that class as a new score. These new scores constitute a *max score* vector. The class with the maximum *max score* is selected." The status of this method with respect to our formalism can be summarized by two facts.

FACT III.1 *As a predefined decision method, the Highest Rank method coincides with a specific fixed set of values for the problem variables b_{jn}. The cases where*

Table III.2: Illustration of the partitioning establishing the relation with the Highest Rank method.

$$
\begin{array}{ll}
\left\{\left(S_{1,2,3};\begin{bmatrix}2&2\\1&1\\0&0\end{bmatrix}\right)\right\}\mapsto & \left(S_{1,2,3};\begin{bmatrix}2\\1\\0\end{bmatrix}\right)\mapsto \underline{g}_{\mathcal{W}}=1,2,3\\
\left\{\left(S_{1,2,3};\begin{bmatrix}2&2\\1&0\\0&1\end{bmatrix}\right);\left(S_{1,2,3};\begin{bmatrix}2&2\\0&1\\1&0\end{bmatrix}\right)\right\}\mapsto & \left(S_{1,2,3};\begin{bmatrix}2\\1\\1\end{bmatrix}\right)\mapsto \underline{g}_{\mathcal{W}}=4,5,6\\
\vdots & \vdots \qquad\qquad \vdots\\
\left\{\left(S_{1,2,3};\begin{bmatrix}2&0\\1&0\\0&2\end{bmatrix}\right);\left(S_{1,2,3};\begin{bmatrix}2&0\\1&1\\0&2\end{bmatrix}\right);\left(S_{1,2,3};\begin{bmatrix}2&0\\0&1\\1&2\end{bmatrix}\right);\right. & \\
\left.\left(S_{1,2,3};\begin{bmatrix}1&2\\0&1\\2&0\end{bmatrix}\right);\left(S_{1,2,3};\begin{bmatrix}0&2\\1&1\\2&0\end{bmatrix}\right);\left(S_{1,2,3};\begin{bmatrix}0&2\\1&0\\2&1\end{bmatrix}\right)\right\}\mapsto & \left(S_{1,2,3};\begin{bmatrix}2\\1\\2\end{bmatrix}\right)\mapsto \underline{g}_{\mathcal{W}}=10,11,12\\
\vdots & \vdots \qquad\qquad \vdots\\
\left\{\left(S_{1,2,3};\begin{bmatrix}0&0\\2&1\\1&2\end{bmatrix}\right);\left(S_{1,2,3};\begin{bmatrix}0&0\\1&2\\2&1\end{bmatrix}\right)\right\}\mapsto & \left(S_{1,2,3};\begin{bmatrix}0\\2\\2\end{bmatrix}\right)\mapsto \underline{g}_{\mathcal{W}}=40,41,42
\end{array}
$$

the Highest Rank method is unable to make a decision because of more than one maximum in the sum score matrix correspond to more than one fixed sets of b_{jn} *values.*

FACT III.2 *The Highest Rank method does not make use of any observation on classifier ensemble behavior and hence is not in general optimum for the combination of non-ideal classifiers.*

For the Highest Rank method, a partitioning $\mathcal{W}$ can be defined where each partition corresponds to a possible *max score vector*. Consider the illustrative case of $P = 3$ source classes and $Q = 2$ classifiers. There are a total of $M_{\mathcal{W}} = 42$ partitions, some of which are illustrated in Table III.2 where the summary notation introduced in Section III.3.3 is used.

The equivalence relation between the optimum solution resulting from the objective function expansion with respect to the partitioning $\mathcal{W}$ and the Highest

Rank solution can be stated as a theorem.

THEOREM III.1 *The Highest Rank method and the optimum solution to the classifier combination problem coincides in the context of maximizing the probability of correct decision expressed as in Eq. (III.16) only for a set of classifier observation statistics* $P\{\underline{g}_{\mathcal{W}} = \mathcal{W}(j,n)\}$ *which satisfy a fixed set of constraints.*

PROOF. This theorem will be proved by directly specifying the set of constraints that should be satisfied by the statistics estimated from the joint behavior of the classifiers. The Highest Rank method inherently specifies a partitioning on the event space. For each max-score vector (which correspond to a set of $(\underline{r} = n)$ values), Highest Rank method decides on a unique class S_k except for the cases with a max-score collision. In order for this decision method to be optimal, for each specific $(\underline{r} = n)$, a condition of the form

$$P\{\underline{r} = n, \underline{s}_x = k | \underline{g}_{\mathcal{W}} = \mathcal{W}(k,n)\} P\{\underline{g}_{\mathcal{W}} = \mathcal{W}(k,n)\} \geq$$
$$P\{\underline{r} = n, \underline{s}_x = j | \underline{g}_{\mathcal{W}} = \mathcal{W}(j,n)\} P\{\underline{g}_{\mathcal{W}} = \mathcal{W}(j,n)\}, \text{ for } j = 1, 2, \cdots, P, \tag{III.18}$$

should be satisfied. If this is the case, then based on its inherent partitioning, the Highest Rank decision method is optimal in the sense of the POS theory. □

This theorem can be illustrated by means of a simple example. Consider the case of $P = 2$ classes and $Q = 2$ classifiers. For this simple case, by the Highest Rank partitioning, one has exactly $M = 6$ partitions due to three distinct sum-score vectors, which are $[1\ 0]^T, [1\ 1]^T, [0\ 1]^T$. For this case the Highest Rank decisions are illustrated in Table III.3 where assuming a random decision, the

Table III.3: The Highest Rank decision method (i.e., values of the decision variables b_{jn}).

X_1X_2	1 1 0 0	1 0 0 1	0 1 1 0	0 0 1 1
S_1	1	1	0	0
S_2	0	0	1	1

Table III.4: Joint behavior of the classifiers given in the form of objective function expansion coefficients.

X_1X_2	1 1 0 0	1 0 0 1	0 1 1 0	0 0 1 1
S_1	0.12	0.48	$\alpha = 0.08$	0.32
S_2	0.05	0.45	$\beta = 0.05$	0.45

score collisions are arbitrarily resolved one in favor of class S_1 and the other in favor of S_2. Now assume that the joint behavior of the two classifiers is such that one obtains a set of coefficients of the objective function expansion as given in Table III.4. In this table, the third column violates the conditions of the theorem by having $\alpha > \beta$. This column necessitates an optimal decision which is different than the Highest Rank decision. Therefore, for the given classifier joint behavior, the Highest Rank decision method is not optimal.

III.4.2 The Borda Count Method

This is yet another popular method of rank-based multiple classifier decision. It is a slightly generalized majority voting technique from Group Decision Theory [2, 50]. Since it is simple to implement, it has been used as a popular rank-based

Table III.5: Illustration of the partitioning establishing the relation with the Borda Count method.

$$
\begin{array}{ll}
\left\{\left(S_{1,2,3};\begin{bmatrix}2&2\\1&1\\0&0\end{bmatrix}\right)\right\}\mapsto & \left(S_{1,2,3};\begin{bmatrix}4\\2\\0\end{bmatrix}\right)\mapsto \underline{g}_{\mathcal{W}}=1,2,3\\
\left\{\left(S_{1,2,3};\begin{bmatrix}2&2\\1&0\\0&1\end{bmatrix}\right);\left(S_{1,2,3};\begin{bmatrix}2&2\\0&1\\1&0\end{bmatrix}\right)\right\}\mapsto & \left(S_{1,2,3};\begin{bmatrix}4\\1\\1\end{bmatrix}\right)\mapsto \underline{g}_{\mathcal{W}}=4,5,6\\
\vdots & \vdots \qquad\qquad \vdots\\
\left\{\left(S_{1,2,3};\begin{bmatrix}2&0\\1&1\\0&2\end{bmatrix}\right);\left(S_{1,2,3};\begin{bmatrix}2&0\\0&2\\1&1\end{bmatrix}\right);\left(S_{1,2,3};\begin{bmatrix}1&1\\2&0\\0&2\end{bmatrix}\right);\right. & \\
\left.\left(S_{1,2,3};\begin{bmatrix}1&1\\0&2\\2&0\end{bmatrix}\right);\left(S_{1,2,3};\begin{bmatrix}0&2\\2&0\\1&1\end{bmatrix}\right);\left(S_{1,2,3};\begin{bmatrix}0&2\\1&1\\2&0\end{bmatrix}\right)\right\}\mapsto & \left(S_{1,2,3};\begin{bmatrix}2\\2\\2\end{bmatrix}\right)\mapsto \underline{g}_{\mathcal{W}}=16,17,18\\
\vdots & \vdots \qquad\qquad \vdots\\
\left\{\left(S_{1,2,3};\begin{bmatrix}0&0\\1&1\\2&2\end{bmatrix}\right)\right\}\mapsto & \left(S_{1,2,3};\begin{bmatrix}0\\2\\4\end{bmatrix}\right)\mapsto \underline{g}_{\mathcal{W}}=55,56,57
\end{array}
$$

technique. The Borda Count method uses the full rank score matrix $\mathbf{R}$ such that rank scores for each class are summed up across different classifiers. Therefore a *sum score* is obtained for each candidate class and the decision is made by choosing the candidate class having the maximum sum score. The partitioning $\mathcal{W}$ is again illustrated in Table III.5 for an example case of $P = 3$ classes and $Q = 2$ classifiers. There are a total of $M_{\mathcal{W}} = 57$ partitions instead of the original 108. The arguments of Fact 1, Fact 2 and Theorem 1 are again valid for this method which does not use any observation about classifier behavior. It coincides with the optimum decision method only for a set of classifier observation statistics, satisfying a set of conditions similar to Eq. (III.18).

III.4.3 The Logistic Regression Method

Unlike the previous two approaches, the logistic regression method attempts to capture and model the joint behavior of the classifiers and is therefore a much

more important and interesting rank-based method to investigate. The method is introduced in [2], where the fundamental motivation is expressed as *to obtain a linearly weighted Borda Count method where the weights reflect the relative significance of the classifiers.* This method can be interpreted as follows: While the system computes the similarity of a pattern to a class model, only rank scores assigned to that class by all classifiers are used. Let S_k be the class model considered. The next step of this method is to find the probability of correct decision if the decision is S_k given the rank scores for S_k. This probability should be estimated based on the experiments done on the cross-validation data. Finally, the class leading to the highest probability of correct decision is selected. In [2], these probabilities are approximated by first counting frequencies of the related events and then further *smoothing* them by the best hyper-plane approximation. This hyper-plane approximation is determined by formulating and solving a prediction problem. The parameters of the hyper-plane determine the bias and the optimum weighting factors for the classifier outputs, effectively resulting in a *linearly-weighted* Borda Count method extension.

Before introducing the partitioning and the resulting statistics which coincide with this specific method, a number of observations should be discussed. The prediction problem in [2] is first introduced as a class dependent problem, i.e., *to predict the probability of correct decision for a class based on the rank scores generated for that class.* Then, the class dependence is dropped for a *simplicity in notation.* However, this is not the case. Dropping the class index conveniently reduces the problem into predicting a single variable instead of P variables, and at the same time, introduces the important assumption that *the joint behavior*

of the classifiers is independent of the true source class involved. This difference will be illustrated by two corresponding partitionings.

Another point is the use of a hyper-plane fit to the classifier observation statistics. Such an hyper-plane fit is probably the result of the initial motivation in [2], namely to obtain a linearly weighted Borda Count. This is a parametric model which effectively smoothes the estimated statistics, with the assumption that their reliability are uniformly poor. For this discrete case however, the reliability of each statistic should be considered separately, since the number of data points leading to such estimates may be different. The hyper-plane fit treats all such statistics uniformly and may be an over-smoothing for the statistics which are more reliable than others. In this sense, the partitioning approach presented in the present study may be considered as a more controlled way of "smoothing" unreliable estimates, especially if this partitioning is done by considering structure of the training data. As discussed in Section III.3.1, such an *automatic partitioning* may be done by means of optimization methods such as Genetic Algorithms or by some other clustering techniques. One may even use the estimates from a coarser partitioning to smooth the estimates of a finer partitioning.

Now we will relate the unifying theory presented to the Logistic Regression Method by a specific partitioning of the event space which leads to the statistics described and used in [2]. Motivated by the Borda Count method, in [2], for each source class, only the scores of the classifiers for that class are considered for observation. If source class dependence were kept, the resulting partitioning for an example case of $P = 3$ classes and $Q = 2$ classifiers would have been as given in Table III.6. The number of partitions for this case is $P^{(Q+1)}$ in general and the

Table III.6: Illustration of a generalization of the partitioning done in the Logistic Regression method by means of preserving the class dependence.

$$\left\{\left(S_1,\begin{bmatrix}2&2\\1&1\\0&0\end{bmatrix}\right);\left(S_1,\begin{bmatrix}2&2\\1&0\\0&1\end{bmatrix}\right);\left(S_1,\begin{bmatrix}2&2\\0&1\\1&0\end{bmatrix}\right);\left(S_1,\begin{bmatrix}2&2\\0&0\\1&1\end{bmatrix}\right)\right\}\mapsto\left(S_1;\begin{bmatrix}2\\2\end{bmatrix}\right)\mapsto \underline{g}_{\mathcal{W}}=1$$

$$\left\{\left(S_2,\begin{bmatrix}1&1\\2&2\\0&0\end{bmatrix}\right);\left(S_2,\begin{bmatrix}1&0\\2&2\\0&1\end{bmatrix}\right);\left(S_2,\begin{bmatrix}0&1\\2&2\\1&0\end{bmatrix}\right);\left(S_2,\begin{bmatrix}0&0\\2&2\\1&1\end{bmatrix}\right)\right\}\mapsto\left(S_2;\begin{bmatrix}2\\2\end{bmatrix}\right)\mapsto \underline{g}_{\mathcal{W}}=2$$

$$\left\{\left(S_3,\begin{bmatrix}0&0\\1&1\\2&2\end{bmatrix}\right);\left(S_3,\begin{bmatrix}0&1\\1&0\\2&2\end{bmatrix}\right);\left(S_3,\begin{bmatrix}1&0\\0&1\\2&2\end{bmatrix}\right);\left(S_3,\begin{bmatrix}0&0\\0&0\\2&2\end{bmatrix}\right)\right\}\mapsto\left(S_3;\begin{bmatrix}2\\2\end{bmatrix}\right)\mapsto \underline{g}_{\mathcal{W}}=3$$

$$\left\{\left(S_1,\begin{bmatrix}2&1\\1&2\\0&0\end{bmatrix}\right);\left(S_1,\begin{bmatrix}2&1\\1&0\\0&2\end{bmatrix}\right);\left(S_1,\begin{bmatrix}2&1\\0&2\\1&0\end{bmatrix}\right);\left(S_1,\begin{bmatrix}2&1\\0&0\\1&2\end{bmatrix}\right)\right\}\mapsto\left(S_1;\begin{bmatrix}2\\1\end{bmatrix}\right)\mapsto \underline{g}_{\mathcal{W}}=4$$

$$\vdots \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \vdots \qquad \vdots$$

$$\left\{\left(S_2,\begin{bmatrix}2&2\\0&0\\1&1\end{bmatrix}\right);\left(S_2,\begin{bmatrix}2&1\\0&0\\1&2\end{bmatrix}\right);\left(S_2,\begin{bmatrix}1&2\\0&0\\2&1\end{bmatrix}\right);\left(S_2,\begin{bmatrix}1&1\\0&0\\2&2\end{bmatrix}\right)\right\}\mapsto\left(S_2;\begin{bmatrix}0\\0\end{bmatrix}\right)\mapsto \underline{g}_{\mathcal{W}}=26$$

$$\left\{\left(S_3,\begin{bmatrix}2&2\\1&1\\0&0\end{bmatrix}\right);\left(S_3,\begin{bmatrix}2&1\\1&2\\0&0\end{bmatrix}\right);\left(S_3,\begin{bmatrix}1&2\\2&1\\0&0\end{bmatrix}\right);\left(S_3,\begin{bmatrix}1&1\\2&2\\0&0\end{bmatrix}\right)\right\}\mapsto\left(S_3;\begin{bmatrix}0\\0\end{bmatrix}\right)\mapsto \underline{g}_{\mathcal{W}}=27$$

number of elements in each partition is $[(P-1)!]^Q$. However, since the source class dependence is dropped in [2], the actual partitioning done corresponds to the one illustrated in Table III.7. For this case, the number of partitions is reduced to P^Q for the general case and the number of elements inside each partition is increased to $P[(P-1)!]^Q$.

The logistic regression method, in its original form, does not use the event counts directly to estimate the statistics resulting from the partitioning illustrated in Table III.7. It further applies a smoothing by using an hyper-plane fit to the count estimates, in addition to the partitioning involved. Considering the promising results discussed in [2] the smoothing used seems to be justified for the machine printed character recognition task.

Apart from the restrictions imposed by this specific partitioning and the addi-

Table III.7: Illustration of the partitioning done for the Logistic Regression method.

$$\left\{ \left(S_1, \begin{bmatrix} 2 & 2 \\ 1 & 1 \\ 0 & 0 \end{bmatrix}\right); \left(S_1, \begin{bmatrix} 2 & 2 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}\right); \left(S_1, \begin{bmatrix} 2 & 2 \\ 0 & 1 \\ 1 & 0 \end{bmatrix}\right); \left(S_1, \begin{bmatrix} 2 & 2 \\ 0 & 0 \\ 1 & 1 \end{bmatrix}\right); \left(S_2, \begin{bmatrix} 1 & 1 \\ 2 & 2 \\ 0 & 0 \end{bmatrix}\right); \left(S_2, \begin{bmatrix} 1 & 0 \\ 2 & 2 \\ 0 & 1 \end{bmatrix}\right); \right.$$
$$\left. \left(S_2, \begin{bmatrix} 0 & 1 \\ 2 & 2 \\ 1 & 0 \end{bmatrix}\right); \left(S_2, \begin{bmatrix} 0 & 0 \\ 2 & 2 \\ 1 & 1 \end{bmatrix}\right); \left(S_3, \begin{bmatrix} 0 & 0 \\ 1 & 1 \\ 2 & 2 \end{bmatrix}\right); \left(S_3, \begin{bmatrix} 0 & 1 \\ 1 & 0 \\ 2 & 2 \end{bmatrix}\right); \left(S_3, \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 2 & 2 \end{bmatrix}\right); \left(S_3, \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 2 & 2 \end{bmatrix}\right) \right\} \mapsto \left(\begin{bmatrix} 2 \\ 2 \end{bmatrix}\right) \mapsto \underline{g}_{\mathcal{W}} = 1$$

$$\left\{ \left(S_1, \begin{bmatrix} 2 & 1 \\ 1 & 2 \\ 0 & 0 \end{bmatrix}\right); \left(S_1, \begin{bmatrix} 2 & 1 \\ 1 & 0 \\ 0 & 2 \end{bmatrix}\right); \left(S_1, \begin{bmatrix} 2 & 1 \\ 0 & 2 \\ 1 & 0 \end{bmatrix}\right); \left(S_1, \begin{bmatrix} 2 & 1 \\ 0 & 0 \\ 1 & 2 \end{bmatrix}\right); \left(S_2, \begin{bmatrix} 1 & 2 \\ 2 & 1 \\ 0 & 0 \end{bmatrix}\right); \left(S_2, \begin{bmatrix} 1 & 0 \\ 2 & 1 \\ 0 & 2 \end{bmatrix}\right); \right.$$
$$\left. \left(S_2, \begin{bmatrix} 0 & 2 \\ 2 & 1 \\ 1 & 0 \end{bmatrix}\right); \left(S_2, \begin{bmatrix} 0 & 0 \\ 2 & 1 \\ 1 & 2 \end{bmatrix}\right); \left(S_3, \begin{bmatrix} 0 & 0 \\ 1 & 2 \\ 2 & 1 \end{bmatrix}\right); \left(S_3, \begin{bmatrix} 0 & 2 \\ 1 & 0 \\ 2 & 1 \end{bmatrix}\right); \left(S_3, \begin{bmatrix} 1 & 0 \\ 0 & 2 \\ 2 & 1 \end{bmatrix}\right); \left(S_3, \begin{bmatrix} 1 & 2 \\ 0 & 0 \\ 2 & 1 \end{bmatrix}\right) \right\} \mapsto \left(\begin{bmatrix} 2 \\ 1 \end{bmatrix}\right) \mapsto \underline{g}_{\mathcal{W}} = 2$$

$$\vdots \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \vdots \qquad \vdots$$

$$\left\{ \left(S_1, \begin{bmatrix} 0 & 0 \\ 2 & 2 \\ 1 & 1 \end{bmatrix}\right); \left(S_1, \begin{bmatrix} 0 & 0 \\ 2 & 1 \\ 1 & 2 \end{bmatrix}\right); \left(S_1, \begin{bmatrix} 0 & 0 \\ 1 & 2 \\ 2 & 1 \end{bmatrix}\right); \left(S_1, \begin{bmatrix} 0 & 0 \\ 1 & 1 \\ 2 & 2 \end{bmatrix}\right); \left(S_2, \begin{bmatrix} 2 & 2 \\ 0 & 0 \\ 1 & 1 \end{bmatrix}\right); \left(S_2, \begin{bmatrix} 2 & 1 \\ 0 & 0 \\ 1 & 2 \end{bmatrix}\right); \right.$$
$$\left. \left(S_2, \begin{bmatrix} 1 & 2 \\ 0 & 0 \\ 2 & 1 \end{bmatrix}\right); \left(S_2, \begin{bmatrix} 1 & 1 \\ 0 & 0 \\ 2 & 2 \end{bmatrix}\right); \left(S_3, \begin{bmatrix} 2 & 2 \\ 1 & 1 \\ 0 & 0 \end{bmatrix}\right); \left(S_3, \begin{bmatrix} 2 & 1 \\ 1 & 2 \\ 0 & 0 \end{bmatrix}\right); \left(S_3, \begin{bmatrix} 1 & 2 \\ 2 & 1 \\ 0 & 0 \end{bmatrix}\right); \left(S_3, \begin{bmatrix} 1 & 1 \\ 2 & 2 \\ 0 & 0 \end{bmatrix}\right) \right\} \mapsto \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}\right) \mapsto \underline{g}_{\mathcal{W}} = 9$$

tional linear smoothing, the Logistic Regression method is optimal in the sense of the POS theory introduced in the present paper. However, using the partitioning interpretation instead of viewing the method as a simple Borda Count generalization shows that many other partitionings (hence new methods) are possible some of which may be better suited for a particular task. The motivation of finding better ones opens new direction for future research and the introduced formalism can be used as a tool to analyze them.

III.5 Links With Type 1 Systems: First Rank Partitioning

The final decision based multiple classifier systems, termed as Type 1 by Xu et.al. [9], use Q individual final decisions in the form $(\underline{d}_j = i_j)$ from the classifiers taking part in the system and combines them to generate the overall system decision $\underline{d}$.

III.5.1 The Traditional Bayesian Formalism

Under the Bayesian formalism, the combined or overall decision is often performed using the maximum aposteriori (MAP) criterion. Therefore, the combined decision of the system is based on the posterior class probabilities and can be expressed as

$$\underline{d} = \underset{1 \leq j \leq P}{\operatorname{argmax}} P\{\underline{S}_x = j | \underline{d}_1 = i_1, \underline{d}_2 = i_2, \cdots, \underline{d}_Q = i_Q\}. \tag{III.19}$$

Let us temporarily denote the posterior probabilities in III.19 by P_j. To compute the P_j values needed for the maximum aposteriori decision, Bayes rule is used on Eq. (III.19) to yield

$$P_j = \frac{P\{\underline{S}_x = j, \underline{d}_1 = i_1, \underline{d}_2 = i_2, \cdots, \underline{d}_Q = i_Q\}}{P\{\underline{d}_1 = i_1, \underline{d}_2 = i_2, \cdots, \underline{d}_Q = i_Q\}}, \tag{III.20}$$

which is transformed to

$$P_j = \frac{P\{\underline{d}_1 = i_1, \underline{d}_2 = i_2, \cdots, \underline{d}_Q = i_Q | \underline{S}_x = j\} P\{\underline{S}_x = j\}}{P\{\underline{d}_1 = i_1, \underline{d}_2 = i_2, \cdots, \underline{d}_Q = i_Q\}}. \tag{III.21}$$

At this point, what is generally done is to introduce the *statistical independence assumption* about the classifiers, whose implications are discussed in detail in Chapter IV. By this assumption, Eq. (III.21) can be reduced to the form

$$P_j = \frac{\prod_{k=1}^{Q} P\{\underline{d}_k = i_k | \underline{S}_x = j\} P\{\underline{S}_x = j\}}{\prod_{k=1}^{Q} P\{\underline{d}_k = i_k\}}. \tag{III.22}$$

Therefore, if we also consider the fact that the denominator term is constant across the possible values of j, the maximum aposteriori decision under the Bayesian formalism together with the independence assumption becomes

$$\underline{d} = \underset{1 \leq j \leq P}{\operatorname{argmax}} \prod_{k=1}^{Q} P\{\underline{d}_k = i_k | \underline{S}_x = j\} P\{\underline{S}_x = j\}. \tag{III.23}$$

If the independence assumption were not introduced, Eq. (III.23) would be

$$\underline{d} = \underset{1 \leq j \leq P}{\operatorname{argmax}} P\{\underline{d}_1 = i_1, \underline{d}_2 = i_2, \cdots, \underline{d}_Q = i_Q | \underline{S}_x = j\} P\{\underline{S}_x = j\}. \tag{III.24}$$

Each decision output behavior of each individual classifier, i.e., the errors made by each classifier X_k can be modeled by its *confusion matrix*

$$\mathbf{C}_k = \begin{bmatrix} n_{11}^{(k)} & n_{12}^{(k)} & \cdots & n_{1P}^{(k)} \\ n_{21}^{(k)} & n_{22}^{(k)} & \cdots & n_{2P}^{(k)} \\ \vdots & \vdots & \ddots & \vdots \\ n_{P1}^{(k)} & n_{P2}^{(k)} & \cdots & n_{PP}^{(k)} \end{bmatrix} \tag{III.25}$$

which is a count table accumulating the cross-validation test patterns with respect to the source class and the classifier class decisions. Row j corresponds to the patterns from class $\underline{S}_x = j$ while column i corresponds to the patterns for which the classifier made a class decision $\underline{d}_k = i$.

The independence assumption made while going from Eq. (III.21) to Eq. (III.22) allows one to use the confusion matrices for each individual classifier to estimate the posterior probabilities P_j, so that the maximum likelihood combined decision in Eq. (III.19) can be made. The posterior probabilities can be

estimated from the counts in the confusion matrices. As an example, consider the conditional probability $P\{\underline{d}_k = i_k | \underline{S}_x = j\}$. Denoting i_k by i for simplicity in notation, it can be estimated as

$$P\{\underline{d}_k = i | \underline{S}_x = j\} = \frac{n_{ji}^{(k)}}{\sum_{i=1}^{P} n_{ji}^{(k)}}. \tag{III.26}$$

The main problem with this approach is the statistical independence assumption for the decision outputs of the individual classifiers, which may not always be a valid one. If this assumption is left out of the derivation, then we have Eq. (III.21) and we can no longer use the individual confusion matrices of the classifiers and statistics about the joint behavior of the classifiers is needed. Modeling the joint behavior for a reasonable observation resolution is what is being attempted by our POS theory.

III.5.2 First Rank Partitioning

Now it will be shown that for a specific partitioning of the observation space, namely for one which take into account only the final decisions from each classifier, the optimum decision method given by the POS theory reduces to the maximum aposteriori decision given in Eq. (III.19).

The *first rank partitioning* is done by considering only the topmost rank for each classifier, i.e., the final decision of each classifier. This corresponds to the new rank score matrix definition based on the elements of the original rank score matrix, as

Table III.8: Illustration of the First Rank Partitioning leading to the relations with the final decision based Bayesian classifier combination.

$$\left\{\left(S_{1,2,3},\begin{bmatrix}2&2\\1&1\\0&0\end{bmatrix}\right);\left(S_{1,2,3},\begin{bmatrix}2&2\\1&0\\0&1\end{bmatrix}\right);\left(S_{1,2,3},\begin{bmatrix}2&2\\0&1\\1&0\end{bmatrix}\right);\left(S_{1,2,3},\begin{bmatrix}2&2\\0&0\\1&1\end{bmatrix}\right)\right\}\mapsto\left(S_{1,2,3},\begin{bmatrix}1&1\\0&0\\0&0\end{bmatrix}\right)\mapsto \underline{g}_{\mathcal{W}}=1,2,3$$

$$\left\{\left(S_{1,2,3},\begin{bmatrix}2&1\\1&2\\0&0\end{bmatrix}\right);\left(S_{1,2,3},\begin{bmatrix}2&0\\1&2\\0&1\end{bmatrix}\right);\left(S_{1,2,3},\begin{bmatrix}2&1\\0&2\\1&0\end{bmatrix}\right);\left(S_{1,2,3},\begin{bmatrix}2&0\\0&2\\1&1\end{bmatrix}\right)\right\}\mapsto\left(S_{1,2,3},\begin{bmatrix}1&0\\0&1\\0&0\end{bmatrix}\right)\mapsto \underline{g}_{\mathcal{W}}=4,5,6$$

$$\vdots \qquad\qquad \vdots$$

$$\left\{\left(S_{1,2,3},\begin{bmatrix}1&1\\0&0\\2&2\end{bmatrix}\right);\left(S_{1,2,3},\begin{bmatrix}1&0\\0&1\\2&2\end{bmatrix}\right);\left(S_{1,2,3},\begin{bmatrix}0&1\\1&0\\2&2\end{bmatrix}\right);\left(S_{1,2,3},\begin{bmatrix}0&0\\1&1\\2&2\end{bmatrix}\right)\right\}\mapsto\left(S_{1,2,3},\begin{bmatrix}0&0\\0&0\\1&1\end{bmatrix}\right)\mapsto \underline{g}_{\mathcal{W}}=25,26,27$$

$$\hat{r}_{ij}=\begin{cases}1, & \text{if } r_{ij}=P-1,\\ 0, & \text{otherwise.}\end{cases} \tag{III.27}$$

For an illustrative example case of $P=3$ classes and $Q=2$ classifiers, the partitioning is demonstrated in Table III.8. As discussed in Section III.3, for a partitioned observation space, the optimal decision method is given by

$$\underline{d}=\underset{1\le j\le P}{\operatorname{argmax}}\, P\{\underline{s}_x=j,\underline{r}=n|\underline{g}_{\mathcal{W}}=\mathcal{W}(j,n)\}P\{\underline{g}_{\mathcal{W}}=\mathcal{W}(j,n)\}. \tag{III.28}$$

But for the specific partitioning defined as *first rank partitioning* we have the following two properties:

(1) The observation statistics $P\{\underline{g}_{\mathcal{W}}=\mathcal{W}(j,n)\}$ are in fact given by

$$\begin{aligned}P\{\underline{g}_{\mathcal{W}}=\mathcal{W}(j,n)\} &= P\{\underline{g}_{\mathcal{W}}=\mathcal{W}(j,n(i_1,i_2,\cdots,i_Q))\}\\ &= P\{\underline{d}_1=i_1,\underline{d}_2=i_2,\cdots,\underline{d}_Q=i_Q,\underline{S}_x=j\}.\end{aligned}$$

where $\underline{d}_i$ is the decision of classifier X_i and $n(i_1,i_2,\cdots,i_Q)$ is a function of these decisions.

(2) The transition terms $P\{\underline{s}_x = j, \underline{r} = n | \underline{g}_{\mathcal{W}} = \mathcal{W}(j,n)\}$ *are constant since the cardinality of each partition is* $|W_m| = 1/\mu, \forall m \in \{1, 2, \cdots, M_{\mathcal{W}}\}$ *for the first rank partitioning, i.e., we have*

$$P\{\underline{s}_x = j, \underline{r} = n | \underline{g}_{\mathcal{W}} = \mathcal{W}(j,n)\} = \mu, \quad \forall j \in \{1, 2, \cdots, P\}, n \in \{1, 2, \cdots, N\}.$$

Using the aforementioned two properties of this specific partitioning, the optimum decision resulting from the POS theory as given in Eq. (III.28) can be expressed as

$$\begin{aligned} \underline{d} &= \underset{1 \le j \le P}{\operatorname{argmax}} \{\mu \cdot P\{\underline{g}_{\mathcal{W}} = \mathcal{W}(j,n)\}\} && \text{(III.29)} \\ &= \underset{1 \le j \le P}{\operatorname{argmax}} P\{\underline{g}_{\mathcal{W}} = \mathcal{W}(j,n)\} && \text{(III.30)} \\ &= \underset{1 \le j \le P}{\operatorname{argmax}} P\{\underline{d}_1 = i_1, \underline{d}_2 = i_2, \cdots, \underline{d}_Q = i_Q | \underline{S}_x = j\} P\{\underline{S}_x = j\} && \text{(III.31)} \end{aligned}$$

The last form given in Eq. (III.31) is clearly equivalent to the Bayesian decision rule without the independence assumption given in Eq. (III.24), establishing our claim in this section.

III.6 Discussion

In this chapter, the partitioned observation space theory is introduced as a promising tool for understanding rank-based multiple classifier decision systems. During the discussion in this part, it has been deliberately avoided to introduce a specific rank-based multiple classifier decision method based on the presented theory but a clear methodology of obtaining a number of such methods by means of different partitionings of the observation space is provided. Specific partitionings are

introduced only to establish the links with some existing rank-based approaches. It has been shown that three rank-based multiple classifier decision methods can be analyzed using the POS theory. Additionally, it has also been shown that the final decision based (Type 1) systems can also be analyzed under this formalism and that the Bayesian approach without the independence assumption is a special case of the developed approach.

Under the developed formalism, specific methods result from specific partitionings and such partitionings should be considered within the context of a specific task domain. The application of the theory in two tasks from the speech processing domain are considered and the computational model necessary for these applications are developed in Chapters VI and VII. The common speech processing concepts and the individual classifiers are developed in Chapter V. In the following chapter, the thesis elaborates on the important concepts of independence and complementariness among classifiers and develops an interpretation based on Information Theory.

CHAPTER IV

Independence and Complementariness of Classifiers

Two closely related concepts arise while using a multiple-classifier system. These are the *independence* and the *complementariness* of the classifiers involved within the system.

While constructing a multiple classifier decision combination system, one is faced with several important problems. It is often not clear whether there will really be an improvement over the performance of the best classifier by the use of more than one classifier. This clearly depends on the individual performances of the classifiers involved and their interaction during the classification process. One is faced with the problem of determining the potential improvement possible by making collective use of multiple classifiers.

Another issue of considerable importance is the computational load implied by the parallel use of multiple classifiers. Given a large set of potential classifiers

(e.g., using different features extraction methods, different modeling/similarity scoring methods), using all of them in parallel may guarantee performance improvement but may not be computationally feasible with the hardware capabilities at hand. Practical considerations often necessitate selecting a suitable subset of classifiers which satisfy a certain performance gain/classification speed tradeoff.

Finally, one should be interested in understanding theoretically *when* and *why* a given set of classifiers, when combined, lead to improved performance, while others do not. Loosely stated, the aforementioned objectives may only be achieved if it is possible to quantify the potential of the combiner to improve the classification performance. The *complementariness* concept and an associated measure may be used to quantify such an ability.

Independence and *complementariness* concepts have been around in the pattern recognition literature for a long time. Unfortunately, the concepts have been often used loosely, without any attempt for a solid definition and the development of a quantifying measure. For example, the dependence between the set of classifiers is often ignored and a statistical independence assumption is used in the development [31, 15]. Some other researchers have argued that statistical independence of the classifier outputs is not really the useful measure for quantifying improved performance but the *independence of the errors made* should be considered instead. This is also left as a verbal argument [12, 13]. There have also been solid contributions such as by Tumer and Ghosh [29, 29, 47]. They have shown the relations between classifier output correlation and the deviation from the optimal Bayesian decision boundary for classifiers which are combined by linear averaging or by order statistics. Unfortunately, their results apply to

classifiers with continuous outputs in measurement form and cannot be extended trivially to rank-based classifier systems.

The present chapter will attempt to introduce a formal treatment of classifier *independence* and *complementariness* concepts for rank-based multiple-classifier systems [1] studied in Chapter III. For this purpose, some relevant basic concepts of Information Theory will be used.

IV.1 Relevant Concepts of Information Theory

Information theory gives us a promising tool to explore the complementariness of multiple classifiers. To illustrate this, we will first summarize some relevant basic results using the notation of Chapter III [57].

Consider a finite event space $\mathcal{R} = \{\mathbf{R}_1, \mathbf{R}_2, \cdots, \mathbf{R}_N\}$ and let $\underline{r}$ be an integer random variable defined such that $\underline{r} = n$ denotes the realization of the rank score matrix $\mathbf{R}_n$. This event space can be thought as a source of information. One can define a *measure* for the information conveyed by the realization of the event $\underline{r} = n$ in terms of its probability as

$$I(\underline{r} = n) = \log \frac{1}{P\{\underline{r} = n\}}. \tag{IV.1}$$

The expected value of the information acquired by the observation of $\mathcal{R}$ is

$$\begin{aligned} H(\underline{r}) &= E\{I(\underline{r} = n)\} \\ &= \textstyle\sum_{n=1}^{N} P\{\underline{r} = n\} \log \frac{1}{P\{\underline{r}=n\}}, \end{aligned} \tag{IV.2}$$

[1] In Chapter III, rank-based systems have been shown to include also the systems where the classifier outputs are final class decisions (i.e., Type 1 systems). Therefore, the results of this chapter will be readily applicable for both group of systems.

which is also known as the *entropy* of this information source. This quantity can be interpreted as a number of properties of the event space $\mathcal{R}$ or the associated random variable $\underline{r}$ [57]. These are *the amount of average "information" conveyed by an observation of* $\underline{r}$, *our uncertainty about* $\underline{r}$ or *the randomness of* $\underline{r}$. The units of these information measures depend on the base of the log(.) operator. For a base 2 logarithm, the unit of information is *bit*. The well known Theorem IV.1 establishes the minimum and maximum values for the entropy function and its proof can be found in [57].

THEOREM IV.1 *Let* $\underline{r} \in \{1, 2, \cdots, N\}$; *then one has* $0 \leq H(\underline{r}\} \leq \log N$. *Furthermore* $H(\underline{r}) = 0$ *iff* $\exists j \in \{1, 2, \cdots, N\}$ *such that* $P\{\underline{r} = j\} = 1$ *and* $H(\underline{r} = \log N)$ *iff* $\forall j \in \{1, 2, \cdots, N\}$ *we have* $P\{\underline{r} = j\} = 1/N$.

Consider now that there are two random variables $\underline{r}_1$ and $\underline{r}_2$ with probability mass functions $P\{\underline{r}_1 = n_1\}$ and $P\{\underline{r}_2 = n_2\}$, representing two related event spaces $\mathcal{R}_1$ and $\mathcal{R}_2$. The relation between these two probability distributions is given by the the conditional probability $P\{\underline{r}_1 = n_1 | \underline{r}_2 = n_2\}$. Now if we define the information conveyed by observing the realization $(\underline{r}_1 = n_1)$ given that we have already observed $(\underline{r}_2 = n_2)$ as

$$I(\underline{r}_1 = n_1 | \underline{r}_2 = n_2) = \log \frac{1}{P\{\underline{r}_1 = n_1 | \underline{r}_2 = n_2\}}, \tag{IV.3}$$

then the *entropy* of $\underline{r}_1$ after observing $\underline{r}_2$ can be found [57] as

$$H(\underline{r}_1 | \underline{r}_2) = \sum_{n_1, n_2} P\{\underline{r}_1 = n_1, \underline{r}_2 = n_2\} \log \frac{1}{P\{\underline{r}_1 = n_1 | \underline{r}_2 = n_2\}}. \tag{IV.4}$$

This *conditional entropy* may be interpreted as a number of properties of $\underline{r}_1$ and $\underline{r}_2$: *The amount of average "information" conveyed by an observation of* $\underline{r}_1$ *given that we have already observed* $\underline{r}_2$*, our uncertainty remaining about* $\underline{r}_1$ *given that we have resolved our uncertainty about* $\underline{r}_2$ or *the randomness of* $\underline{r}_1$ *after observing* $\underline{r}_2$. Since we know our uncertainty about $\underline{r}_1$ both *before* and *after* observing $\underline{r}_2$, we can derive the amount of average information we have acquired about the former by observing the latter. This symmetric quantity is known as the *mutual information* between $\underline{r}_1$ and $\underline{r}_2$ and is given by

$$I(\underline{r}_1, \underline{r}_2) = H(\underline{r}_1) - H(\underline{r}_1|\underline{r}_2). \tag{IV.5}$$

which can be expressed in explicit form as

$$I(\underline{r}_1, \underline{r}_2) = \sum_{n_1,n_2} P\{\underline{r}_1 = n_1, \underline{r}_2 = n_2\} \log \frac{P\{\underline{r}_1 = n_1, \underline{r}_2 = n_2\}}{P\{\underline{r}_1 = n_1\} P\{\underline{r}_2 = n_2\}}. \tag{IV.6}$$

THEOREM IV.2 *We have* $I(\underline{r}_1, \underline{r}_2) \geq 0, \forall \underline{r}_1, \underline{r}_2$ *and* $I(\underline{r}_1, \underline{r}_2) = 0$ *if and only if the two random variables are statistically independent.*

Theorem IV.2 whose proof can be found in [57] asserts that the mutual information as defined in Eq. (IV.5) is a well suited measure of statistical dependence between the random variables $\underline{r}_1$ and $\underline{r}_2$ hence between the underlying events [57]. These concepts can be applied in the context of multiple classifier systems as discussed in the following sections.

IV.2 An Information Theoretic Interpretation of Classifiers

Information theory defines a *discrete memoryless communication channel* (DMC) as *an object that accepts, every unit of time, one of P input symbols and outputs one of N output symbols.* The output can be thought of as a noisy version of the input [57]. A classifier on the other hand, is an object which accepts patterns, whose class labels are known to a *supervisor*, and outputs its best estimates of these class labels.

A classifier can be interpreted as analogous to a DMC if we argue that the true realization of the class label is transformed by the classifier into a noisy output form. The source of the noise is not important for this interpretation but it may be the result of the feature extraction and/or the similarity scoring algorithm. The actual source of information we are interested in (the input to the DMC interpretation of the classifier) is the true label of the class emitting the patterns. However, what we have access to is only the noisy output of this DMC as illustrated by Figure IV.1.

When more than one classifiers are involved, we may consider them as multiple DMCs transmitting the same information source whose outputs are to be considered to acquire information about this source.

IV.3 Output Independence of Classifiers

A multiple classifier decision combination system with observation space partitioning can be visualized as a set of interrelated random variables as illustrated

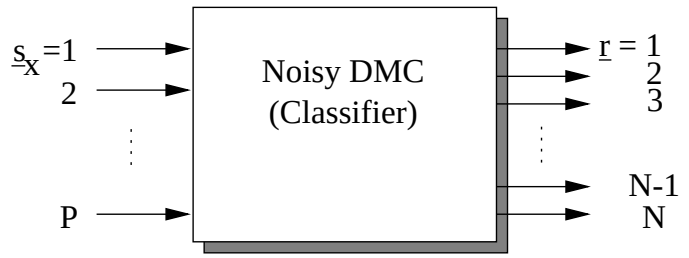

Figure IV.1: The discrete memoryless channel interpretation of a classifier. The input to the DMC is the true label of the pattern while the output of the DMC is the classifier output. The exact number of outputs depends on the level of information supplied by the classifier.

in Figure IV.2. With the random variable definitions given in Figure IV.2, we are at a point to introduce a formal definition of independence among the outputs of classifiers both before and after observation space partitioning as described in Chapter III. Consider two classifiers whose rank-based outputs represented by the random variables $\underline{r}_1, \underline{r}_2$. In view of Theorem IV.2 we can make the following definition which can easily be extended to more than two classifiers.

DEFINITION IV.1 *Classifiers X_1, X_2 are said to be* output independent in the rank-based sense *if and only if we have $I(\underline{r}_1, \underline{r}_2) = 0$ with $I(\underline{r}_1, \underline{r}_2)$ defined by*

$$I(\underline{r}_1, \underline{r}_2) = \sum_{n_1,n_2,j} P\{\underline{r}_1 = n_1, \underline{r}_2 = n_2 | \underline{s}_x = j\} \log \frac{P\{\underline{r}_1=n_1, \underline{r}_2=n_2 | \underline{s}_x=j\}}{P\{\underline{r}_1=n_1 | \underline{s}_x=j\} P\{\underline{r}_2=n_2 | \underline{s}_x=j\}}. \tag{IV.7}$$

Otherwise, the two classifiers are output dependent *with $I(\underline{r}_1, \underline{r}_2)$ being a measure of dependence between them.*

If one uses the random variables $\underline{r}_1, \underline{r}_2, \cdots, \underline{r}_Q$ in this definition, then the output dependence of the original classifiers is computed. However, it is also possible to compute the output dependence, after a partitioning of the observation

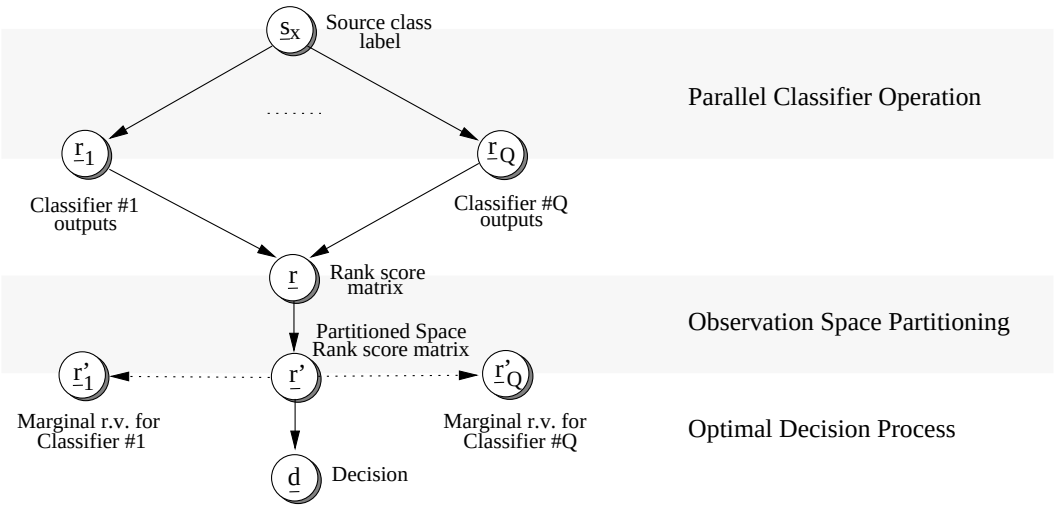

Figure IV.2: Random variable representation of the multiple classifier decision combination system. The events within the system can be represented by a number of interrelated random variables. The random variables are transformed from one to another either by means of the classifiers, or by means of the partitioning and the optimal decision process. $\underline{r}'_1, \underline{r}'_2, \cdots, \underline{r}'_Q$ are the marginal random variables reflecting the individual classifier outputs after the observation space partitioning is applied.

space as described in Chapter III. For this, the marginal output random variables $\underline{r}'_1, \underline{r}'_2, \cdots, \underline{r}'_Q$, derived after the partitioning $\mathcal{W}$, should be used instead of $\underline{r}_1, \underline{r}_2, \cdots, \underline{r}_Q$. Note that the numerical measure of dependence among the outputs of the classifiers will be different depending on whether this is computed for the original outputs or after each specific partitioning

Output independence of classifiers is an important parameter in itself. However, as is shown by the following example, it is not necessarily a measure of *complementariness*.

EXAMPLE IV.1 Suppose we consider two rank-based classifiers X_1 and X_2 operating on a simple two class problem where the class labels are S_1 and S_2. Assume that these classifiers are operated in parallel on patterns from these two classes

Table IV.1: True joint probability distribution of the classifier observation space. Columns denote the *rank score matrices* while rows denote pattern classes. Each cell represent the estimate of the probability that patterns from a class lead to a specific rank score matrix at the outputs of the classifiers.

X_1X_2	$\begin{bmatrix}1&1\\0&0\end{bmatrix}$	$\begin{bmatrix}1&0\\0&1\end{bmatrix}$	$\begin{bmatrix}0&1\\1&0\end{bmatrix}$	$\begin{bmatrix}0&0\\1&1\end{bmatrix}$
S_1	0.45	0.35	0.15	0.05
S_2	0.15	0.05	0.05	0.75

and the joint probabilities in Table IV.1 are obtained. These will be called as the *true* joint distribution of the classifier behavior. The marginal probabilities for the individual classifiers can be obtained from this joint distribution and are given in Table IV.2 (a) and (b).

Probability of the errors made by the individual classifiers may be analyzed from these two marginal tables. Considering Table IV.2 (a), the jointly optimum decision [2] selects class S_1 if the rank score matrix $(1\ 0)^T$ occurs and S_2 if $(0\ 1)^T$ occurs at the classifier outputs. Denoting the decision by $\underline{d}$ and using the random variable notations of Chapter III, the total probability of error for classifier X_1 is

$$\begin{aligned} P^e_{X_1} &= P\{\underline{d}=1|\underline{s}_x=2\}P\{\underline{s}_x=2\} + P\{\underline{d}=2|\underline{s}_x=1\}P\{\underline{s}_x=1\} \\ &= 0.2\times 0.5 + 0.2\times 0.5 \\ &= 0.2 \end{aligned}$$

By a similar computation for classifier X_2, we have $P^e_{X_2} = 0.3$. Therefore, it can be argued that X_1 is the best of the two classifiers.

[2] The *jointly optimum decision* is in the sense of Chapter III. In this sense, the *jointly optimum decision* and the *optimum combination* is synonymous. For this example, the class label with the largest probability for a given column is selected.

Table IV.2: Marginal probabilities for individual classifiers (a) X_1 and (b) X_2. Columns denote the *rank score vectors* at classifier outputs while rows denote pattern classes. These tables are rank-based generalized forms of the matrices known as classifier *confusion matrices*.

(a)

X_1	$(1\ 0)^T$	$(0\ 1)^T$
S_1	0.8	0.2
S_2	0.2	0.8

(b)

X_2	$(1\ 0)^T$	$(0\ 1)^T$
S_1	0.6	0.4
S_2	0.2	0.8

Table IV.3: Joint probability distribution of the classifier observation space computed from the marginal distributions in Table IV.2, under the *independence assumption*.

X_1X_2	$\begin{bmatrix}1&1\\0&0\end{bmatrix}$	$\begin{bmatrix}1&0\\0&1\end{bmatrix}$	$\begin{bmatrix}0&1\\1&0\end{bmatrix}$	$\begin{bmatrix}0&0\\1&1\end{bmatrix}$
S_1	0.48	0.32	0.12	0.08
S_2	0.04	0.16	0.16	0.64

Let the true joint distribution in Table IV.1 be used for jointly optimal decision. From this table, it can be seen that one has $\underline{d} = 1$ if $\underline{r} \in \{1, 2, 3\}$ and $\underline{d} = 2$ if $\underline{r} = 4$, where $\underline{r}$ denotes the realization of the rank score matrix. The total probability of error for the jointly optimal decision is

$$\begin{aligned} P^e_{X_1X_2} &= (0.15 + 0.05 + 0.05) \times 0.5 + 0.75 \times 0.5 \\ &= 0.15 \end{aligned}$$

which is lower than the probability of error $P^e_{X_1} = 0.2$ for the best individual classifier. Therefore, an improvement in performance over the best individual classifier is achieved by the jointly optimal decision. Now suppose we assume that the classifiers are independent. Then, we can construct a joint probability distribution by making use of this assumption. This derived distribution is given in Table IV.3.

When this derived joint distribution is considered for optimal decision, one has now $\underline{d} = 1$ if $\underline{r} \in \{1, 2\}$ and $\underline{d} = 2$ if $\underline{r} \in \{3, 4\}$. In this case, the total probability of error would clearly be

$$\begin{aligned} \hat{P}^e_{X_1X_2} &= (0.04 + 0.16) \times 0.5 + (0.12 + 0.08) \times 0.5 \\ &= 0.20, \end{aligned}$$

which shows no improvement over the performance of the best classifier X_1.

This simple example shows that the independence assumption may hide a potential for improvement for classifiers which are in fact dependent. It also shows that independence of classifiers is not a necessary condition for such an

Table IV.4: Joint probability distribution of the classifier observation space for the two classifiers of Example IV.2.

X_1X_2	$\begin{bmatrix}1&1\\0&0\end{bmatrix}$	$\begin{bmatrix}1&0\\0&1\end{bmatrix}$	$\begin{bmatrix}0&1\\1&0\end{bmatrix}$	$\begin{bmatrix}0&0\\1&1\end{bmatrix}$
S_1	0.12	0.48	0.08	0.32
S_2	0.05	0.45	0.05	0.45

Table IV.5: Marginal probabilities for individual classifiers (a) X_1 and (b) X_2 in Example IV.2.

(a)

X_1	$(1\ 0)^T$	$(0\ 1)^T$
S_1	0.6	0.4
S_2	0.5	0.5

(b)

X_2	$(1\ 0)^T$	$(0\ 1)^T$
S_1	0.2	0.8
S_2	0.1	0.9

improvement. For dependent classifiers, the jointly optimal decision process in the sense of the theory developed in Chapter III may achieve an improvement over the best individual classifier while methods based on the independence assumption will fail to do so. An interesting question at this point is whether or not an improvement is still possible for the case of classifiers which are truly output independent. The following example gives a positive answer.

EXAMPLE IV.2 Again consider a simple problem with two classifiers X_1 and X_2, operating on patterns from two classes S_1 and S_2. The joint distribution of the classifier observation space is given in Table IV.4 while the marginal distributions for the individual classifiers are given in Table IV.5 (a) and (b).

For this example, we have $I(\underline{r}_1, \underline{r}_2)) = 0$ and therefore, the classifiers are output independent. The total probability of error for both individual classifiers

Table IV.6: Class dependent error probabilities for classifiers in Example IV.2.

Classifier	X_1	X_2
$P\{\underline{y} = 0 \mid \underline{s}_x = 1\}$	0.4	0.8
$P\{\underline{y} = 0 \mid \underline{s}_x = 2\}$	0.5	0.1

are $P^e_{X_1} = P^e_{X_2} = 0.45$. However, when the joint distribution is considered for optimal decision, the decisions are $\underline{d} = 1$ when $\underline{r} \in \{1, 2, 3\}$ and $\underline{d} = 2$ when $\underline{r} = 4$ effectively leading to a total probability of error of $P^e_{Comb} = 0.435$. This is smaller than the probability of error for both of the classifiers denoting an improved performance for the case of output independent classifiers.

An interesting observation can be made about these classifiers if one inspects the *class dependent error probabilities* $P\{\underline{y} = 0 \mid \underline{s}_x = 1\}$ and $P\{\underline{y} = 0 \mid \underline{s}_x = 2\}$ where $\underline{y}$ is the *indicator* of correct decision (see Section III.2.1). These are given in Table IV.6. From these probabilities, it can be concluded that classifier X_1 cannot successfully classify patterns from class S_2 while classifier X_2 cannot classify patterns from class S_1. The fact that the errors of the two classifiers are concentrated on different classifiers support the ideas in [12, 13].

IV.4 A Condition for Complementariness

The joint distribution in Table IV.3 is obtained under the assumption of independence from the marginal distributions. However, this could as well have been the true joint distribution of the classifier observation space. Given the true joint distribution and the marginal distributions, one important task is to find the conditions on these distributions so that there will be an improvement by using the

jointly optimal decision. Such a general condition is introduced by the following Definition and Theorem.

DEFINITION IV.2 *In a multiple classifier system, a classifier is called as the* dominating classifier *if the jointly optimal decision is a function of only the rank score vector of that classifier.*

THEOREM IV.3 *If one classifier dominates the others, then the jointly optimal performance of the multiple classifier system becomes exactly equal to the performance of the dominating classifier.*

PROOF. Consider the first classifier and let the classifier decide on class S_k for a specific rank score vector $\underline{r}_1 = n_1$, where $\underline{r}_1$ denotes the rank-based output of classifier X_1. Using $\underline{y}$ again as the indicator of correct decision, the probability of the error made by this decision is given by

$$P\{\underline{y}=0,\underline{r}_1=n_1\} = \sum_{\substack{j=1\\ j\neq k}}^{P} P\{\underline{r}_1=n_1|\underline{s}_x=j\}P\{\underline{s}_x=j\}. \tag{IV.8}$$

Defining $\mathcal{R}_1^{X_1}, \mathcal{R}_2^{X_1}, \cdots \mathcal{R}_P^{X_1}$ as the rank score vector sets for which the classifier decides on class labels $S_1, S_2, \cdots, S_P$ respectively, the total probability of error for the first classifier is given by

$$\begin{aligned} P_{X_1}^e &= \sum_{n_1\in\mathcal{R}_1^{X_1}} P\{\underline{y}=0,\underline{r}_1=n_1\} + \sum_{n_1\in\mathcal{R}_2^{X_1}} P\{\underline{y}=0,\underline{r}_1=n_1\} \\ &\quad + \cdots + \sum_{n_1\in\mathcal{R}_P^{X_1}} P\{\underline{y}=0,\underline{r}_1=n_1\}. \end{aligned} \tag{IV.9}$$

By using Eq. (IV.8) in Eq. (IV.9), and assuming uniform class distribution $P\{\underline{s}_x = j\} = 1/P$ we obtain

$$P^e_{X_1} = \frac{1}{P}\left\{ \sum_{n_1 \in \mathcal{R}_1^{X_1}} \sum_{j=2}^{P} P\{\underline{r}_1 = n_1 | \underline{s}_x = j\} + \sum_{n_1 \in \mathcal{R}_2^{X_1}} \sum_{\substack{j=1 \\ j \neq 2}}^{P} P\{\underline{r}_1 = n_1 | \underline{s}_x = j\} + \cdots + \sum_{n_1 \in \mathcal{R}_P^{X_1}} \sum_{j=1}^{P-1} P\{\underline{r}_1 = n_1 | \underline{s}_x = j\} \right\} \tag{IV.10}$$

The same error probability analysis may be done for the remaining classifiers. Now consider the multiple classifier system instead of the individual classifiers. Assume without loss of generality that classifier X_1 is the dominating classifier. Let a random vector $\underline{\mathbf{r}} = [n_1\ n_2\ \cdots\ n_Q]^T$ represent the rank score output of the classifier set. Here each entry represents the index of some rank score vector of a specific classifier among Q classifiers. Also let Ω be the set of all allowable rank score vectors for a single classifier. Finally, define $\mathcal{R}_p^C$ for $p = 1, 2, \cdots, P$ as

$$\mathcal{R}_p^C = \left\{ \mathbf{n} = [n_1\ n_2\ \cdots\ n_Q] \,|\, n_1 \in \mathcal{R}_p^{X_1}, n_k \in \Omega\ ; k = 2, \cdots, P \right\} \tag{IV.11}$$

With these definitions, the total probability of error for the combined system can be expressed as

$$P^e_{Comb} = \frac{1}{P}\left\{ \sum_{\mathbf{n} \in \mathcal{R}_1^C} \sum_{j=2}^{P} P\{\underline{\mathbf{r}} = \mathbf{n} | \underline{s}_x = j\} + \sum_{\mathbf{n} \in \mathcal{R}_2^C} \sum_{\substack{j=1 \\ j \neq 2}}^{P} P\{\underline{\mathbf{r}} = \mathbf{n} | \underline{s}_x = j\} + \cdots + \sum_{\mathbf{n} \in \mathcal{R}_P^C} \sum_{j=1}^{P-1} P\{\underline{\mathbf{r}} = \mathbf{n} | \underline{s}_x = j\} \right\} \tag{IV.12}$$

$$
\begin{aligned}
P^e_{Comb} \;=\; \tfrac{1}{P}\Bigg\{ & \sum_{n_1 \in \mathcal{R}_1^{X_1}} \sum_{j=2}^{P} \sum_{\substack{n_k \in \Omega \\ k=2,3,\cdots,P}} P\{\underline{\mathbf{r}} = [n_1\ n_2\ \cdots\ n_P] | \underline{s}_x = j\} \\
& + \sum_{n_1 \in \mathcal{R}_2^{X_1}} \sum_{\substack{j=1 \\ j\neq 2}}^{P} \sum_{\substack{n_k \in \Omega \\ k=2,3,\cdots,P}} P\{\underline{\mathbf{r}} = [n_1\ n_2\ \cdots\ n_P] | \underline{s}_x = j\} \\
& + \cdots + \sum_{n_1 \in \mathcal{R}_P^{X_1}} \sum_{j=1}^{P-1} \sum_{\substack{n_k \in \Omega \\ k=2,3,\cdots,P}} P\{\underline{\mathbf{r}} = [n_1\ n_2\ \cdots\ n_P] | \underline{s}_x = j\} \Bigg\}
\end{aligned}
\tag{IV.13}
$$

The inner summations being over all the rank score vectors of the remaining classifiers, they are equal to the probability $P\{\underline{r}_1 = n_1 | \underline{s}_x = j\}$. Therefore Eq. (IV.13) becomes equal to Eq. (IV.10) and this establishes $P^e_{Comb} = P^e_{X_1}$, proving the Theorem. □

Theorem IV.3 shows for the general case that if one classifier dominates the others, no improvement can be expected from the combination of the classifiers. In other words, for improvement by combination, no classifier should dominate, i.e., the jointly optimal decision should favor each classifier's decision in turn, for some rank score matrices. This is expressed by Theorem IV.4 which makes use of Lemma IV.1.

LEMMA IV.1 *Due to the joint optimality of the combined decision, the combined performance cannot be lower than the performance of the best classifier within a multiple classifier system.*

PROOF. To show this, let X_1 be the best individual classifier and let $\mathcal{R}_1^C$ be the set of all rank score matrices for which X_1 decides on class label S_1 as given by

$$
\mathcal{R}_1^C = \Big\{ \mathbf{n} = [n_1\ n_2\ \cdots\ n_Q] \,|\, n_1 \in \mathcal{R}_1^{X_1}, n_k \in \Omega\ ; k = 2, \cdots, P \Big\}. \tag{IV.14}
$$

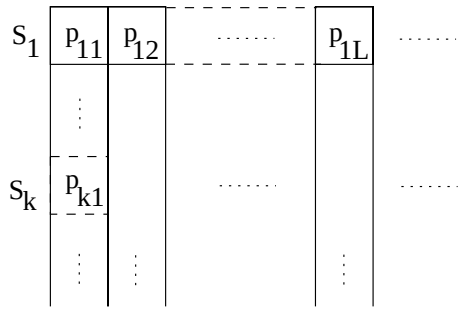

Figure IV.3: Part of the joint distribution of the observation space used for Lemma IV.1.

Without loss of generality, the rank score matrices can be ordered such that these $L = |\mathcal{R}_1^C|$ rank score matrices correspond to $\underline{r} = 1, 2, \cdots, L$. The corresponding part of the joint distribution of the observation space is illustrated in Figure IV.3. If the conditions $p_{1n} > p_{jn}$ for $j = 2, 3, \cdots, P$ and $n = 1, 2, \cdots, L$ are satisfied, then the jointly optimal decision is equivalent to the decision of X_1 for this set of $\underline{r}$ values.

Suppose we try to disturb this condition by letting $p_{k1} > p_{11}$ for the $\underline{r} = 1$. This largest probability term will contribute to the probability of error made by X_1. However, it will not contribute to the probability of error made by the optimal decision since the optimal decision will select S_k for $\underline{r} = 1$. Therefore, the error for the optimum decision will necessarily be *lower* than the error for the best classifier X_1. □

THEOREM IV.4 *If none of the classifiers in a Q classifier ensemble dominate the ensemble, then we necessarily have* $P^e_{Comb} < \min\{P^e_{X_1}, P^e_{X_2}, \cdots, P^e_{X_Q}\}$.

PROOF. Without loss of generality, assume that classifier X_1 is the best performing individual classifier. However, it is not a dominating classifier since there is

none. Define $\mathcal{D}^q(\mathbf{r}_q^l)$ to represent the decision of classifier X_q for the specific rank score vector $\mathbf{r}_q^l$ while $\mathcal{D}^C(\mathbf{R}_l)$ denotes the jointly optimal decision for the specific rank score matrix $\mathbf{R}_l = [\mathbf{r}_1^l\ \mathbf{r}_2^l\ \cdots\ \mathbf{r}_Q^l]$.

The fact that X_1 is not a dominating classifier means that there exist at least one or more rank score matrices $\mathbf{R}_l$ such that $\mathcal{D}^C(\mathbf{R}_l) \neq \mathcal{D}^1(\mathbf{r}_1^l)$. For each such rank score matrix, an intermediate, *partially optimal* decision process $\hat{\mathcal{D}}^l$ can be designed which satisfies $\hat{\mathcal{D}}^l(\mathbf{R}_l) = \mathcal{D}^C(\mathbf{R}_l)$ while for all other rank score matrices its decisions coincide with the decision of classifier X_1, i.e., $\hat{\mathcal{D}}^l(\mathbf{R}_k) = \mathcal{D}^1(\mathbf{r}_1^k)$, $\forall \mathbf{R}_k \in \mathcal{R},\quad k \neq l$. By Lemma IV.1, the partially optimized decision process cannot yield a performance lower than the performance of the best individual classifier. Therefore, such a decision process which is *different* than the best individual classifier should necessarily yield to an improved performance. □

Another result of this section about dominance is given by Corollary IV.1.

COROLLARY IV.1 *If there is a dominating classifier within a multiple classifier system, then this is necessarily the best performing individual classifier.*

PROOF. By Theorem IV.3, the performance of the dominating classifier equals the performance of the combination. However, by Lemma IV.1, the performance of the combination cannot be lower than the best individual performance. Therefore, the performance of the dominating classifier equals the performance of the best classifier, proving the Corollary. □

The above discussion suggests that output independence plays no exclusive role in assessing the potential for improvement by the combination of classifiers. However, a different concept the thesis defines as the *dominance of a classifier*

gives a condition on classifier complementariness. Namely, one should have no dominating classifier in a given classifier ensemble in order to have performance improvement by optimal combination in the sense of Chapter III.

IV.5 Complementariness of Classifiers

The previous section defined a condition for achieving complementary behavior among classifiers and hence, to obtain an improvement from classifier combination. However, the fact that none of the classifiers are *dominating*, does not give one, a measure on the potential improvement possible by the combination of a set of classifiers. In the present section, an attempt is made to introduce such a measure.

Consider again Figure IV.2. Apart from the probability of correct classification, another measure on the performance of an individual classifier X_k may be given by means of the mutual information $I(\underline{r}_k, \underline{s}_x)$ between the classifier output $\underline{r}_k$ and the source class $\underline{s}_x$. I.e., it may be argued that the *amount of information acquired about the true class label by observing the outputs of classifier* X_k is a reasonable measure on that classifier's performance.

Now consider that while using X_k individually, one asks the question: *How much does classifier* X_l *has a potential to complement the present classifier* X_k*?*. This depends on the ability of X_l to provide *additional* information about the source class label. I.e., one should be interested in *the amount of new information provided by the output of* X_l *which was not present in the output of* X_k. This quantity can be expressed as a difference

$$\Delta I_{X_k X_l} \doteq I(\underline{r}_k, \underline{r}_l; \underline{s}_x) - I(\underline{r}_k, \underline{s}_x), \tag{IV.15}$$

where the first term represents the amount of information acquired about the source class label $\underline{s}_x$ by observing both classifier outputs $\underline{r}_k$ and $\underline{r}_l$ while the last term represents the amount of information acquired about the source class label by observing the output of classifier X_k alone. Replacing both mutual information terms by their entropy definitions as given in Eq. (IV.5) one gets

$$\Delta I_{X_k X_l} = H(\underline{s}_x|\underline{r}_k) - H(\underline{s}_x|\underline{r}_k, \underline{r}_l). \tag{IV.16}$$

which can be expressed in expanded form as

$$\Delta I_{X_k X_l} = \sum_{j,n_1,n_2} P\{\underline{s}_x = j, \underline{r}_1 = n_1, \underline{r}_2 = n_2\} \log \frac{P\{\underline{s}_x = j|\underline{r}_1 = n_1, \underline{r}_2 = n_2\}}{P\{\underline{s}_x = j|\underline{r}_1 = n_1\}}. \tag{IV.17}$$

The quantity we have defined in Eq. (IV.15) is not symmetric, namely, we have $\Delta I_{X_k X_l} \neq \Delta I_{X_l X_k}$. This is a reasonable behavior since for classifiers with different performances, the amount of information contributed by X_l to X_k cannot be the same as the amount contributed by X_k to X_l. One expects the contribution of the better performing classifier to be larger.

The quantity defined by $\Delta I_{X_k X_l}$ can be proposed as a measure of the complementariness of classifier X_l with respect to classifier X_k. This proposal is supported by investigating the behavior of the aforementioned measures on several simulated examples with two classifiers and two classes. The joint distributions and the derived marginal distributions for these five examples are illustrated in

Table IV.7. Three of these distributions can be recognized from Examples IV.1 and IV.2. The given simulated cases are selected such that the performance and the marginal classifier observation space distribution for classifier X_1 is always the same, while they vary for the second classifier X_2.

Consider the following scenario while investigating Tables IV.7 and IV.8. One is restricted to use only two classifiers in parallel for this two class illustrative problem. Five different classifiers are available and the best classifier is labeled X_1. The task is to select the second classifier X_2 among the available ones which is the *most complementary* with respect to the best classifier X_1. I.e., the largest performance improvement over the performance of the best classifier is sought. For this purpose, each alternative classifier is operated in parallel with the best one and the distributions in Table IV.7 are obtained. From these distributions, the measures in Table IV.8 are obtained where all logarithms are Base 2 logarithms. This gives a measurement unit of *Bits*. One can make the following discussions.

For this two class problem with uniform class distribution, the entropy of the source random variable $\underline{s}_x$ is 1 bit, which is hence the maximum value for all measures in Table IV.8 based on Information Theory. For Case 1, the best classifier is dominating the pair since the optimal decision on the joint distribution is the same as the decision of the best classifier X_1 for all cases. Therefore, the candidate classifier cannot contribute to the best classifier and so here is no performance improvement. However, it is interesting to note that the $\Delta I_{X_1X_2}$ column still reports a positive value. It can be argued that the dominance condition may not be reflected in $\Delta I_{X_1X_2}$.

For the remaining cases which are ordered with respect to the actual perfor-

Table IV.7: Five simulated example cases. Joint and individual classifier observation space distributions are illustrated as three columns. Measures computed from these distributions are given in Table IV.8.

	Joint				X_1		X_2	
Case 1	0.48	0.32	0.12	0.08	0.80	0.20	0.60	0.40
	0.04	0.16	0.16	0.64	0.20	0.80	0.20	0.80
Case 2	0.51	0.29	0.09	0.11	0.80	0.20	0.60	0.40
	0.12	0.08	0.08	0.72	0.20	0.80	0.20	0.80
Case 3	0.70	0.10	0.07	0.13	0.80	0.20	0.77	0.23
	0.08	0.12	0.11	0.69	0.20	0.80	0.19	0.81
Case 4	0.66	0.14	0.14	0.06	0.80	0.20	0.80	0.20
	0.09	0.11	0.11	0.69	0.20	0.80	0.20	0.80
Case 5	0.45	0.35	0.15	0.05	0.80	0.20	0.60	0.40
	0.15	0.05	0.05	0.75	0.20	0.80	0.20	0.80

Table IV.8: Intermediate measures of interest for the examples with two classes and two classifiers, given in Table IV.7. The complementariness of classifier X_2 with respect to classifier X_1 is given in the column labeled as $\Delta I_{X_1X_2}$ and is the primary measure of interest.

	Ind. Perf.		Indiv. Infor.		Independence	Joint Infor.	Complementariness		Improv.
Case	$P^e_{X_1}$	$P^e_{X_2}$	$I(\underline{r}_1, \underline{s}_x)$	$I(\underline{r}_2, \underline{s}_x)$	$I(\underline{r}_1, \underline{r}_2)$	$I(\underline{r}_1, \underline{r}_2; \underline{s}_x)$	$\Delta I_{X_1X_2}$	$\Delta I_{X_2X_1}$	ΔP^e_{Comb}
Case 1	0.2	0.3	0.2781	0.1245	0.0000	0.3888	0.0807	0.2343	0.000
Case 2	0.2	0.3	0.2781	0.1245	0.0846	0.3204	0.0423	0.1959	0.005
Case 2	0.2	0.21	0.2781	0.2591	0.1008	0.3592	0.0811	0.1001	0.010
Case 4	0.2	0.2	0.2781	0.2781	0.0359	0.4033	0.1252	0.1252	0.015
Case 5	0.2	0.3	0.2781	0.1245	0.1538	0.4319	0.1538	0.3073	0.050

mance improvement over the best, the best classifier is not dominating. Also, the $\Delta I_{X_1X_2}$ column seems to reflect the potential improvement achievable by combination. Investigating the output independence column $I(\underline{r}_1, \underline{r}_2)$ supports that output independence is not necessarily a desired condition for complementariness. Case 5 shows that the maximum improvement given in the Table is for the candidate classifier which has the maximum dependence with the best classifier. Again a considerable improvement is possible for Case 4, where the output dependence between classifiers is quite low. A last observation on IV.8 is that the complementing classifier performance need not necessarily be very close to the performance of the best classifier for improvement to be possible. Again the maximum improvement is achieved by a complementing classifier with $p^e = 0.3$ while a much smaller improvement could be achieved with a much better performing classifier with $p^e = 0.21$.

IV.6 Discussion

In the present chapter, the thesis attempted to clarify the concepts of *output independence* and *complementariness* and their relations with the actual performance improvement achievable by optimal combination. The following have been the main contributions. Firstly, an Information Theoretic interpretation of a multiple classifier system is introduced and this enabled the use of measures from information theory to quantify relations between random variables representing events within such a system. A measure for classifier output dependence is developed under this framework and it is shown that output independence plays no exclusive role in determining how much a classifier can complement another. A new

concept called as *dominance of classifier* is introduced to give a critical condition for performance improvement. Finally, another Information Theoretic measure is introduced to quantify the potential for improvement in such a system which have been supported by empirical justification. However, not all the questions raised within the scope of this Chapter could be answered and there exist several issues open for further research. For example, the concept of *error independence* and its relation with performance improvement through combination remains an open issue. Also, the theoretical relation between the complementariness measure $\Delta I_{X_1 X_2}$ and the actual improvement remains to be established. These points are viable directions for future research.

CHAPTER V

Speech Feature Extraction and Individual Classifiers for Speech Pattern Recognition

Part of any classifier is a feature extraction stage which is followed by a similarity scoring stage. During feature extraction, the pattern is transformed into a set of *descriptive features* for that pattern while during similarity scoring, this set of features is compared with a set of *class models* to determine the similarity between the pattern and the pattern classes whose models are known to the classifier.

This basic operating model which is illustrated in Figure V.1 is also valid for classifiers operating on speech patterns, both for speaker identification and for speech recognition. In this chapter, the thesis discusses these two building blocks of a classifier for speech pattern recognition, which form the common computational basis for the experiments in both speaker identification and speech recognition. The specific considerations for each task and the experimental results will be discussed in detail in Chapters VI and VII.

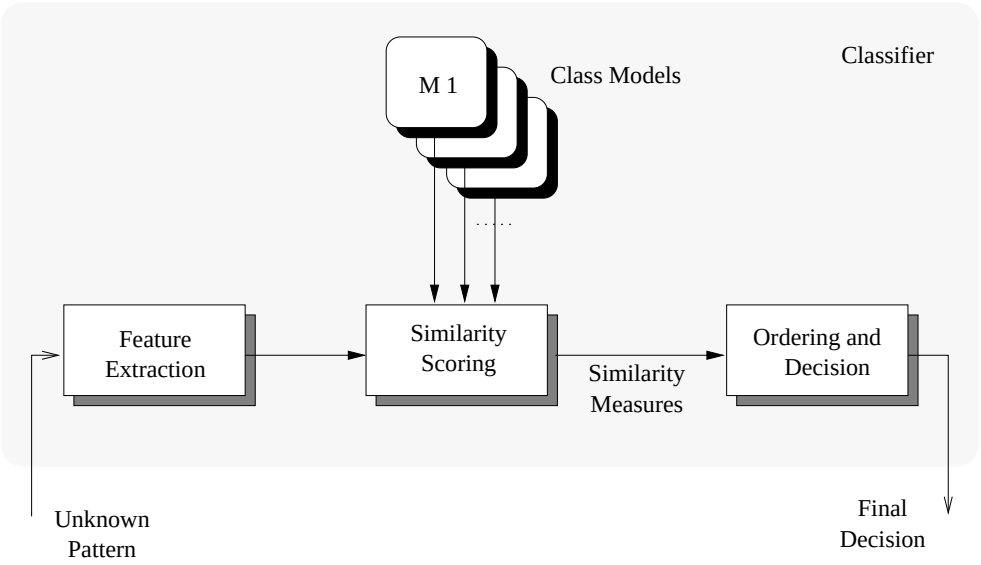

Figure V.1: The typical classifier operation model. In most classifiers, the pattern is first transformed into a set of descriptive features, then a similarity scoring method is used to compare these features against a set of stored models of pattern classes. The class with the highest similarity score is identified by the classifier.

V.1 Feature Extraction for Speech Signals

The *speech pattern* is basically the speech signal itself, sampled at a suitable sampling rate and quantized to a suitable bit-length so that the intelligibility of the human speech is preserved. The speech signal is not exactly band-limited. However, its exponentially decaying average spectrum [58] allows an assumption of band-limitedness. The so called *telephone-quality* speech is assumed to be band-limited to $3.4kHz$ and hence most telephone lines low-pass filter the speech signal with a cut-off frequency of $3.4kHz$. Also a quantization of 8 $bits/sample$ with A-law encoding, is often used for telephone-quality speech.

Since the telephone quality speech preserves both the message and the personality as perceived by a human listener, both automatic speaker identification and speech recognition should be theoretically possible by using a speech signal

sampled at the Nyquist rate of 8000 *samples/second* and at a quantization of 8 *bits/sample*. Therefore, for this study, these parameters are chosen for the digitization of the speech signal.

Followed by this digitization procedure, the sampled and quantized speech waveform is processed to extract a set of descriptive features. These are often in vector form for this pattern. Since human speech is an information bearing signal, its behavior and descriptive *features* change over time, i.e., it is not stationary [59]. Therefore, to capture the behavior of the speech signal over time, speech features are often selected as a *time sequence of feature vectors* where each vector is extracted through necessary processing, from successive short segments of the speech signal. These uniform length segments of speech are called as *frames*. The framed speech can be defined for all time as

$$s_m[n] = s[n+m]w[n], \tag{V.1}$$

where $w[n]$ is a *windowing function* and $s[n]$ is the speech signal which is shifted by m samples so that segment being analyzed overlaps with the non-zero window. This definition generates a new signal which is defined for $-\infty < n < +\infty$ and which is formally suitable for analysis methods which inherently assume that the signal is defined for all n. A formal treatment of this topic can be found in [59]. The windowing function $w[n]$ determines the frame length N and is often chosen such that it tapers the speech signal corresponding to the frame towards the frame end-points. The rationale behind the tapering shape of most windowing functions is to minimize the adverse effects of *cutting-out* a segment of the signal from

its surroundings on the frequency resolution of spectral estimates, which form the basis of most speech signal features. A discussion of windowing function selection and its consequences on the spectral estimates can be found in [60]. This framing process is illustrated in Figure V.2 on an actual speech signal with a Hamming windowing function. In this study, we use a frame length of $N = 512$ samples where successive frames overlap by 128 samples which are determined by preliminary experiments in speaker identification discussed in Chapter VI.

By this framing process, the entire speech signal is transformed into a set of *speech frames* which often overlap by the selection of successive m values satisfying $m < N$. Each speech frame is transformed into a *feature vector* whose number of elements are much smaller than the number of nonzero samples in the speech frame. This transformation is a data reduction process which tries to preserve information in the speech frame relevant for the problem and task considered, while irrelevant information (including the effects of noise) is tried to be eliminated [1]. In this thesis, three such feature extraction methods are considered to form the basis of the three classifiers used for the experimental evaluations. They will be discussed in the following three sections.

V.1.1 FFT Derived Cepstral Coefficients (FFTCep)

The *cepstral coefficients* rely on a signal processing concept known as *homomorphic signal processing* [60]. Homomorphic systems form a class of nonlinear systems where a generalized superposition is satisfied with different input and output operators. Namely, we have

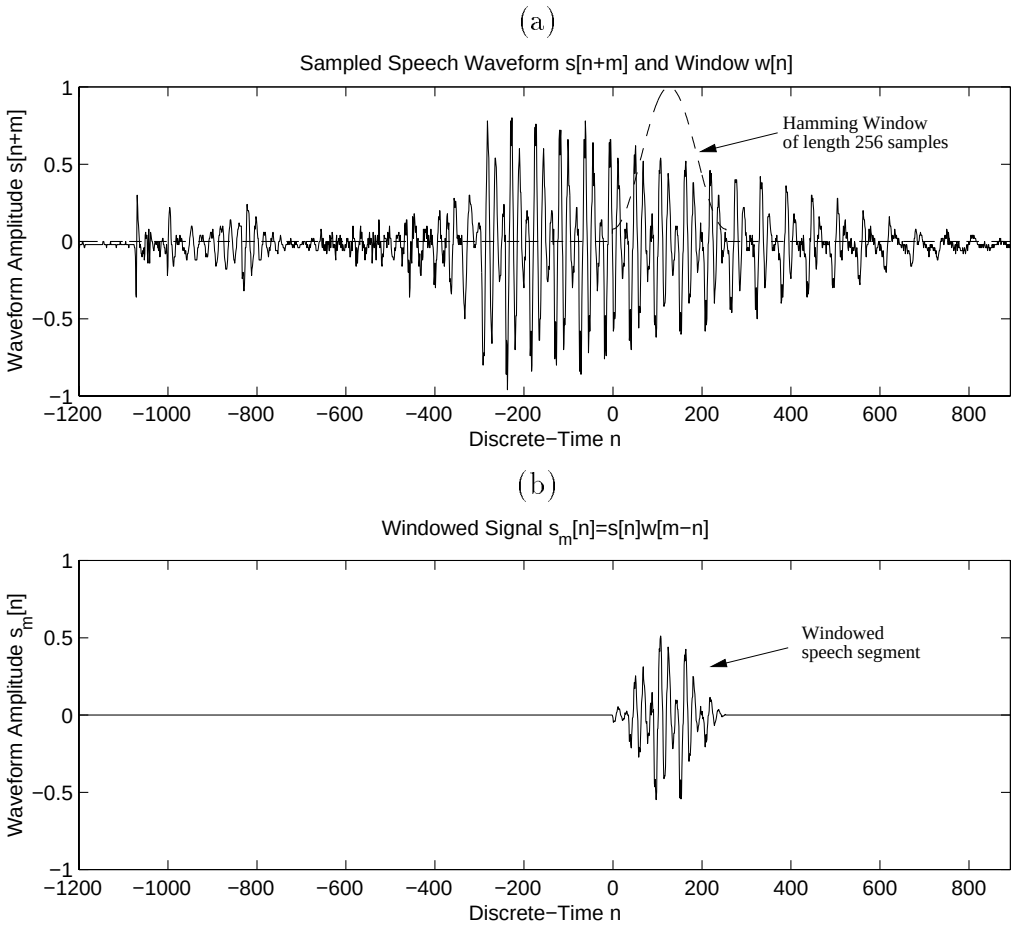

Figure V.2: Framing process for feature extraction. (a) The speech signal together with a shifted multiplying windowing function of a suitable shape. (b) The new signal obtained by windowing the original signal. A short-time feature can be obtained by various transformations on this new signal.

$$H\{x_1[n] \diamond x_2[n]\} = H\{x_1[n]\} \circ H\{x_2[n]\} \tag{V.2}$$

Consider the convolution of two signals $x_1[n] * x_2[n]$. Using a homomorphic transformation with an appropriate choice of the *characteristic system* $H[.]$, it is possible to transform this convolution to a simple *addition* between two transformed signals. In this transformed domain, the two transformed signals appear to be linearly combined, which is much easier to separate.

The speech signal is often considered to be the result of a *glottal excitation signal* driving a linear all-pole filter which is a simplified and linearized model of the human vocal tract. This model of speech production is illustrated in Figure V.4. If we consider the *discrete-time Fourier transform* operator $\mathcal{F}\{.\}$, we have the property

$$\mathcal{F}\{g[n] * v[n]\} = \mathcal{F}\{g[n]\} \cdot \mathcal{F}\{v[n]\}, \tag{V.3}$$

while the $\log|.|$ operator has the property

$$\log(|G(e^{-jw}) \cdot V(e^{-jw})|) = \log(|G(e^{-jw})|) + \log(|V(e^{-jw})|). \tag{V.4}$$

Using properties V.3 and V.4, the convolution operation between two signals in the time-domain, can be transformed into a simple addition in a new domain as

$$\log(|\mathcal{F}\{g[n] * v[n]\}|) = \log(|\mathcal{F}\{g[n]\}|) + \log(|\mathcal{F}\{v[n]\}|). \tag{V.5}$$

This new domain is called the *cepstral domain* and the transformed signal $c[n]$ the *real cepstrum*. The normalized time index n in the cepstral domain is denoted

as *Quefrency* a term coined-up to describe "frequencies" in this new "frequency domain" [59]. Consider the assumption that the speech signal $s[n]$ is produced as the convolution of the glottal excitation with the vocal tract filter, as illustrated in Figure V.4. One important motivation behind the use of the cepstrum sequence as feature vector is to transform the speech signal so that the effects of the glottal excitation and vocal tract shape are more easily separated.

Since the *Discrete-Fourier Transform* for finite length sequences also has property V.3, the practical computation of the cepstrum of the speech frame is done by means of the *Fast Fourier Transform* algorithm. This results in a cepstrum sequence of length N samples where N is the length of the original frame. However, this sequence is often truncated to a much smaller length and the first term $c[0]$, which reflects the signal energy, is often dropped from the feature vector based on empirical observations of its unreliable behavior [1]. The cepstral features obtained by means of the FFT algorithm are called as the *FFT Derived Cepstral Coefficients*. In the present study, 12 cepstral coefficients $c[1]$ to $c[13]$ are used to form the first set of feature vectors. The extraction of this feature set is illustrated in Figure V.3 where all $N = 128$ samples of the real cepstrum is illustrated.

V.1.2 LPC Derived Cepstral Coefficients (LPCCep)

This set of features is based on the linear model of human speech production illustrated in Figure V.4. The all-pole linear filter $V(z^{-1})$ in this model is estimated through a process which is often called as the *Linear Prediction Analysis*. This estimation problem can be interpreted in a number of ways as argued

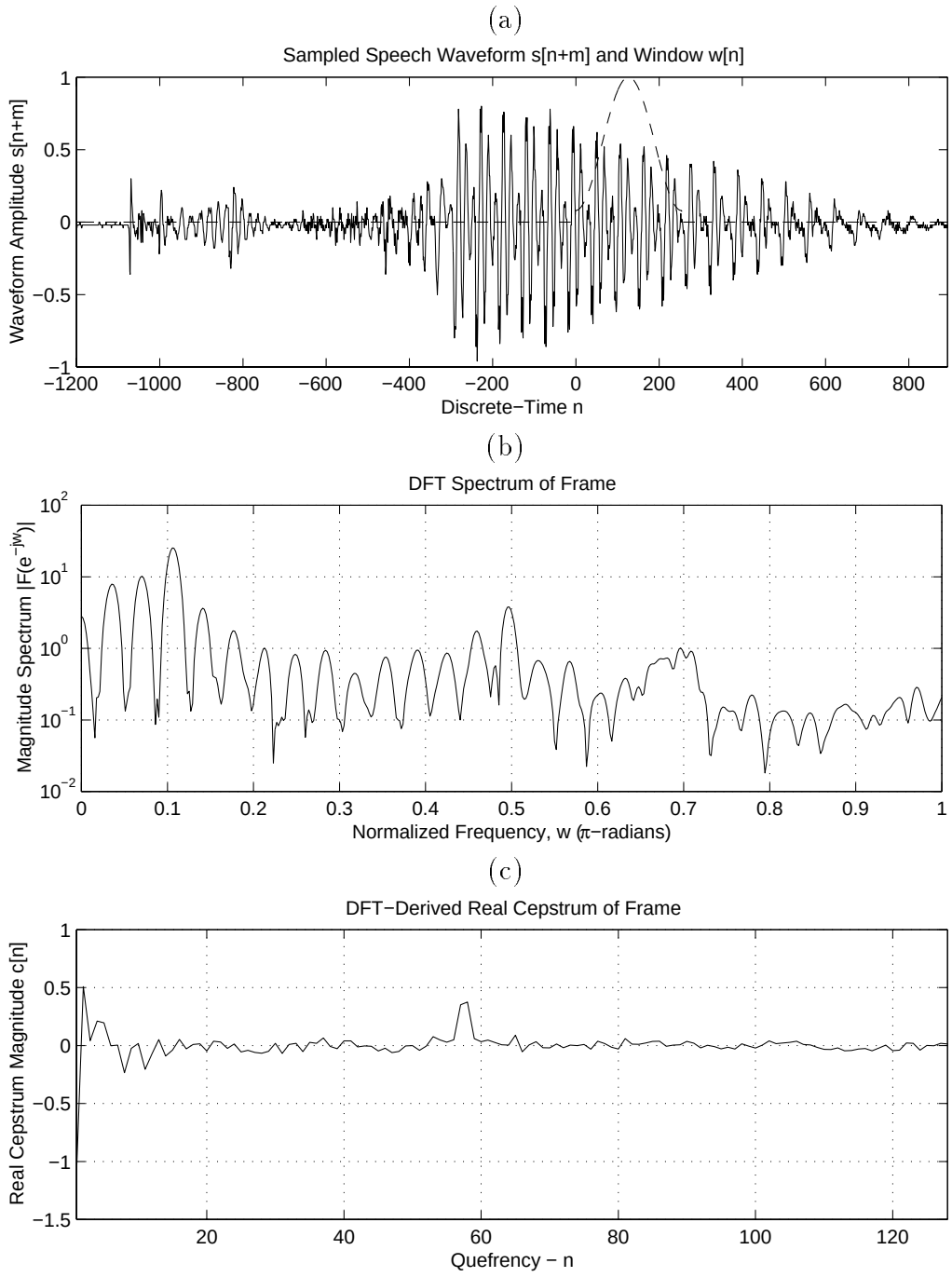

Figure V.3: FFT Derived Cepstral feature extraction. (a) A voiced speech frame with the corresponding windowing function. (b) The frame DFT spectrum which is an intermediate step in cepstrum extraction. (c) The FFT derived real cepstrum sequence of frame, illustrated for all $N = 128$ samples of the cepstrum. Only the coefficients $c[1], \cdots, c[13]$ are used as the feature vector. Note that both low and high quefrency components are present in the $c[n]$ sequence and the peak around $n = 58$ denotes the *pitch period* of the voiced frame, hence the excitation component in the speech.

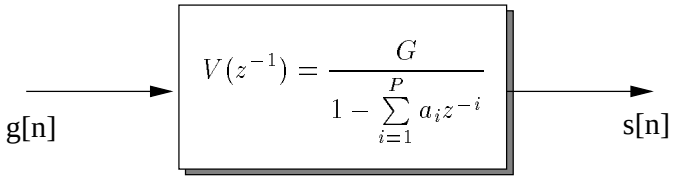

Figure V.4: The linearized speech production model. A glottal excitation signal drives a linear all-pole filter model of the human vocal tract. The speech signal is considered to be the output of this system, which is the convolution of the excitation signal with the filter impulse response.

by Deller et.al. [59], all leading to the same solution for the model parameters G and a_i, $i = 1, 2, \cdots, P$ where P is known as the *LP filter order*. One such interpretation is the *prediction interpretation* where the idea is to find the best linear predictor of the current speech sample $s[n]$ from its P past samples $s[n-1], s[n-2], \cdots, s[n-P]$. A short-time *average square prediction error* is defined as

$$E_m = \sum_n e_m^2[n] = \sum_n \left[s_m[n] - \sum_{i=1}^{P} a_k s_m[n-i] \right]^2 . \qquad \text{(V.6)}$$

where the analysis frame is defined as in Eq. (V.1). Assuming summation limits are $-\infty < n < +\infty$ and using Eq. (V.1) in Eq. (V.6) we obtain the error

$$E_m = \sum_{n=-\infty}^{+\infty} e_m^2[n] = \sum_{n=-\infty}^{+\infty} \left[s[n+m]w[n] - \sum_{i=1}^{P} a_i s[n+m-i]w[n-i] \right]^2 . \qquad \text{(V.7)}$$

The optimum a_i values minimizing the prediction error function in Eq. (V.7) can be found by solving the set of optimality equations $\delta E_m / \delta a_j = 0$, $j = 1, 2, \cdots, P$ which are given by

$$\sum_{n=-\infty}^{+\infty} s[n+m]w[n]s[n+m-j]w[n-j] = \sum_{i=1}^{P} a_i \sum_{n=-\infty}^{+\infty} s[n+m-i]w[n-i]s[n+m-j]w[n-j]. \tag{V.8}$$

for $j = 1, 2, \cdots, P$. This set of equations can be expressed in a more compact form by the definition of the *Autocorrelation function estimate*

$$R_m(j) = \sum_{-\infty}^{+\infty} s[n+m]w[n]s[n+m-j]w[n-j]. \tag{V.9}$$

Using also the fact that $R(.)$ is an even function, the optimality equations take the form

$$\sum_{i=1}^{P} a_i R_m(|i-j|) = R_m(j), \quad \text{for } j = 1, 2, \cdots, P. \tag{V.10}$$

which can be expressed in matrix form as

$$\begin{bmatrix} R_m(0) & R_m(1) & \cdots & R_m(P-1) \\ R_m(1) & R_m(0) & \cdots & R_m(P) \\ \vdots & \vdots & \ddots & \vdots \\ R_m(P-1) & R_m(P-2) & \cdots & R_m(0) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_P \end{bmatrix} = \begin{bmatrix} R_m(1) \\ R_m(2) \\ \vdots \\ R_m(P) \end{bmatrix}. \tag{V.11}$$

Since the autocorrelation matrix on the left hand side of Eq. (V.11) is *Toeplitz*, the optimality equations can be very efficiently solved using the *Levinson-Durbin recursion* [58], hence giving the model parameters $a_i, \quad i = 1, 2, \cdots, P$ which are called as the *Linear Prediction Coefficients* (LPC). The computation of the gain is done by considering the energy of the frame and is discussed in [58]. Since it will not be used as an element of the feature vector, the procedure is not detailed here.

The frequency response of the LP filter gives us the *envelope* of the short-time Fourier spectrum of the speech frame and represents the vocal tract part of the speech production model. Our second set of features called as the *LPC derived cepstral coefficients* considers only these parameters to compute a set of cepstral coefficients. Since the LP filter derived from the autocorrelation analysis is stable, these cepstral coefficients can be obtained directly from the LP coefficients by a recursive procedure [1],

1. *Initialize* $c_{LP}[1] = a_1$;

2. *For* $2 \leq i \leq N_c$ *Compute*

$$c_{LP}[i] = a_i + \sum_{j=1}^{i-1}(1 - \frac{j}{i})a_j c_{LP}[i-j].$$

Note that by this recursion, it is possible to compute an infinite number of cepstral coefficients. However, since these are derived from a limited number of LPC coefficients, it is only meaningful to keep a number of coefficients comparable to the LPC filter order. The extraction of the LPC Derived Cepstral features is illustrated in Figure V.5 .

V.1.3 LPC Residual Derived Cepstral Coefficients (LPCResCep)

The all-pole LPC filter is a linearized model of speech production. When the actual speech signal is passed through the inverse of the LPC filter, a *residual* (or *error*) signal is obtained. If the LPC model were able to capture all the information about the vocal tract, this residual would reflect the *excitation signal* at the vocal folds. However, LPC filter is only a linearized model. Therefore the

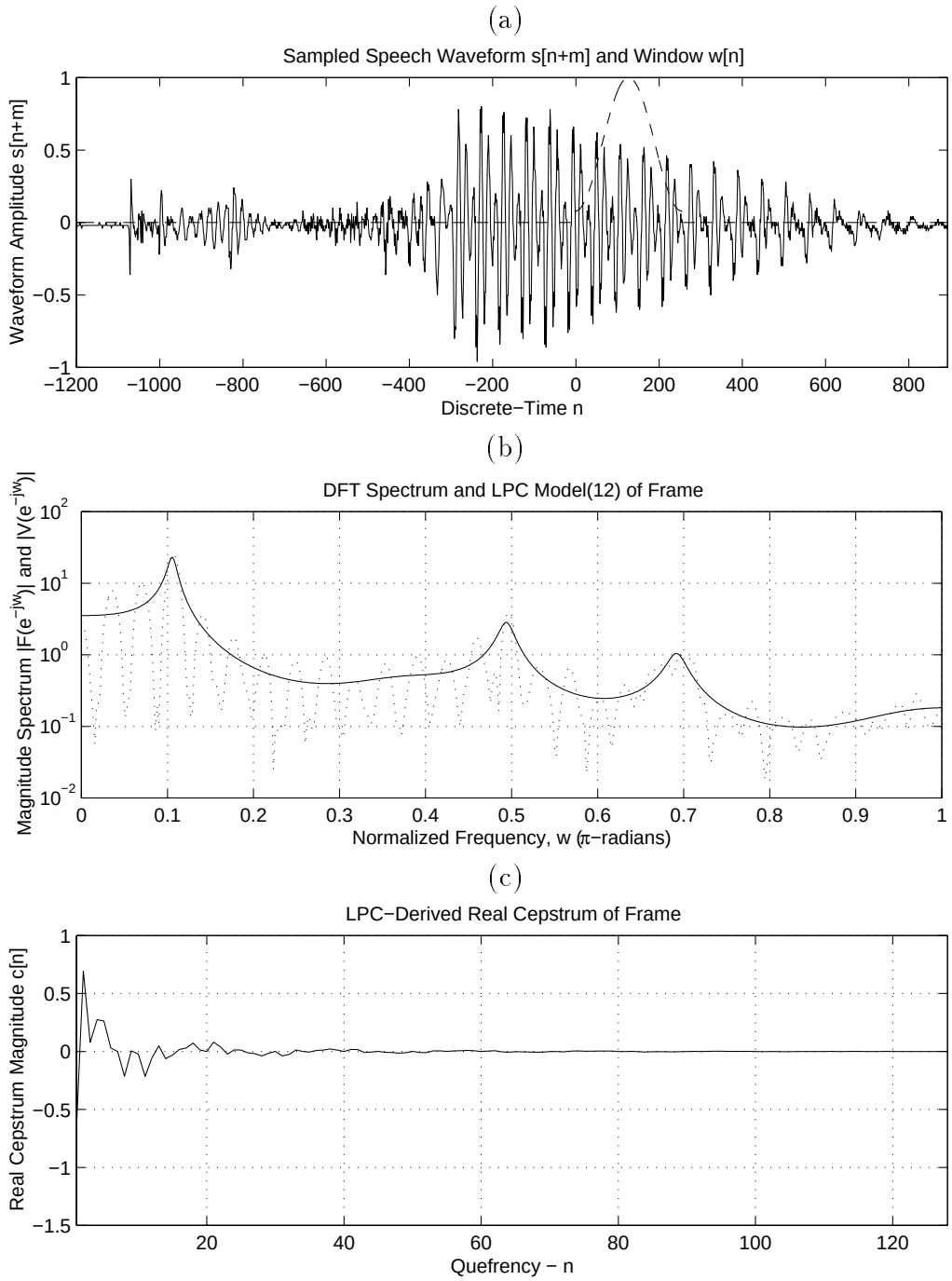

Figure V.5: LPC Derived Cepstral feature extraction. (a) The same voiced frame as in Figure V.3. (b) The LPC all-pole filter magnitude response for a filter order of $P = 12$ together with the frame DFT spectrum. (c) The LPC derived real cepstrum sequence of frame, illustrated for all $N = 128$ samples of the cepstrum. Only the coefficients $c[1], \cdots, c[13]$ are used as the feature vector. Note that high quefrency components (excitation) are removed by the use of the LPC model while low quefrency components (vocal tract) are preserved.

residual signal bears important information about the excitation *and* the vocal tract, which was not modeled by the LPC filter. In this sense, the residual signal can be thought to carry information which is complementary to the LPC model. A feature set can be constructed from the residual signal to represent this complementary information.

Following these arguments, we propose a feature set called as the *Residual Derived Cepstral Coefficients*. These are FFT Derived Cepstral Coefficients, extracted from the LPC residual signal. The feature extraction process is as follows: The speech frame is subjected to an LPC analysis and the LPC coefficients of a specified order are obtained. During this estimation process, the frame *prediction error signal* is obtained. This is equivalent to the signal which would be obtained by passing the speech frame from the inverse LPC filter, except for a scaling factor G. It is used to extract the FFT Derived Cepstral Coefficients, as described in Section V.1.1. Since FFT Derived Cepstrum does not involve any linearization of the underlying process, it is suitable to capture the nonlinear information present in the residual signal. This process is illustrated in Figure V.6.

V.2 Modeling and Similarity Scoring Methods

Once a time-sequence of feature vectors are obtained for all the training patterns of a class by a specific feature extraction procedure, this information is often transformed into a representative *compact model* for that class. This is called as the *modeling phase*. There exist a variety of modeling methods for speech pattern classes. These include *Vector Quantization*, *Hidden Markov Models* and a rich set of *Neural Network* architectures. Some of these methods can model the time

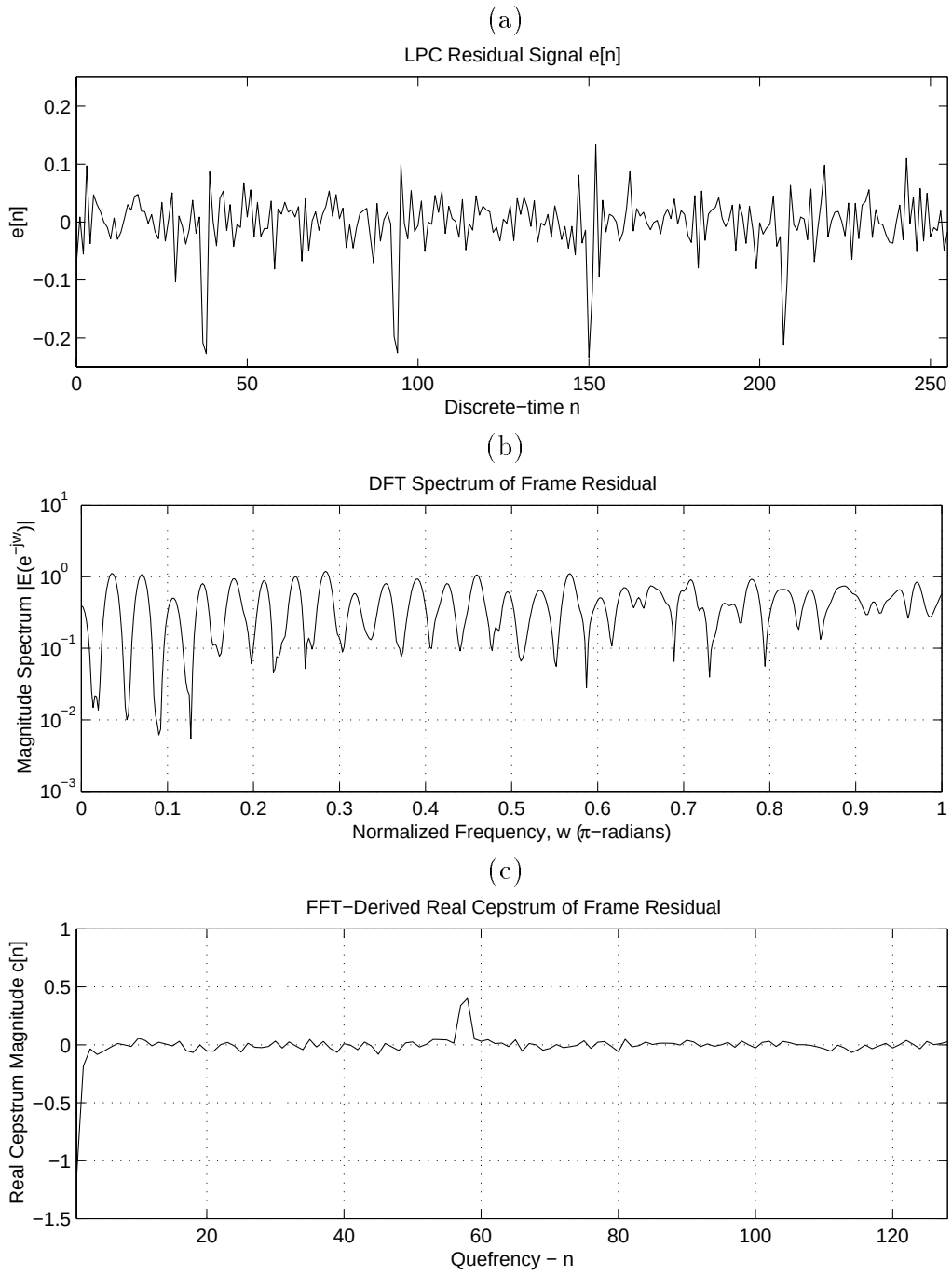

Figure V.6: LPC Residual Derived Cepstral feature extraction. The same voiced frame as in Figure V.3 is considered. (a) The LPC residual signal. (b) The FFT magnitude spectrum of the residual signal. (c) The FFT derived real cepstrum sequence of the frame residual, illustrated for all $N = 128$ samples of the cepstrum. Only the coefficients $c[1], \cdots, c[13]$ are used as the feature vector.

evolution of the speech patterns while some of them cannot. In the present thesis, the Vector Quantization, which does not model the time-evolution, is considered. [1]

V.2.1 Vector Quantization Class Models

A *Vector Quantizer* is a mapping $\mathcal{Q}$ which assigns each input vector $\mathbf{r} = \{r_1, r_2, \cdots, r_N\}$, a reproduction vector $\mathbf{y} = \mathcal{Q}(\mathbf{r})$ drawn from a finite reproduction alphabet $\hat{\mathcal{A}} = \{\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_M\}$ [61]. The reproduction alphabet $\hat{\mathcal{A}}$ is also known as the vector quantizer *codebook*. The quantizer $\mathcal{Q}$ partitions the input vector space into disjoint sets of vectors $V_i = \{\mathbf{r} : \mathcal{Q}(r) = y_i\}$ making up a partitioning $\mathcal{V} = \{V_1, V_2, \cdots, V_M\}$. The mapping $\mathcal{Q}$ is constructed by an optimal procedure such that the codebook vectors form a representative set for all the input vectors. In a multi-dimensional feature space where we have speech feature vectors, the *codewords* are often *centroids* of the clusters of the feature distribution.

In the present work, all the feature vectors from all patterns for a specific class are used to train a VQ codebook. For a P class problem, a VQ codebook model is trained for each pattern class S_j known to the system. When a pattern is given, the set of vectors extracted by the feature extraction algorithm are matched against these models by the similarity scoring procedure described in Section V.2.3.

[1] The choice of Vector Quantization leads to competitive performance for text-independent speaker identification task since the time evolution of the speech patterns are not very important due to independence from textual content. However, the performance for the BDEV task is below the figures reported in the literature since speech recognition is inherently tied to the time evolution of the speech signal. Nevertheless, this does not degrade the significance of the example since our focus is not on individual classifier performance but on the *improvement* over the best performing classifier by making use of classifier combination.

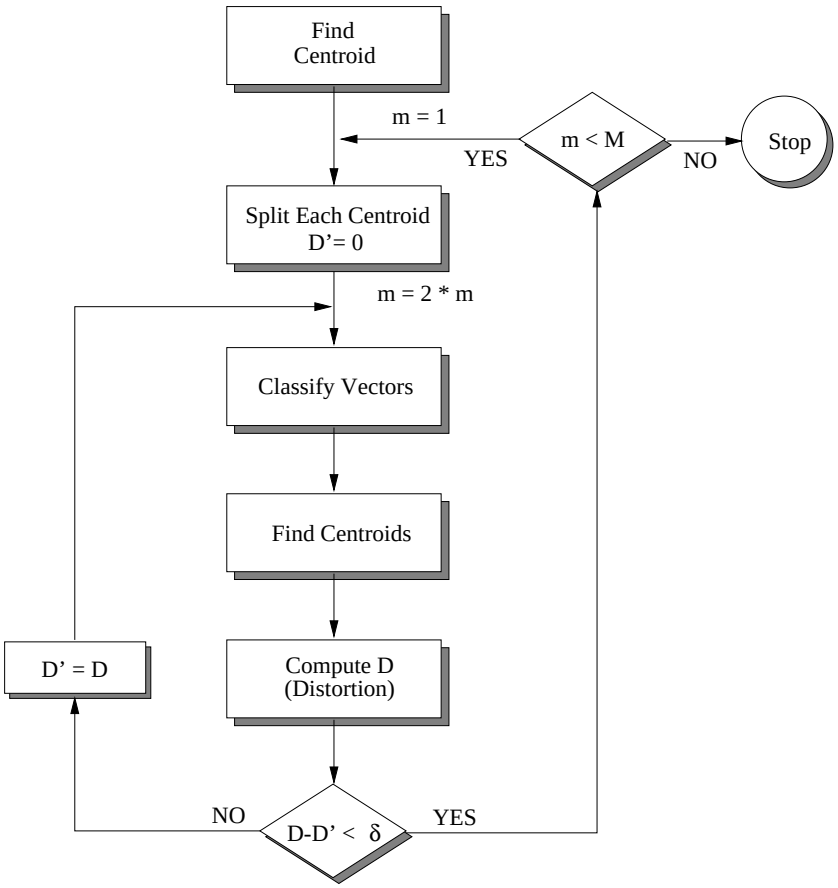

Figure V.7: Flow diagram for vector quantizer codebook training by the Binary-Splitting LBG algorithm. The algorithm operates by *splitting* the codebook into two at each iteration and computing the codebook distortion. A maximum codebook size can be specified to stop the training process. The codebook size is always a power of two.

V.2.2 Codebook Training (Modeling) Method

The construction of a VQ codebook which is optimal with respect to some distortion measure is studied in the literature. A study by Linde et.al., presents an algorithm which is a Binary-Splitting variant of the LBG algorithm well known in the speech processing literature [61]. The general flow diagram of the algorithm is illustrated in Figure V.7.

The algorithm is initialized by letting the reproduction alphabet size to $M = 1$

and finding the *centroid* or center of gravity of all the training vectors as the initial alphabet $\hat{\mathcal{A}}_0$. Then an outer iterative process is started where the alphabet size is doubled at each iteration. This is done by *splitting* each vector $\mathbf{r}_i$ into two close vectors $\mathbf{r}_i + \epsilon$ and $\mathbf{r}_i - \epsilon$ by introducing a small perturbation vector ϵ. This is repeated until a final alphabet size is reached. For each cycle of the iteration where the codebook size is m centroids, the following inner procedure, summarized in Figure V.7 as three inner blocks, is applied:

1. *Given the reproduction alphabet* $\hat{\mathcal{A}}_m = \{\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_m\}$, *find the minimum distortion partition* $\mathcal{P}(\hat{\mathcal{A}}_m) = \{Y_1, Y_2, \cdots, Y_m\}$ *of the training vectors so as to have* $\mathbf{r}_j \in Y_i$ *if* $d(\mathbf{r}_j, \mathbf{y}_i) \leq d(\mathbf{r}_j, \mathbf{y}_l), \forall l$, *where* $d(.)$ *is a suitable distance measure.*

2. *Find the optimum reproduction alphabet for this partitioning* $\mathcal{P}(\hat{\mathcal{A}}_m)$ *by computing a center of gravity, hence a centroid for each partition* Y_j, $j = 1, 2, \cdots, m$.

3. *Compute the average distortion for the partitioning* $\mathcal{P}(\hat{\mathcal{A}}_m)$ *with respect to the new reproduction alphabet* $\hat{\mathcal{A}}_m$ *as*

$$D = \frac{1}{N} \sum_{j=1}^{N} \min_{\mathbf{y} \in \hat{\mathcal{A}}_m} d(\mathbf{r}_j, \mathbf{y}).$$

4. *Compute the reduction in the distortion. Finish inner loop if* $D - D' \leq \delta$. *Otherwise, goto Step 1*

Note that the training algorithm necessitates a distance measure defined between two feature space vectors $\mathbf{r}, \hat{\mathbf{r}}$ as $d(\mathbf{r}, \hat{\mathbf{r}})$. The procedure is independent of the distance measure definition and the use of different measures is discussed in [61].

For the present thesis, the Euclidean distance, i.e., the l_2-norm is used and is given by

$$d(\mathbf{r}, \hat{\mathbf{r}}) = \sum_{i=1}^{p} |r_i - \hat{r}_i|^2. \tag{V.12}$$

V.2.3 Similarity Scoring Method

Once a VQ codebook is obtained for each pattern class, one needs to compute the similarity of a given set of feature vectors to each of the stored models. Let the class models (hence the VQ codebooks) be denoted by $S_1, S_2, \cdots, S_P$. The distance of a single vector $\mathbf{r}_i$ to the class model S_j can be measured by the distance to the nearest codebook vector of that model,

$$d(\mathbf{r}_i, S_j) = \min_{\mathbf{y} \in S_j} d(\mathbf{r}_i, \mathbf{y}) \tag{V.13}$$

For a set of unknown vectors, one has to consider a cumulative distance to the class models. For an unknown pattern x consisting of N feature vectors, the averaged cumulative distance to class model S_j will be given by the expression

$$D(x, S_j) = \frac{1}{N} \sum_{n=1}^{N} \min_{\mathbf{y} \in S_j} d(\mathbf{r}_n, \mathbf{y}). \tag{V.14}$$

The *similarity* of a set of feature vectors representing a pattern to a class model can be computed as the inverse of the distance given in Eq. (V.14). The class models can be ordered according to these *similarity scores* to provide a ranking of the candidate classes with respect to the unknown pattern x.

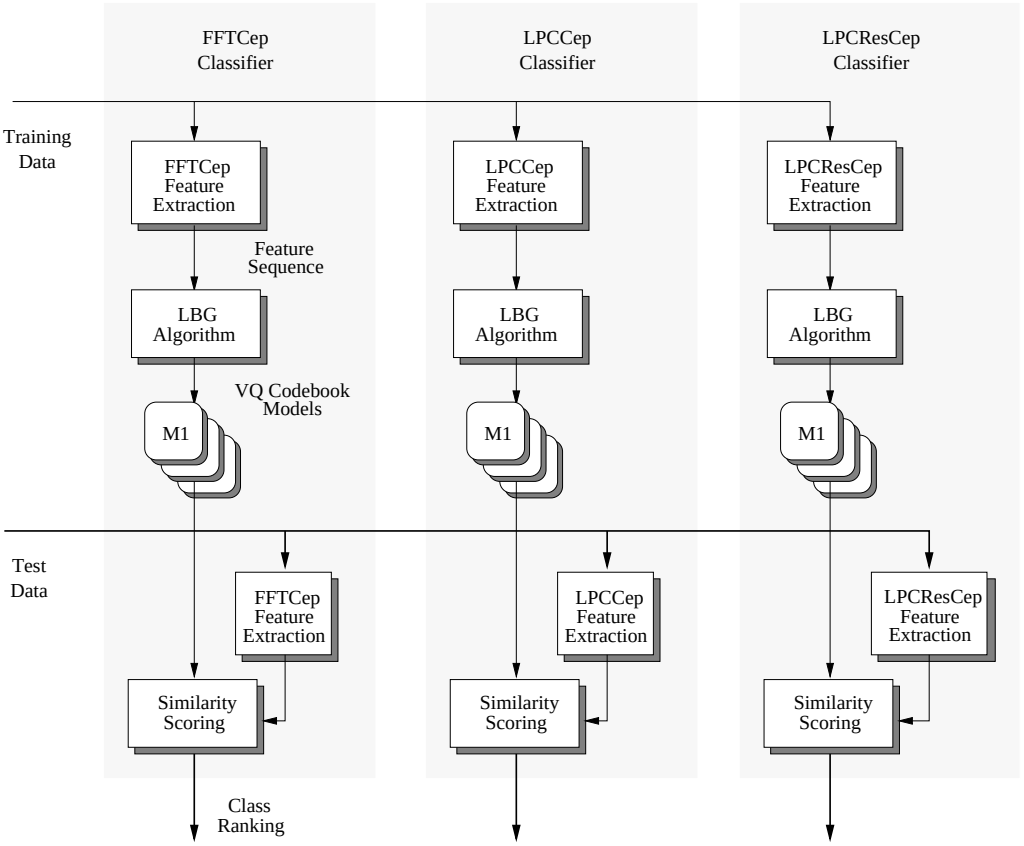

Figure V.8: The three individual classifiers based on three feature extraction methods. The training (modeling) and testing (similarity scoring) stages of the classifiers are illustrated.

V.3 Three Classifiers for Speech Pattern Recognition

The use of different feature extraction methods to generate meaningfully complementary behaving (i.e., in the sense of Chapter IV) classifiers is proposed by many researchers in the literature [46, 10, 14, 15].

Considering the three different speech feature extraction methods and the modeling and similarity scoring procedure just described, the three classifiers used in the present study can be defined. All three classifiers use the same modeling and similarity scoring procedure but differ in the feature extraction phase. They are illustrated in Figure V.8 for both the training phase and the testing phase.

The following two chapters presents the application of these three classifiers and their combination for two different pattern classification tasks from speech processing.

CHAPTER VI

Experiments on Closed-Set Text-Independent Speaker Identification Task

Two tasks from speech processing are considered as a test bed for the rank-based multiple classifier decision combination methods discussed in this thesis. The first of these is *automatic speaker identification*. This is the task of determining the identity of a person from his/her speech signal and is the topic of the present chapter.

VI.1 Task Description

Human speech sound, if appropriately sampled and digitized, contains a large amount of information. The most dominant information is clearly the spoken message. However, this is not the only information present in the signal. The identity of the speaking person, the language spoken, speaking disorders and emotional state of the speaker are all contained in this complex signal[5]. Specifically,

Speaker Identification is the process of using a machine to process a person's speech signal with the aim of automatically extracting his/her identity. The correlation between the identity of a person and his/her speech signal can be attributed to the *physiological properties* of the person as well as his/her *behavioral properties*[5].

One can try either to *identify* or to *verify* a speaker's identity. Identification is the task of determining the unknown identity while verification is task of determining whether or not a speech signal is uttered by a specific person. Also, these tasks can be attempted with unconstrained speech or with speech of known textual content, leading to tasks known as *text-independent* and *text-dependent* respectively. The present thesis considers the speaker *identification* task with *text-independent* system operation. Also, *closed-set* mode of operation is used, i.e., the system must make a decision on one of the known speaker classes and remaining undecided is not allowed. The speaker identification process is illustrated in Figure VI.1.

Here, the speech collected from *known* speakers are used to build models characterizing each of these speakers. For this purpose, properly sampled speech signal from each speaker is passed from a preprocessing or feature extraction stage where the speech pattern is compressed into a sequence of descriptive feature vectors. These feature vectors are then used by a modeling method to build the speaker models which are stored by the classifier. During operation, speech sample from an unknown identity speaker is passed from the same preprocessing stage and the feature vectors are matched against each of the speaker models by a similarity scoring procedure inside the classifier. The label of the model which

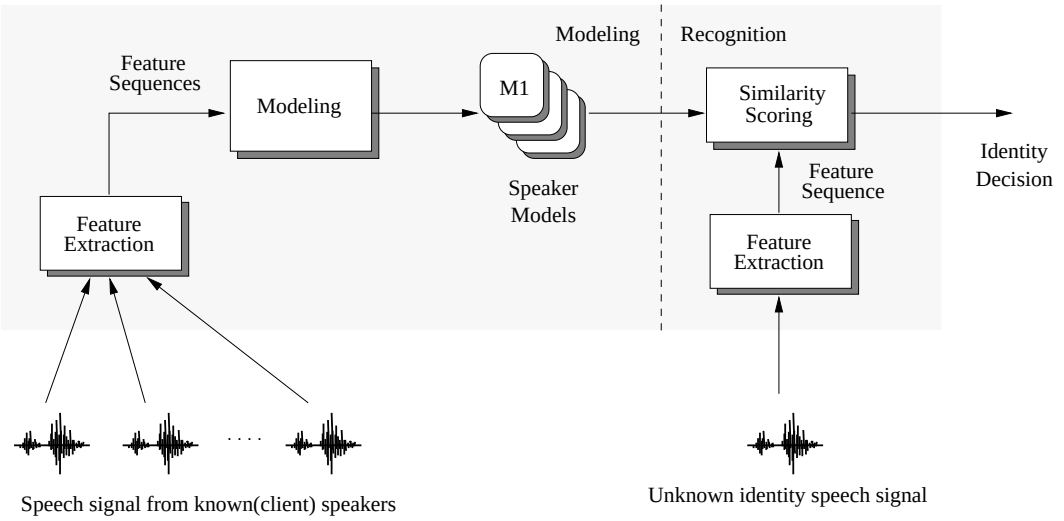

Figure VI.1: The speaker identification process. The speech from labelled speakers is used to build models for each speaker after they have been transformed into descriptive feature vectors. The models are then used in a similarity scoring procedure where feature vectors from an unknown speaker are matched against these models. The label of the best matching model identifies the speaker.

is the most similar to the given feature vectors identifies the speaker.

Speaker identification is a well studied problem in pattern recognition, especially within the speech processing community [4, 3, 62, 63, 64]. Many feature extraction and modeling/similarity scoring methods have been proposed in the literature, including *vector quantization* [5, 65], *hidden Markov models* [63, 66, 67], *Gaussian mixture models* [68, 69, 70, 71] and various methods based on artificial neural networks [72, 18, 73]. As described in Chapter V, three different speech feature extraction methods combined with vector quantization are used for the experiments in the present thesis.

VI.2 Database Description

A database tailored specifically for speaker identification and verification is used for the experiments of the present chapter. This is the POLYCOST Database which is the result of the joint efforts of the European COST-250 Action. [1] The database is described in detail in [74]. Its properties can be summarized as follows. The database consist of speech recordings over long distance international telephone lines. Speakers from the COST-250 member countries participated in the recording process. Approximately 10 speaker per country form a set of 74 male and 59 female speakers which called the recording system 1285 times. Each person has approximately 10 recording sessions. Each session for each speaker is composed of different types of recordings which are given in Table VI.1. The 10 recording sessions for each speaker is spread over the period of February-April 1996 with a minimum spacing of three days. All recordings are done with 8 kHz sampling rate with 8 bits/sample quantization. Several interesting variabilities are present in the database. Most of the speakers are non-native speakers of English and the MOT recordings are entirely in the mother tongue. International telephone lines introduce difficult variabilities between sessions. Also, the classical variabilities such as in frequential, temporal, intra and inter-speaker variabilities are present.

Speaker sets of 30 male speakers are used for the individual and multiple classifier experiments. These 30 male speaker subsets are selected among the 69 male speakers consisting the first CD-ROM of the database. An experiment

[1] Speech Processing Laboratory of the Department of Electrical and Electronics Engineering is a member of the European COST-250 Action titled "Speaker Recognition in Telephony" and has contributed to the POLYCOST Database.

Table VI.1: Recording items in the POLYCOST Database

Type	Items#	Description
CLI	4	7-digit client code
DIG	5	Sequence of 10 digits
SEN	2	Sentence
PHO	1	International phone number
MOT	2	Mother tongue utterance

series consist of 7 experiments where 10 new speakers are introduced into the set while 10 old speakers are removed. By adding and removing different speakers to the 30 speaker test set, the best and worst cases of performance are tried to be observed.

VI.3 Training, Testing and Cross-Validation Data

The longer mother tongue free speech recording `mot02.alw` files from the first 4 sessions of each speaker are used to train the individual classifiers within the system. The testing data is composed of the shorter `mot01.alw` files which contain constrained mother tongue speech recordings. Each recording session is used as a single *token* of speech data for which individual classifiers generate outputs.

Apart from the training data from which the individual classifiers are trained, the theory necessitates a set of data on which the trained classifiers are operated during a set of *cross-validation experiments* to give classifier ensemble observation data. This information is then used to build a model of the classifier behavior in the form of *classifier observation statistics* as described in Chapter III.

In a typical speaker identification task, as is the case for the POLYCOST database, there is a limited amount of training data available for each speaker

model. Therefore, dividing the training data to obtain a cross validation data set is not feasible. However, the cross-validation data should be such that for each identification experiment, the unknown identity data given to the classifiers should never be seen by the classifiers during the training. Since during actual test sessions, all tokens are previously unseen, this is an important principle to be able to observe a realistic behavior of the classifiers.

One method of generating such a cross-validation data set is described in [23] and will be called as the *leave-one-out* method. In this scheme, the training data is used to generate a cross-validation data set as follows: There are 4 session data for each speaker model. Let each cross-validation token to be supplied to the system be an entire session data. Then for each such cross-validation test, a session data is left out of the training data for that speaker model. The speaker model is trained on the remaining sessions data. Then the data left out of model training is used as a cross-validation token and given to the system. The classifier outputs for this unseen token make *one* cross-validation test result. All training sessions for all speakers are processed by this scheme. Each time, the session left out of model training is used as a test token. Note that although the same training set is used, this scheme guarantees that the system is tested always with previously unseen tokens.

VI.4 Individual Classifiers

A speech pattern is often the sampled speech waveform of an utterance as described in Chapter V. For speaker identification, this is an entire recording from a single person. Each of the three classifiers described in Chapter V are used for the

experiments of the present chapter. They consist of individual feature extraction methods used in combination with a common modeling/similarity scoring method based on Vector Quantization(VQ) and the Binary-Splitting LBG VQ-codebook training algorithm. The three resulting classifiers are denoted by *FFTCep*, *LPC-Cep* and *LPCResCep* and have been illustrated in Figure V.8.

VI.5 Combination based on the POS Theory Formalism

In order to apply the POS Theory described in Chapter III to the present task, the following computational model is proposed. This is also illustrated in Figure VI.2.

1. First, an appropriate partitioning $\mathcal{W}$ is chosen for the classifier observation space. This determines the transition terms $P\{\underline{s}_x = j, \underline{r} = n | \underline{g}_{\mathcal{W}} = \mathcal{W}(j,n)\}$.

2. The labelled training data used to train the individual classifiers is transformed into a cross-validation data set by the leave-one-out method. Each pattern is left out of the training set during classifier training and used as a *test pattern* to test the trained classifiers. Hence, L cross-validation tests outputs are obtained using the classifiers (where L is the number of patterns in the database) with each test pattern being previously unseen by the classifiers.

3. The results of these cross-validation tests are used to determine the distribution of the source classes and rank scores among the partitions by means of *partition accumulators*. Hence, partition occurrence statistics $P\{\underline{g}_{\mathcal{W}} = m\}$

are estimated for $m = 1, 2, \cdots, M$. This is the statistical combination model.

4. Given an unknown test pattern, the classifiers are operated on the pattern and the rank score matrix $\mathbf{R}$ is obtained. The computational model of the optimum solution (Section III.3.2) is applied for the given $\mathbf{R}$ and the statistical coefficients $P\{\underline{s}_x = j, \underline{r} = n | \underline{g}_{\mathcal{W}} = \mathcal{W}(j,n)\} P\{\underline{g}_{\mathcal{W}} = \mathcal{W}(j,n)\}$ for all candidate classes $j = 1, 2, \cdots, P$ are computed. The candidate class with the maximum coefficient is selected by the rule

$$\underline{d} = \underset{j=1,2,\cdots,P}{\operatorname{argmax}} P\{\underline{s}_x = j, \underline{r} = n | \underline{g}_{\mathcal{W}} = \mathcal{W}(j,n)\} P\{\underline{g}_{\mathcal{W}} = \mathcal{W}(j,n)\}$$

Three methods are proposed under the POS Theory formalism. Each of these methods are based on a specific partitioning of the classifier observation space. These are described in the following sections.

VI.5.1 Method 1: First Rank

This method basically relies on the *first rank* based partitioning rule discussed in Chapter III with a modification for the consensus decision case. Firstly, only the highest ranking class label is considered from each classifier. This effectively discards the intermediate rank information from the lower ranking classes and degenerates into a Type 1 system. However, such a method may be justified when lower ranks are highly unreliable or cross-validation data is very limited for estimating partition statistics for a finer partitioning. Secondly, when the two classifiers are in consensus(i.e., their top ranking class labels are the same) it is reasonable to trust this consensus decision. In addition to considering the

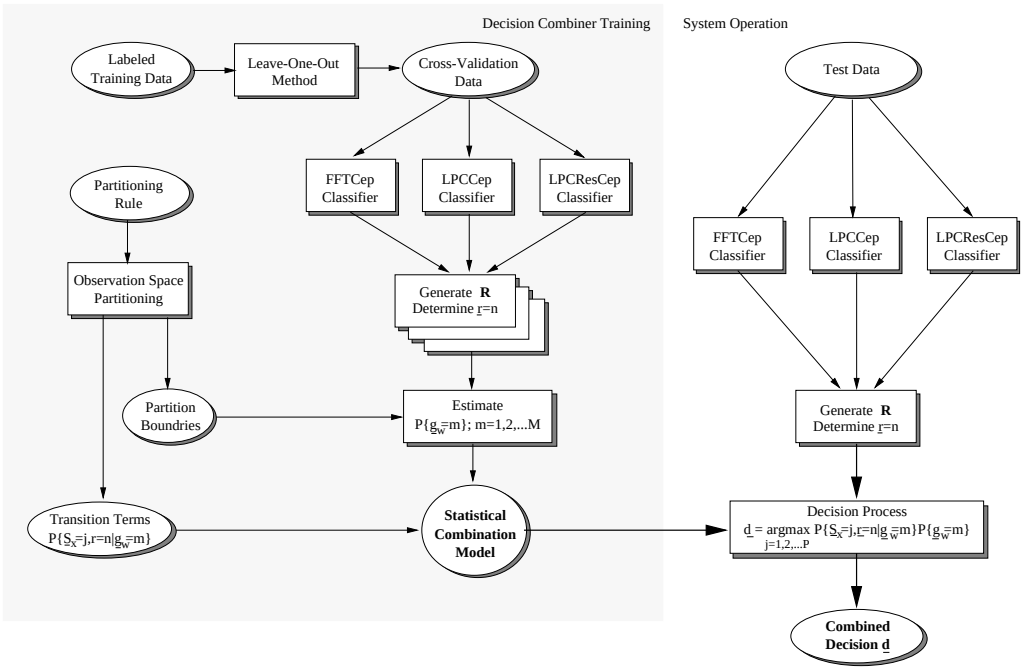

Figure VI.2: Training and operation block diagram of the Multiple Classifier Decision Combiner. Combiner system training consist of the determination of the observation space partitioning and the estimation of the resulting statistical combination parameters. The estimation is done by running the set of classifiers on a generated cross-validation data. During operation, the set of classifiers are run in parallel on unknown data and the statistical combination model is used to apply the optimal decision process.

top rank from each classifier, the partitioning is modified to accommodate this observation.

The partitioning described is independent of the actual data observed on a specific task but reflect a general understanding of the task. It constrains a possible solution within a certain region of the solution space by effectively achieving a *partitioning of the observation space* in the sense of the theory developed in Chapter III. A considerable compression of the classifier observation space is achieved. One can expect unacceptable performance if the actual behavior of the classifiers violate the assumptions.

VI.5.2 Method 2: First Two Ranks

This method relies on the *first two ranks* based partitioning described in Chapter III again with a modification for consensus. The highest ranking two class labels are considered from each classifier. Again, the consensus decision is accepted whenever there is one.

VI.5.3 Method 3: First Two Ranks with Variable Ordering

This method is a variation on Method 2, where the ordering of the class labels within the first two ranks is considered unimportant. This attempt is in response to the fact that possible rank score matrices increase with the rank depth and the limited cross-validation data may be insufficient to estimate observation The possible observation partitions are further reduced by suppressing the ordering detail within the first two ranks but keeping the resolution at a *membership* level instead.

VI.6 Experimental Results

In this section, the performance of the individual classifiers and their combination using existing and proposed techniques are presented. The 7 speaker sets are considered for these experiments.

VI.6.1 Performance of Individual Classifiers

Tables VI.2 and VI.3 illustrate the performance of the three individual classifiers on the speaker sets considered. As expected, there is a general performance drop from cross-validation to actual test since the test data is further separated in time

Table VI.2: Cross-validation test results for the three individual classifiers, FFTCep, LPCCep and LPCResCep. The identification experiments are performed for 7 different speaker sets, each composed of $L = 30$ speakers. The figures in the table are percent classification rates.

Classifier	FFTCep	LPCCep	LPCResCep
Set 1	88.4	90.2	82.7
2	88.3	89.2	73.3
3	91.7	93.3	82.5
4	93.3	91.7	80.8
5	92.5	94.2	82.5
6	93.3	95.0	80.8
7	93.3	95.0	78.3
Average	91.6	92.6	80.1

Table VI.3: Actual test results for the three individual classifiers FFTCep, LPCCep and LPCResCep.

Classifier	FFTCep	LPCCep	LPCResCep
Set 1	88.4	90.2	82.7
2	90.2	89.0	77.5
3	91.8	86.8	82.4
4	85.4	81.5	79.8
5	85.1	83.4	77.1
6	81.3	79.5	75.3
7	89.7	89.7	82.8
Average	87.4	85.7	79.6

from the data used for training and cross-validation. Therefore, a performance drop occurs due to the temporal variations in the individual's speech signal. Another interesting observation is that the best performing classifier is changed from cross-validation to actual test. The FFTCep classifier which is the second best classifier with an average identification rate of 91.6% during cross-validation becomes the best classifier with 87.4% identification rate during actual test. The highest performance achieved by this classifier during test is 89.7% which occurs for speaker set 7.

VI.6.2 Combined Performances using Existing Methods

The performances of the two-classifier rank-based decision combination systems using the three combination methods from the literature are illustrated in Tables VI.4, VI.5 and VI.6. The results of a statistical significance test, under the assumption of Gaussian distribution, are given as the last three rows of these Figures and will be discussed at the end of this Section.

The performances of the Highest Rank and the Borda Count methods are not acceptable for this task. The Borda Count method have rarely exhibited an improvement over the best classifier performance and the average over the speaker sets shows a consistent loss of performance for all combined classifier pairs. The Borda Count method performed slightly better but still well below acceptable. Occasional improvements over the best classifier are seen for some speaker sets but on the average, there is still a loss of performance. The Logistic Regression method performed better for the combination of classifier pairs FFT-Cep/LPCResCep and LPCCep/LPCResCep with a positive average improvement over the best classifier. However, the combination model estimation method of Logistic Regression described in Chapter II led to defective models for 3 speaker set while combining the classifier pair FFTCep/LPCCep. These cases are marked on the table with dashes. The logistic regression model combines the rank scores from individual classifiers by a linear regression model of the form $y = a + br_1 + cr_2$ where r_1 and r_2 are rank scores from the two classifiers and a, b and c are the regression parameters [2]. The logic behind the model assumes that the two parameters b and c be positive for all cases. The deficient models have one of these

Table VI.4: Test results for combination using the Highest Rank Method for pairwise combination of available classifiers. The experiments are performed for 7 different speaker sets, each composed of $L = 30$ speakers. The last three rows of the table illustrate the results of a statistical significance test for the data of columns representing the improvement over the best performing individual classifier. The 95% and 90% values indicate the desired confidence level on the truth of hypothesis H_1 and a yes/no value indicate whether or not the truth can be guaranteed with the specified confidence.

Combined Classifiers	FFTCep and LPCCep	Δ from Best	FFTCep and LPCResCep	Δ from Best	LPCCep and LPCResCep	Δ from Best
Set 1	89.6	-0.57	85.0	-3.47	85.0	-5.20
2	90.7	0.58	85.5	-4.62	86.1	-2.89
3	87.9	-3.85	84.6	-7.14	83.0	-3.84
4	83.7	-1.68	84.8	-0.56	83.7	2.25
5	83.4	-1.71	80.0	-5.14	78.9	-4.57
6	81.3	0.00	75.3	-6.03	77.1	-2.41
7	89.7	0.00	83.3	-6.33	85.6	-4.03
Average	86.6	-1.03	82.7	-4.76	82.8	-2.96
$P\{data\|H_1\}$	-	0.06	-	0.00	-	0.01
H_1? (95%)	-	no	-	no	-	no
H_1? (90%)	-	no	-	no	-	no

parameters as negative upon performing an approximation to the cross-validation data hence deteriorating the performance. The scores for these cases are left out of the averaging process. However, the average performance for this classifier pair is still poor.

To establish the significance of the given results, a statistical significance test is performed. Assume that the actual distribution of the performance difference figures is Gaussian. Let the averages of the "Δ from Best" columns be indicators of whether or not there is an improvement over the best individual classifier. An hypothesis test can be performed to determine this. Define the default or *Null Hypothesis* (H_0) as "The improvement mean is zero." while the *Alternative Hy-*

Table VI.5: Test results for combination using the Borda Count Method for pairwise combination of classifiers.

Combined Classifiers	FFTCep and LPCCep	Δ from Best	FFTCep and LPCResCep	Δ from Best	LPCCep and LPCResCep	Δ from Best
Set 1	89.6	-0.57	90.2	1.73	89.6	-0.57
2	90.2	0.00	86.1	-4.04	87.3	-1.74
3	87.9	-3.85	86.3	-5.50	85.2	-1.65
4	83.2	-2.22	84.8	-0.56	81.5	0.00
5	82.9	-2.28	86.3	1.15	83.4	0.00
6	78.9	-2.41	81.3	0.00	81.7	2.21
7	87.9	-1.73	89.7	0.00	90.8	1.14
Average	85.8	-1.87	86.4	-1.03	85.6	-0.09
$P\{data\|H_1\}$	-	0.00	-	0.18	-	0.44
H_1? (95%)	-	no	-	no	-	no
H_1? (90%)	-	no	-	no	-	no

Table VI.6: Test results for combination using the Logistic Regression Method for pairwise combination of classifiers.

Combined Classifiers	FFTCep and LPCCep	Δ from Best	FFTCep and LPCResCep	Δ from Best	LPCCep and LPCResCep	Δ from Best
Set 1	88.4	-1.73	88.4	0.00	90.7	0.58
2	-	-	90.2	0.00	89.0	0.00
3	86.8	-4.95	90.7	-1.10	86.8	0.00
4	82.6	-2.81	86.0	0.57	83.1	1.69
5	-	-	86.7	2.29	82.3	-1.14
6	79.5	-1.81	83.1	1.80	79.5	0.00
7	-	-	90.8	1.14	89.7	0.00
Average	84.3	-2.83	88.0	0.57	85.9	0.16
$P\{data\|H_1\}$	-	0.02	-	0.90	-	0.68
H_1? (95%)	-	no	-	no	-	no
H_1? (90%)	-	no	-	no	-	no

Table VI.7: Test results for combination using Method 1: First Rank.

Combined Classifiers	FFTCep and LPCCep	Δ from Best	FFTCep and LPCResCep	Δ from Best	LPCCep and LPCResCep	Δ from Best
Set 1	90.7	0.58	89.0	0.58	91.3	1.16
2	90.7	0.58	89.6	-0.57	88.4	-0.58
3	92.3	0.55	91.8	0.00	86.8	0.00
4	85.4	0.00	86.5	1.13	81.5	0.00
5	85.7	0.57	85.1	0.00	82.7	-0.77
6	81.3	0.00	80.1	-1.21	79.5	0.00
7	91.9	2.29	89.7	0.00	90.8	1.14
Average	88.3	0.65	87.4	-0.01	85.9	0.14
$P\{data\|H_1\}$	-	0.97	-	0.49	-	0.86
H_1? (95%)	-	yes	-	no	-	no
H_1? (90%)	-	yes	-	no	-	no

pothesis (H_1) as "The improvement mean is positive," or equivalently "There is an improvement." Denote the probability of observing the column data, given that the Alternative Hypothesis is true as $P\{data|H_1\}$. The value of $P\{data|H_1\}$ as well as the *truth* of the Alternative Hypothesis (hypothesis test result) for two different reference confidence levels are given for all three tables. From these results, it can be concluded that for all three existing methods, one cannot establish, even with 90% confidence, that there will be an an actual improvement in performance.

VI.6.3 Combined Performances using Proposed Methods

The performances of the two-classifier rank-based decision combination systems using the three new combination methods are illustrated in Tables VI.7, VI.8 and VI.9.

Method 1 based on the First Ranks only achieves good performance for the

Table VI.8: Test results for combination using Method 2: First Two Ranks.

Combined Classifiers	FFTCep and LPCCep	Δ from Best	FFTCep and LPCResCep	Δ from Best	LPCCep and LPCResCep	Δ from Best
Set 1	90.7	0.58	89.0	0.58	91.3	1.16
2	90.2	0.00	90.2	0.00	89.0	0.00
3	91.8	0.00	91.8	0.00	86.8	0.00
4	85.4	0.00	85.4	0.00	82.0	0.56
5	85.7	0.57	85.1	0.00	83.4	0.00
6	81.3	0.00	81.3	0.00	79.5	0.00
7	92.0	2.29	89.7	0.00	89.7	0.00
Average	88.2	0.49	87.5	0.08	86.0	0.25
$P\{data\|H_1\}$	-	0.91	-	0.82	-	0.75
H_1? (95%)	-	no	-	no	-	no
H_1? (90%)	-	yes	-	no	-	no

Table VI.9: Test results for combination using Method 3: First Two Ranks with Variable Ordering.

Combined Classifiers	FFTCep and LPCCep	Δ from Best	FFTCep and LPCResCep	Δ from Best	LPCCep and LPCResCep	Δ from Best
Set 1	91.3	1.16	91.9	3.47	91.3	1.16
2	90.8	0.58	89.0	-1.15	86.7	-2.31
3	91.8	0.00	91.8	0.00	86.8	0.00
4	85.4	0.00	85.4	0.00	82.0	0.56
5	85.7	0.57	85.1	0.00	83.4	0.00
6	81.3	0.00	80.7	-0.61	79.5	0.00
7	92.5	2.87	92.5	2.87	90.2	0.57
Average	88.4	0.74	88.1	0.65	85.7	0.00
$P\{data\|H_1\}$	-	0.95	-	0.82	-	0.50
H_1? (95%)	-	no	-	no	-	no
H_1? (90%)	-	yes	-	no	-	no

combination of the classifier pair FFTCep/LPCCep with a maximum improvement of 2.29% over the best individual classifier. The average improvement is 0.65% for all speaker sets. This is the only case where an improvement can be guaranteed with a 95% confidence. The performance improvement remains marginal for the combination of the two other classifier pairs.

Considering the first two ranks by Method 2 introduces a slight improvement for the combination of the latter two classifier pairs. However, this is not significant. Although the maximum performance improvement for the FFTCep/LPCCep pair is still 2.29% over the best individual classifier, the average improvement drops to 0.49%. Also, one's confidence that there is an actual improvement is dropped and can no longer meet a 95% confidence level. Improvement can not be guaranteed for the other two classifier pairs even with a 90% confidence level.

The major factor behind this performance loss by considering first two ranks can be given as follows. The speaker identification is inherently a task with sparse cross-validation data. The amount of speech data required for a classifier to reach a speaker identity decision is large. A speech segment of acceptable length for a classifier is called a speech token. As the duration of the token becomes shorter, the number of cross-validation test samples becomes larger but the classification performance of individual classifiers deteriorate. Therefore one is not able to observe the actual behavior of the classifier ensemble. When the duration is kept longer for acceptable individual performance, the total number of cross-validation tests that can be performed gets limited and this critically effects the reliability of the classifier observation statistics. For the present experiments, the token

length is selected as the recording session length. With the available database, this leads to 4 sessions per speaker, i.e., 4 cross-validation tests per speaker. The cross-validation data is clearly very sparse.

By considering the first two ranks partitioning instead of the first rank partitioning considerably increases the problem dimensionality and the number of classifier observation statistics to be estimated. Therefore the estimations of these statistics becomes poor. Therefore, the expected gain from Method 2 using the first two ranks partitioning cannot be achieved for this task.

The performance figures for Method 3, which uses a modified version of the first two ranks partitioning are given in Table VI.9. This modification specifically addresses the limited cross-validation data and reduces the number of partitions by discarding the ordering within the first two ranks and preserving only the membership information. The results obtained for this method show a maximum improvement of 2.87% over the best individual classifier for the FFTCep/LPCCep and FFTCep/LPCResCep classifier pairs. The average improvement for all the speaker sets for these two classifier pairs are increased to 0.74% and 0.65% respectively. However, the significance test results show that although the FFTCep/LPCCep classifier pair confidence for an improvement is increased, one still cannot guarantee an improvement with 95% confidence.

Although the cross-validation data is sparse, it is interesting to note that methods based on classifier observation statistics perform better than the simple methods Highest Rank and Borda Count which does not use any classifier observation statistics. This shows that the inherent assumptions by these two methods cannot be justified for the speaker identification task. It is also noted that the

smoothing model imposed by the Logistic Regression method is not suitable for this case. The only case one is 95% confident that there is an improvement over the best individual classifier is classifier pair FFTCep/LPCCep using Method 1.

VI.7 Discussion

The experimental results in this chapter have shown that simple application of the POS Theory is capable of generating useful methods for pattern classification. These simple methods making use of classifier observation statistics are shown to outperform three existing rank-based decision combination methods. However, the performance gain achieved over the best performing individual classifier is often not very significant. A 95% confidence level hypothesis test can guarantee improved performance for only a single case, namely the FFTCep/LPCCep classifier pair using Method 1. All statistical rank-based methods analyzed and unified by the present thesis necessitate reliable estimation of classifier observation statistics from cross-validation test data. Speaker identification task on the other hand is inherently sparse in this respect. Note that this is not a problem associated with the database used for the experiments but a general characteristic of the task, since the database reflects realistic operating conditions. Therefore, it can be argued that speaker identification does not seem to be a task which is best suited for the application of statistical rank-based methods. The next Chapter discusses an interesting task yet having a constrained size.

CHAPTER VII

Experiments on Turkish BDEV Discrimination Task

In this chapter, the application of the theory presented in Chapter III in another real-life problem from speech pattern classification is considered. This is the discrimination of the highly confusable four letter-words: The Turkish names of the letters b,d,e and v.

VII.1 Task Description and Database

The task of discriminating the BDEV letter-words for English has been first discussed by Lang to introduce the time-delayed neural networks [6]. The discrimination of the Turkish version letter-words is a similar task with the same difficulties. It is an interesting and difficult real-life task having constrained size (only 4 pattern classes) and hence much more cross-validation data per class can be available than the task of speaker identification. This enables more reliable

estimates for the classifier observation statistics for the decision combiner. Also, the manageable size of the problem enables the visualization and examination of the classifier observation statistics within the context of the developed theory.

The aim of this task is to classify an unknown speech signal to be an utterance of one of the 4 letter-words with the highest rate of correctness. This is a closed-set classification task similar to closed-set speaker identification discussed in Chapter VI. Therefore, the individual classifier operation for this task is similar to the one illustrated in Figure VI.1.

A multi-speaker database of 5 speakers is collected in common office environment for the experiments. One training and one testing session are recorded with a 2 to 3 days time separation between them. The speakers are asked to read a random sequence of isolated letter-words with approximately 60 utterances/letter-word, building up a total of 1200 training and 1200 test utterances. The recordings are made in *telephone quality* with 8 kHz sampling rate and 8 bits/sample quantization in a common office environment with computer cooling fans as the dominant background noise. The boundaries between the utterance of each letter-word is then determined by hand. However, the no speech/silence detection is performed and hence, silence sections of the utterances are not eliminated. Actual recordings from this database are illustrated in Figure VII.1 with their corresponding spectrograms.

The main difficulty of the task is that the discriminating consonant sounds at the beginning of the letter-words are of very short duration and are followed by high energy vowel 'e' sounds which are common to all 4 classes. Therefore the high energy parts of the signal are non-discriminative. Also, the SNR is low due

to the slow sampling rate and coarse quantization. The presence of the silence sections within the utterances increases the difficulty of the task.

VII.2 Generation of the Cross-Validation Data

The cross-validation data is again generated from the data collected for the training of the individual classifiers by the *leave-one-out* method detailed in Section VI.3. Namely, for all classifiers, each letter-word utterance is left out of the training of the classifier and the classifier is trained by all the remaining utterances. Then the left-out utterance is used as a test token for the system. This is repeated for all training tokens to generate the cross-validation test results.

VII.3 Individual Classifiers

A speech pattern is often the sampled speech waveform of an utterance as described in Chapter V. For the BDEV discrimination task, such a pattern is an utterance of a specific letter-word.

Each of the three classifiers described in Chapter V are used for the experiments of the present chapter. They consist of three individual feature extraction algorithms used in combination with a common modeling/similarity scoring procedure based on Vector Quantization(VQ) and the Binary-Splitting LBG VQ-codebook training algorithm. The three resulting classifiers are denoted by *FFTCep*, *LPCCep* and *LPCResCep*.

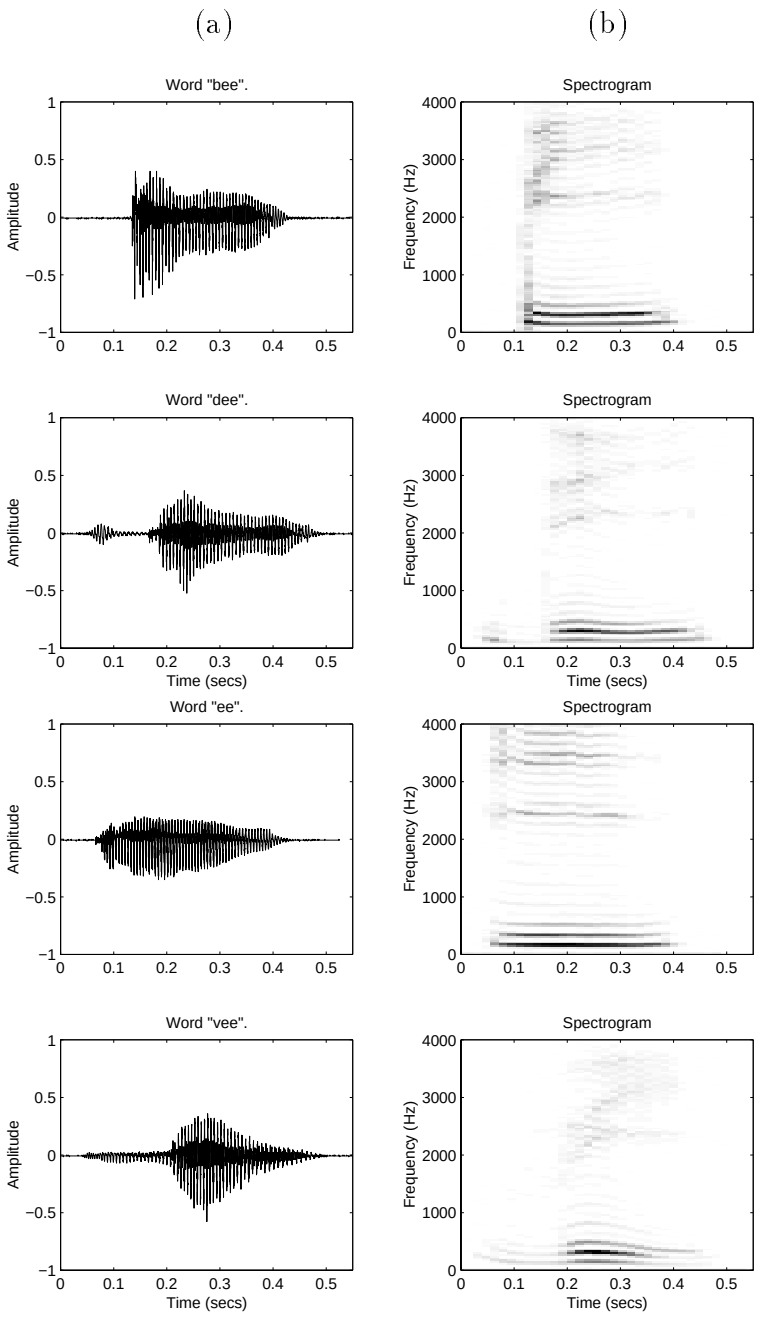

Figure VII.1: Examples for the 4 letter-word classes. (a) The time waveform of the letter-word. (b) The corresponding spectrogram of the letter-word. The spectrogram illustrate the time-frequency behavior of the signal. One can note that the dominant high energy formant frequencies corresponding to the vowel sound (near black lines close to the bottom) are very similar for each letter-word while the differences arise mainly in the low energy (near white) parts.

VII.4 Combination Based on the POS Theory Formalism

Two simple rank-based decision combination strategies are proposed based on the POS Theory formalism. The first one is based on the *First Two Ranks Partitioning* described in Section III.3.3. Here the classifier behavior observation space is partitioned to discriminate only among the top two positions of the candidate class ranking from each classifier. We call this method as *Rank2*. This intuitive partitioning is based on an expectation that the lower ranks in a classifier output may be unreliable due to noisy estimates and is done in a data-independent manner.

The second method is based on *First Rank Based Partitioning* described in Section III.5. It is a degenerate version of a rank-based system by a partitioning which discriminates only the top position of the candidate class rankings. Therefore, it can be called as a Type 1 system. We call this method as *Rank1*.

These two methods are used by the computational model previously described in Section VI.5 and illustrated in Figure VI.2.

VII.5 Experimental Results

VII.5.1 Performance of Individual Classifiers

The classification performances of the three individual classifiers are given in Table VII.1 both for the cross-validation tests and for the actual tests. The best performing classifier for both cross-validation and test data is the FFTCep classifier. One can also observe that the LPCResCep classifier individual performance is very low for this task.

Table VII.1: Classification performances of individual classifiers on the BDEV task.

Classifier	FFTCep	LPCCep	LPCResCep
Cross-Validation Data	62.2	63.8	37.7
Test Data	65.8	61.8	35.0

VII.5.2 Combined Performances

The comparative results of pairwise and collective combination of these individual classifiers by all the rank-based decision combination methods discussed are given in Table VII.2 and Table VII.3.

A serious problem for the Highest Rank method was the excessive score collisions (more than one class with the same max-score) when small number of classes are involved. Highest Rank and Borda Count methods are aided by resolving score collisions with the decision of the best performing classifier instead of a random decision between colliding classifiers.

The last three combination methods show classification improvement over the individual classifiers. The highest performance improvement is achieved by the Rank2 method for all pairwise combinations. The most consistent improvement by rank-based methods seems to be achieved for the classifier pair FFTCep/LPCCep. Also there is a significant performance improvement for the combination of all three classifiers by Rank1 and Rank2 methods. However, the performance when all three classifiers are combined by Rank2 method is a drop over the improvement by Rank1, contrasting with the expectation for higher improvement. For this case, the Rank1 method achieves the highest improvement.

Table VII.2: Classification performance of existing and proposed combination methods on the BDEV task: Pairwise combination of classifiers.

Combined Classifiers	FFTCep LPCCep	Δ from Best	FFTCep LPCResCep	Δ from Best	LPCCep LPCResCep	Δ from Best
Highest Rank	65.8	0.00	65.8	0.00	61.8	0.00
Borda Count	67.6	1.75	58.2	-7.66	55.6	-6.25
Logistic Regression	67.5	1.67	65.8	0.00	61.8	0.00
Rank 1	68.8	3.00	65.8	0.00	61.8	0.00
Rank 2	69.2	3.39	66.8	0.92	62.3	0.45

Table VII.3: Classification performance of existing and proposed combination methods on the BDEV task: All three classifiers combined.

Combined Classifiers	FFTCep, LPCCep and LPCResCep	Δ from Best
Highest Rank	35.0	-30.8
Borda Count	64.7	-1.16
Logistic Regression	66.3	0.42
Rank 1	69.2	3.37
Rank 2	68.8	3.00

It can be argued that the reason for this drop is the increasing number of statistics to estimate for the Rank2 partitioning with three classifiers. This increase in problem dimensionality degrades the reliability of the classifier observation statistics estimated from the fixed amount of cross-validation data and hence may degrade the combination performance.

For the methods *Logistic Regression*, *Rank 1* and *Rank 2*, the classifier observation statistics are extracted from the cross-validation tests. The maximum gain from these methods can be achieved when these statistics are reliably estimated and exactly reflect the behavior on the actual test data. To observe such an

Table VII.4: Classification performance of rank-based statistical combination methods on the BDEV task based on statistics derived from the test data instead of the cross-validation data: Pairwise combination of the three classifiers. These performance figures show the upper bounds in performance possible for an exact statistical match between cross-validation and test.

Combined Classifiers	FFTCep LPCCep	Δ from Best	FFTCep LPCResCep	Δ from Best	LPCCep LPCResCep	Δ from Best
Logistic Regression	67.5	1.67	65.8	0.00	61.8	0.00
Rank 1	69.0	3.17	65.8	0.00	61.8	0.00
Rank 2	73.9	8.09	70.3	4.42	66.9	5.09

Table VII.5: Classification performance of rank-based statistical combination methods on the BDEV task based on statistics derived from the test data instead of the generated cross-validation data: All three classifiers combined.

Combined Classifiers	FFTCep, LPCCep and LPCResCep	Δ from Best
Logistic Regression	66.8	0.92
Rank 1	70.9	5.09
Rank 2	85.8	20.0

upper bound on performance, the behavior statistics are also extracted from the actual test data and the combination results are given in Tables VII.4 and VII.5. It can be observed that the potential improvement in combined results is much larger than the actual improvement. This suggests that there is a mismatch in the classifier behavior from cross-validation to testing and this plays a significant role in limiting the improvement achievable by combination.

This is an expected behavior but shows that there is still a margin for improvement if the behavior statistics can be more reliably estimated. The distribution of the cross-validation test and actual test result samples in the original event

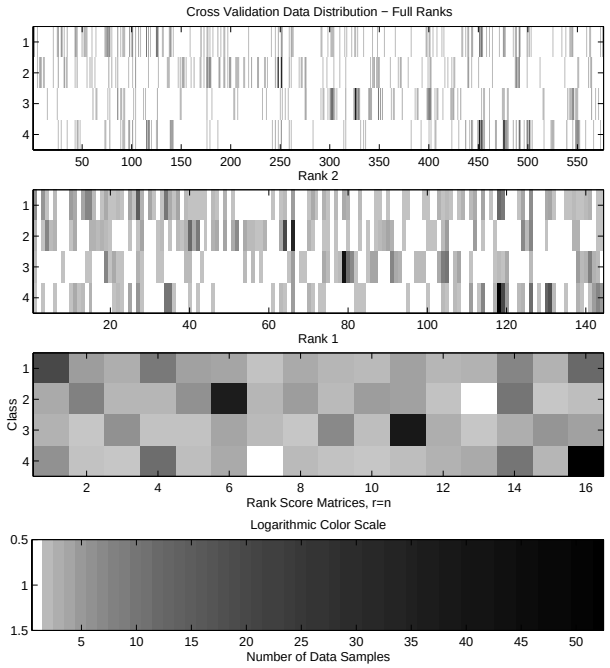

Figure VII.2: The distribution of the cross-validation test patterns among the observation space partitions for 2 classifiers case: No partitioning, Rank1 and Rank2. The nonlinear gray color scale is illustrated at the bottom of the Figure.

space and in the partitions resulting from Rank1 and Rank2 methods is illustrated in Figures VII.2 and VII.3. Color *white* signifies the lack of any samples to estimate a statistic while *black* signifies maximum amount of samples. The distribution of the cross-validation samples in Figure VII.2 for the no-partitioning (full-ranks) case clearly illustrate how such data is sparsely distributed across the uncompressed classifier observation space. From the two figures, the mismatch in the behavior statistics from cross-validation to test can also be clearly observed. Note that the mismatch is more apparent when there is no partitioning at all. With the introduction of the Rank1 and Rank2 partitionings, the available data becomes sufficient for the estimation of the resulting observation statistics. Also, the mismatch from cross-validation to test is partially smoothed-out.

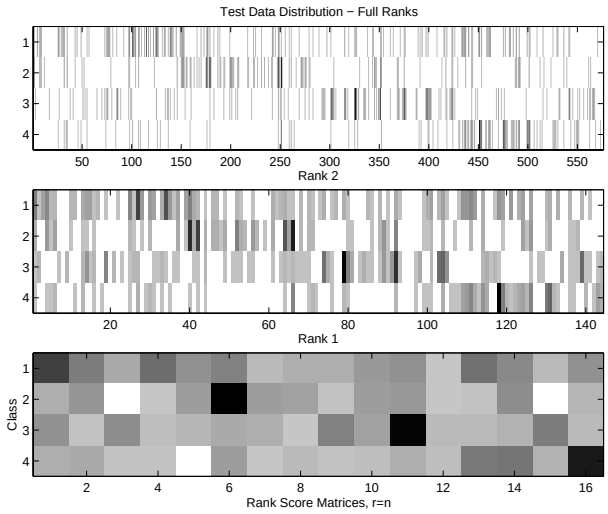

Figure VII.3: The distribution of the actual test patterns among the observation space partitions for 2 classifiers case: No partitioning, Rank1 and Rank2.

Table VII.6: Classifier independence and complementariness measures for the pairwise combination of the three classifiers using Rank1 partitioning. Column $\Delta I_{X_b X_w}$ denotes the extend to which the worse classifier X_w complements the best classifier X_b. Column $I(\underline{r}_1, \underline{r}_2)$ denotes the output independence of the classifier pair.

Classifier Pair	$\Delta I_{X_b X_w}$	$I(\underline{r}_1, \underline{r}_2)$
FFTCep/LPCCep	0.18	0.16
FFTCep/LPCResCep	0.025	0.023
LPCCep/LPCResCep	0.032	0.033

VII.5.3 Evaluation of Independence and Complementariness

For the pairwise combination of the three classifiers by the Rank1 and Rank2 methods, the measures discussed in Chapter IV are applied. The cross-validation statistics are used to compute the output independence between the two classifiers subject to combination and the complementariness of the worse individual classifier with respect to the best classifier.

The results for the Rank1 method are illustrated in Table VII.6. From this

table, it can be observed that the best performing classifier pair FFTCep/LPCCep have shown the largest complementariness while considerably lower values are obtained for the other two classifier pairs. Another interesting observation is that the classifier pair which is the most complementary is at the same the pair which shows the most output dependence. This supports our conclusion in Chapter IV that output independence does not reflect classifier complementariness.

Unfortunately, numerical difficulties are encountered while computing the same measures for the Rank2 method case. These difficulties mainly arise from the fact that the measures make use of *all* joint probabilities for the partitioned classifier observation space. It can be observed from Figure VII.2 that for the Rank2 partitioning, there is still a considerable number of cells of the classifier observation space with no data to estimate the corresponding partition probabilities. These probabilities which are estimated as zero valued in fact correspond to very small probabilities corresponding to rare events in the observation space. Unfortunately, such zero values cause numerical difficulties with the computation of the proposed measures. Note that such zero probabilities are rare in the case of Rank1 partitioning.

An interesting question which arise is, why this measure becomes unstable while the optimal decision still achieves good performance. For the optimal decision to achieve a reliable decision, one must have the following. For each specific rank score matrix (i.e., a column of the observation space), if there is a large difference between the largest and the second largest partition probability, then reliable estimation of the largest probability is sufficient for a reliable decision for that rank score matrix. If there is a small difference, then it is sufficient that *both*

of these two largest probabilities are reliably estimated. For reliable decision, it is not necessary that all probabilities of the column are reliably estimated. Investigation of Figure VII.2 and VII.3 shows that despite the empty partitions, this condition is mostly satisfied for Rank2 partitioning. However, the proposed complementariness measure suffers from this lack of data for certain partitions more severely than the optimal decision process since it uses *all* partition probabilities.

These experimental results suggest that the proposed complementariness measure is open to improvement. Future research may involve developing this measure so that it further takes into account the behavior of the optimal decision process, also making it at least as robust as the optimum decision process, against estimation errors.

VII.6 A Statistical Significance Test for Improvement

Promising improvements are suggested by Tables VII.2 and VII.3 for this task where there is considerable cross-validation samples to estimate the classifier observation statistics for reasonable partitionings of the observation space. However, based on the observed performance, it is important to gain an idea about how much one is confident of improvement by means of combination. In this section such an analysis is presented.

Assume that N independent classification tests are performed to obtain the performance figures about the individual classifiers and their combinations. Let $\underline{x}_i,\ i = 1, 2, \cdots, N$ be the *indicator* of correct classification by the best individual classifier for the i'th test with

$$\underline{x}_i = \begin{cases} 1, & \text{if the } i\text{'th classification is correct,} \\ 0, & \text{otherwise.} \end{cases} \tag{VII.1}$$

Also assume that $P\{\underline{x}_i = 1\} = p$ is the *true* probability of correct classification by the best individual classifier. The sum of all N indicator random variables, given by $\underline{x} = \sum_{i=1}^{N} \underline{x}_i$, is another random variable which takes all integer values between 0 and N. This new random variable has Binomial distribution with

$$P\{\underline{x} = n\} = \binom{N}{n} p^n (1-p)^{N-n}. \tag{VII.2}$$

for $n = 0, 1, \cdots, N$. The expected value of $\underline{x}$ is given by

$$E\{\underline{x}\} = N \cdot p. \tag{VII.3}$$

Therefore, $\frac{1}{N} E\{\underline{x}\}$ is an unbiased estimate of the true probability p of correct classification [75].

Now consider the multiple classifier system decision. Let the indicator of correct combined classification for N test tokens as $\underline{y}_i$, $i = 1, 2, \cdots, N$, defined as in Eq. (VII.1). Once more, $\underline{y}$ is defined as $\underline{y} = \sum_{i=1}^{N} \underline{y}_i$ having binomial distribution. The *true* probability of correct classification for the combined system is assumed as $P\{\underline{y}_i = 1\} = q$. One has $\frac{1}{N} E\{\underline{y}\}$ as an unbiased estimate of this true probability q.

Now one can try to answer the question of how much can one be confident that there is truly an improvement by combination when the individual and combined performances are observed over N tests. This can be answered by computing

Table VII.7: Statistical confidence levels (percent probabilities) to obtain an actual improvement over the best individual classifier in the combination.

Combined Classifiers	FFTCep LPCCep	FFTCep LPCResCep	LPCCep LPCResCep	FFTCep LPCCep LPCResCep
Highest Rank	49.1	49.1	49.1	0.0
Borda Count	79.7	00.0	0.8	29.5
Logistic Regression	78.3	49.1	49.1	58.4
Rank 1	92.3	49.1	49.1	94.5
Rank 2	94.5	67.4	58.2	92.0

the confidence of the Hypothesis: $\{\underline{y} > \underline{x}\}$. This is equivalent to finding the probability $P\{\underline{y} > \underline{x}\}$. This probability can be computed as

$$P\{\underline{y} > \underline{x}\} = \sum_{i=1}^{N} P\{\underline{y} = i, \underline{x} < i\}. \tag{VII.4}$$

Now, assuming statistical independence between the individual classifier tests and the combined system tests, Eq. (VII.4) becomes

$$\begin{aligned} P\{\underline{y} > \underline{x}\} &= \textstyle\sum_{i=1}^{N} p\{\underline{y} = i\} P\{\underline{x} < i\} \\ &= \textstyle\sum_{i=1}^{N} P\{\underline{y} = i\} \sum_{j=0}^{i-1} P\{\underline{x} = j\}, \end{aligned} \tag{VII.5}$$

where $P\{\underline{x} = j\}$ and $P\{\underline{y} = i\}$ are given by the two Binomial distributions $B(n, N, p)$ and $B(n, N, q)$ respectively.

When this computation is applied for the individual classifier performance figures given in Table VII.1 and the combined performance figures in Tables VII.2 and VII.3, the improvement confidence values in Table VII.7 are obtained.

The confidence values in Table VII.7 show that an improvement over the best performing individual classifier have a strong likelihood for several cases. For example, for the combination of the classifier pair FFTCep/LPCCep by the

Rank2 method, one can be 94.5% confident that there is an actual improvement over the best individual performance of 65.8% correct classification.

VII.7 Discussion

In this chapter, the Turkish BDEV letter-word classification task has been considered. It has been shown that statistically significant improvements over the best performing classifier can be obtained for this task by the application of the theory developed in the present thesis. The BDEV task is such that there are limited number of pattern classes and hence the problem dimensionality is small as compared with the speaker identification task considered in Chapter VI. Due to the comparatively small problem dimensionality and the use of a partitioning, the cross-validation data becomes sufficient for the estimation of the classifier observation statistics. As a result, a reliable rank-based statistical combination model can be obtained for the Logistic Regression, Rank1 and Rank2 methods. However, it can be said that the Logistic Regression method cannot perform as well as the Rank1 and Rank2 methods due to the over-smoothing done by an hyper-plane fit to the data, supporting our discussion in Section III.4.3. The computation of the independence and complementariness measures developed in Chapter IV for Rank1 partitioning led to meaningful results in support of the discussions of Chapter IV. However, for Rank2 partitioning, numerical difficulties revealed their sensitivity to estimation errors in the observation space.

The results of Chapter VI and the present chapter establishes the fact that the availability of enough cross-validation data to estimate the classifier observation statistics is very important for the performance improvement by statistical

rank-based decision combination methods. These include the family of optimal combination methods which are the focus of the present thesis. The present task exemplifies pattern recognition tasks with comparatively few pattern classes and relatively large training and testing data per class. One can conclude that pattern recognition tasks exemplified by the present chapter are more suited for rank-based combination methods.

CHAPTER VIII

Conclusions

VIII.1 Summary

A statistical unified theory to analyze and extend rank-based multiple classifier systems has been proposed. The theory has been used to analyze existing rank-based decision combination methods with respect to the global optimum solution. Then the theory has been extended to investigate independence and complementariness among classifiers and their effect in combined performance.

The decision combination problem in rank-based multiple classifier systems can be formulated as a discrete optimization problem which possess a simple global optimum solution. The formulation used a set of statistics about the joint output behavior of the involved classifiers. However, the prohibitive dimensionality of the resulting problem space and the need to integrate prior knowledge and assumptions into the problem necessitated manipulation and compression of this problem space. A partitioning approach has been proposed and elaborated

in this thesis to achieve this end. Under this framework, it has been shown that a number of different partitionings lead to three existing methods, namely Highest Rank, Borda Count and Logistic Regression. The optimality of these methods have been analyzed under the developed framework and their non-optimality have been established. Another important result of this part is that once a partitioning rule is chosen and the associated observation statistics are estimated, the optimal solution can be implemented by an efficient computational model. This model allowed the solution to be computed on-the-fly as the unknown patterns are processed for classification.

Despite the clear distinction between Type 2 (rank-based) and Type 1 systems proposed in [9], the thesis have shown that the theory developed for rank-based systems can be applied to analyze Type 1 systems. In fact, Type 1 systems do result from a simple first rank based partitioning.

It is a common practice to make the assumption of statistical independence between the outputs of classifiers in multiple classifier systems, to facilitate the analysis. However, the effects of this assumption on the combined performance has not been well understood. The actual improvement achievable by combination is another issue which can be called as the complementariness between classifiers. The thesis attempted to use the developed theory in combination with related concepts from Information Theory to present an Information Theoretic interpretation of rank-based multiple classifier systems. Afterwards the interpretation has been used to develop measures for output independence and complementariness. It has been established that output independence plays no direct role in determining the potential improvement by combination. However,

the thesis has also shown that may not benefit from the available potential for improvement, if the analysis is made under the independence assumption. Whether there would be an actual improvement by combination has been shown to be linked to a condition termed as dominance of a classifier. An Information Theoretic measure is proposed for complementariness and justified by several simulated experimental results.

Finally the theory developed by the thesis is applied to two pattern recognition tasks from speech processing. These two tasks had different characteristics and exemplified different classes of pattern recognition problems. The behavior of existing rank-based multiple classifier systems are investigated on these two tasks. Although the proposal of new methods have not been a primary goal of the present thesis, simple application of the developed theory proved to perform better than the existing methods on these two tasks.

The speaker identification task had large number of pattern classes (speakers) and therefore a small amount of training and testing data per class. This led to limited cross-validation data which was required for the estimation of the observation space statistics for combination using the proposed theory. Sparse cross-validation data limited the usefulness of the statistical rank-based methods including those proposed under the proposed theory and led to marginal improvements by combination. This resulted from unreliable estimates for the classifier observation statistics.

One the other hand, problems having relatively few pattern classes, exemplified by the BDEV letter-word recognition task proved to be more suitable for the application of statistical rank-based systems. Application of the developed the-

ory in this task provided statistically significant improvements by combination, over the performance of the best individual classifier. For this case, comparatively large amount of cross-validation data could be generated and the classifier observation statistics could be reliably estimated for the partitionings considered.

VIII.2 Directions for Future Research

The theoretical framework proposed in this thesis relies on the partitioning of the joint classifier observation space as controlled tool of proposing different combination methods for rank-based decision combination. Several partitioning rules are discussed, specifically to analyze existing methods. Also, a number of them are used for the application tasks. However, it is yet not clear how to generate such useful partitionings in an automatic, preferably optimal manner.

It is clear that the availability of the cross-validation data and its distribution among the partitions resulting from a specific partitioning rule is crucially important to provide reliable estimates for the resulting classifier observation statistics. Therefore, a partitioning should ideally take into account the distribution of the cross-validation data over the original unpartitioned observation space. However, such a partitioning should also be able to incorporate one's assumptions and prior knowledge about the task at hand. An interesting research direction is to investigate the use of optimization methods such as Genetic Algorithms to achieve an automatic partitioning. For example, a possible optimization criterion for automatic partitioning can be based on minimizing the variability between observation statistics estimates based on different cross-validation samplings of the training data or on different sub-divisions of the cross-validation data.

Another research direction related with automatic partitioning idea is the investigation of the optimal smoothing of the classifier observation statistics with the aim of improving their reliability. The relation between optimally smoothing the statistics for a fixed partitioning and optimally generating a partitioning is very interesting to establish. These methods will be very useful to automatically generate new rank-based combination methods suitable for a certain task, once the training and testing results are obtained.

The independence and complementariness treatment in this thesis raises many interesting questions as well as answering some others. The dominance condition is developed as an indicator of improvement by combination. However, its relations with the Information Theoretic independence and complementariness measures are yet to be established. These measures, which are based on the joint output behavior of the set of classifiers, can take into account the partitioning of the observation space. However, as is the case for the complementariness measure, the effect of the optimal decision method are not involved. Therefore, improvement potential reported by the proposed complementariness measure may not be realized. Yet another issue open for analysis is the concept of error independence and its relations with complementariness and consist an interesting and promising direction for future research.

A last point arises when considering the cases where there still exist empty partitions for which partition statistics cannot be reliably estimated. It is observed that even for such cases, the jointly optimal decision process can achieve significant performance improvements. However, the proposed complementariness measure may fail to report improvements since it uses all partition statistics and

may become numerically unstable due to zero partition statistics. This suggest two interesting directions for future research. One is to investigate the optimality of the decision under unreliable partition statistic estimates, which may lead to the design of some hierarchical classifier systems. The other one is the extension of the proposed complementariness measure so that it takes into account the dynamics of the optimal decision process and hence becomes numerically as stable as the optimal decision process against unreliable partition statistic estimates.

REFERENCES

[1] J. W. Picone, "Signal modeling techniques in speech recognition," *Proceedings of the IEEE*, vol. 79, pp. 1214–1247, April 1994.

[2] T. K. Ho, J. J. Hull, and S. N. Srihari, "Decision combination in multiple classifier systems," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 16, pp. 66–75, January 1994.

[3] J. P. Campbell(Jr.), "Speaker recognition: A tutorial," *Proceedings of the IEEE*, vol. 85, pp. 1436–1462, September 1997.

[4] H. Gish and M. Schmidt, "Text-independent speaker identification," *IEEE Signal Processing Magazine*, pp. 18–32, October 1994.

[5] A. E. Rosenberg and F. K. Soong, "Recent research in automatic speaker recognition," in *Advances in Speech Signal Processing* (S. Furui and M. M. Sondhi, eds.), pp. 701–738, New York: Marcel Deccer Inc., 1992.

[6] K. J. Lang, *A Time-Delay Neural Network Architecture for Speech Recognition*. PhD thesis, Carnegie Mellon University, School of Computer Science, Pittsburg, PA 15213, July 1989.

[7] N. J. Nilsson, *Learning Machines: Foundations of Trainable Pattern Classifying Systems*. McGraw Hill, 1965.

[8] R. M. Haralick, "The table look-up rule," *Communications in Statistics - Theory and Methods*, vol. A5, no. 12, pp. 1163–1191, 1976.

[9] L. Xu, A. Krzyżak, and C. Y. Suen, "Methods of combining multiple classifiers and their applications to handwriting recognition," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 22, pp. 418–435, May/June 1992.

[10] Y. Huang and C. Suen, "A method of combining multiple experts for the recognition of unconstrained handwritten numerals," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 17, pp. 90–94, January 1995.

[11] T. K. Ho, "Recognition of handwritten digits by combining independent learning vector quantizations," in *Proceedings of 2nd International Conference on Document Analysis and Recognition*, (Tsukuba Science City, Japan), pp. 818–822, October 1993.

[12] R. Battiti and A. M. Colla, "Democracy in neural nets: Voting schemes for classification," *Neural Networks*, vol. 7, no. 4, pp. 691–707, 1994.

[13] G. Rogova, "Combining the results of several neural network classifiers," *Neural Networks*, vol. 7, no. 5, pp. 777–781, 1994.

[14] D. M. Tax, R. P. Duin, and M. van Breukelen, "Comparison between product and mean classifier combination rules," in *Proceedings of 1st International Conference on Statistical Techniques in Pattern Recognition* (P. Pudil, J. Novovicova, and J. Grim, eds.), pp. 165–170, Institute of Information Theory and Automation, June 1997.

[15] J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas, "On combining classifiers," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, pp. 226–239, March 1998.

[16] F. Kimura and M. Shridhar, "Handwritten numerical recognition based on multiple algorithms," *Pattern Recognition*, vol. 24, no. 10, pp. 969–983, 1991.

[17] K. R. Farell and R. J. Mammone, "Data fusion techniques in speaker recognition," in *Modern Methods of Speech Processing* (R. Ramachandran and R. J. Mammone, eds.), ch. 12, pp. 279–297, Boston, Massachusetts: Kluwer Academic Publishers, 1995.

[18] Y. Bennani and P. Gallinari, "Neural networks for discrimination and modelization of speakers," *Speech Communication*, vol. 17, pp. 159–175, 1995.

[19] M. Demirekler and A. Saranlı, "A study on improving decisions in closed set speaker identification," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, (Munich, Germany), pp. 1127–1130, April 1997.

[20] V. Radová and J. Psutka, "An approach to speaker identification using multiple classifiers," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, (Munich, Germany), pp. 1135–1138, April 1997.

[21] B. Achermann and H. Bunke, "Combination of face classifiers for person identification," in *Proceedings of IAPR International Conference on Pattern Recognition*, (Vienna, Austria), pp. 416–420, 1996.

[22] R. Brunelli and D. Falavigna, "Person identification using multiple cues," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 17, pp. 955–966, October 1995.

[23] D. H. Wolpert, "Stacked generalization," *Neural Networks*, vol. 5, pp. 241–259, 1992.

[24] J. A. Benediktsson, J. R. Sveinsson, O. K. Ersoy, and P. H. Swain, "Parallel consensual neural networks," *IEEE Transactions on Neural Networks*, vol. 8, pp. 54–64, January 1997.

[25] J. A. Benediktsson and P. H. Swain, "Consensus theoretic classification methods," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 22, pp. 688–704, July/August 1992.

[26] B. V. Dasarathy, *Decision Fusion.* Los Alamitos: IEEE Computer Society Press, 1994.

[27] J. S.-J. Lee, J.-N. Hwang, D. T. Davis, and A. C. Nelson, "Integration of neural networks and decision tree classifiers for automated cytology screening," in *Proceedings of the International Joint Conference on Neural Networks*, vol. 1, (Seattle), pp. 257–262, July 1991.

[28] Y. H. Hu, S. Palreddy, and W. J. Tompkins, "A patient-adaptable ECG beat classifier using a mixture of experts approach," *IEEE Transactions on Biomedical Engineering*, vol. 44, pp. 891–900, September 1997.

[29] K. Tumer and J. Ghosh, "Analysis of decision boundries in linearly combined neural classifiers," *Pattern Recognition*, vol. 29, pp. 341–348, February 1996.

[30] K. Tumer and J. Ghosh, "Theoretical foundations of linear and order statistics combiners for neural pattern classifiers," Tech. Rep. TR-95-02-98, Department of Electrical and Computer Engineering, University of Texas, Austin, USA, 1995.

[31] M. P. Perrone and L. N. Cooper, "When networks disagree: Ensemble methods for hybrid neural networks," in *Artificial Neural Networks for Speech and Vision* (R. J. Mammone, ed.), pp. 127–142, London, UK: Chapman & Hall, 1993.

[32] K. Al-Ghoneim and B. Kumar, "Learning ranks with neural networks," in *Proceedings of SPIE*, pp. 446–464, 1995.

[33] G. Mani, "Lowering variance of decisions using artificial neural networks portfolios," *Neural Computation*, vol. 3, pp. 484–486, 1991.

[34] L. Hansen and P. Salamon, "Neural network ensembles," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 12, pp. 993–1001, 1990.

[35] J. Pearl, *Probabilistic reasoning in intelligent systems: Networks of plausible inference.* Palo Alto, CA, USA: Morgan Kaufmann Publishers, 1988.

[36] J. Pearl, "Fusion, propagation and structuring in belief networks," in *Readings in Uncertain Reasoning* (G. Shafer and J. Pearl, eds.), Representation and Reasoning, ch. 6, pp. 366–414, San Mateo, California: Morgan Kaufmann Publishers, 1990.

[37] G. Shafer, *A Mathematical Theory of Evidence.* New Jersey, USA: Princeton University Press, 1976.

[38] J. A. Barnett, "Computational methods for a mathematical theory of evidence," in *Proceedings of the 7th International Joint Conference on Artificial Intelligence*, (Vancouver, B.C. Canada), pp. 868–875, August 1981.

[39] E. Mandler and J. Shurmann, "Combining the classification results of independent classifiers based on dempster/shafer theory of evidence," *Pattern Recognition and Artificial Intelligence*, vol. 10, pp. 381–393, 1988.

[40] K. Woods, W. P. Kegelmeyer, and K. Bowyer, "Combination of multiple classifiers using local accuracy estimates," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, pp. 405–410, April 1997.

[41] E. M. Kleinberg, "Stochastic discrimination," *Annals of Mathematics and Artificial Intelligence*, vol. 1, pp. 207–239, 1990.

[42] T. K. Ho and E. M. Kleinberg, "Building projectable classifiers of arbitrary complexity," in *Proceedings of IAPR International Conference on Pattern Recognition*, (Vienna, Austria), pp. 880–885, 1996.

[43] C. Ji and S. Ma, "Combination of weak classifiers," *IEEE Transactions on Neural Networks*, vol. 8, pp. 32–42, January 1997.

[44] D. H. Wolpert, "A mathematical theory of generalization: Part 1," *Complex Systems*, vol. 4, pp. 151–200, 1990.

[45] K. Fukunaga, *Introduction to Statistical Pattern Recognition*. Boston: Academic Press, 2 ed., 1990.

[46] K. Tumer and J. Ghosh, "Classifier combining: Analytical results and implications," in *Proceedings of the National Conference on Artificial Intelligence*, (Portland, USA), August 1996.

[47] K. Tumer and J. Gosh, "Estimating the bayes error rate through classifier combining," in *Proceedings of IAPR International Conference on Pattern Recognition*, (Vienna, Austria), pp. 695–699, 1996.

[48] J. Kittler, M. Halef, and R. Duin, "Combining classifiers," in *Proceedings of IAPR International Conference on Pattern Recognition*, (Vienna, Austria), pp. 897–901, August 1996.

[49] T. K. Ho, *A Theory of Multiple Classifier Systems and Its Application to Visual Word Recognition*. PhD thesis, Department of Computer Science, State University of New York at Buffalo, May 1992.

[50] D. Black, *The Theory of Commitees and Elections*. London: Cambridge University Press, 2nd ed., 1963.

[51] F. K. Soong and A. E. Rosenberg, "On the use of instantaneous and transitional spectral information in speaker recognition," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 36, pp. 871–879, June 1988.

[52] A. Saranlı, "Multi-classifier decision combination techniques in speaker identification and verification, Ph.D. Thesis Proposal," tech. rep., Dept. of Electrical and Electronics Engineering, Middle East Technical University, December 1997.

[53] A. Saranlı, "Rank-based multiple-classifier decision combination problem: A theoretical investigation," Tech. Rep. TR-98-01, Dept. of Electrical and Electronics Engineering, Middle East Technical University, April 1998.

[54] A. Saranlı, "Rank-based multiple-classifier decision combination for speaker identification: Progress on theory and experimental results," Tech. Rep. TR-98-02, Dept. of Electrical and Electronics Engineering, Middle East Technical University, December 1998.

[55] A. Saranlı and M. Demirekler, "A statistical unified framework for rank-based multiple classifier decision combination." To appear in *Pattern Recognition*.

[56] A. Saranlı and M. Demirekler, "Rank-based multiple classifier decision combination: A theoretical study," in *Proceedings of the IEEE International Workshop on Intelligent Signal Processing*, (Budapest, Hungary), pp. 51–56, September 1999.

[57] R. J. McEliece, *The Theory of Information and Coding*, vol. 3 of *Encyclopedia of Mathematics and Its Applications*. London: Addison-Wesley Publishing Co., 1977.

[58] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*. Signal Processing, Prentice Hall Inc., 1978.

[59] J. R. Deller, J. G. Proakis, and J. H. L. Hansen, *Discrete-Time Processing of Speech Signals*. New York: Macmillan Publishing Company, 1993.

[60] A. V. Oppenheim and R. W. Schafer, *Discrete-Time Signal Processing*. Signal Processing, Prentice Hall Inc., 1989.

[61] R. G. Y. Linde, A. Buzo, "An algorithm for vector quantizer design," *IEEE Transactions on Communications*, vol. 20, pp. 84–95, January 1980.

[62] B. Atal, "Automatic recognition of speakers from their voices," *Proceedings of the IEEE*, vol. 64, pp. 460–475, 1976.

[63] Y.-H. Kao, L. Netsch, and P. Rajasekaran, "Speaker recognition over the telephone channels," in *Modern Methods in Speech Processing* (R. Ramachandran and R. J. Mammone, eds.), ch. 13, pp. 299–321, Boston, Massachusetts: Kluwer Academic Publishers, 1995.

[64] S. Furui, "Research on individuality features in speech waves and automatic speaker recognition techniques," *Speech Communication*, vol. 5, pp. 183–197, 1986.

[65] J. He, L. Liu, and G. Palm, "A new codebook training algorithm for vq based speaker recognition," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, (Munich, Germany), pp. 1091–1094, April 1997.

[66] S. Euler, R. Langlitz, and J. Zinke, "Comparison of whole word and subword modeling techniques for speaker verification with limited training data," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, (Munich, Germany), pp. 1079–1082, April 1997.

[67] D. Hardt and K. Fellbaum, "Spectral subtraction and rasta filtering in text-dependent hmm-based speaker verification," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, (Munich, Germany), pp. 867–870, April 1997.

[68] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian Mixture speaker models," *IEEE Transactions on Speech and Audio Processing*, vol. 3, pp. 72–83, January 1995.

[69] D. A. Reynolds, "Speaker identification and verification using Gaussian Mixture speaker models," *Speech Communication*, vol. 17, pp. 91–108, 1995.

[70] D. A. Reynolds *et al.*, "The effects of telephone transmission degradations on speaker recognition performance," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, (Detroit, Michigan, USA), pp. 329–332, May 1995.

[71] C. shi Liu and H. chuan Wang, "A segmental probabilistic model of speech using orthogonal polynomial representation: Application to text-independent speaker verification," *Speech Communication*, vol. 18, pp. 291–304, 1996.

[72] J. Lei and L. O. Hall, "Speaker recognition with a self configuring neural network," in *Proceedings of IEEE International Conference on Neural Networks*, (Houston, Texas, USA), pp. 2351–2354, IEEE, 1997.

[73] T. Artieres and P. Gallinari, "Multi-state predictive neural networks for text-independent speaker recognition," in *Proceedings of the European Conference on Speech Communication and Technology*, (Madrid, Spain), pp. 633–636, September 1995.

[74] J. Hennebert *et al.*, "The polycost 250 database (v1.0)," tech. rep., EPFL, June 1996.

[75] J. T. McClave and T. Sincich, *A First Course in Statistics*. Prentice Hall, 1995.

VITA

Afşar Saranlı was born in Ankara, Turkey on July 10, 1971. He received his B.Sc. degree in Electrical and Electronics Engineering with Honors from Middle East Technical University, Ankara in 1993. He has been awarded the British Council Fellowship and received his M.Sc. degree in Telecommunications and Signal Processing with Distinction from the Department of Electrical and Electronic Engineering, Imperial College of Science, Technology and Medicine, University of London, UK in 1994. In 1995 he has rejoined the Department of Electrical and Electronics Engineering of Middle East Technical University as a Ph.D. student and a research assistant. As of January 2000, he is still working as a research assistant there, also taking part in teaching activities especially in digital signal processing. His areas of interest include digital signal and speech processing for communications, pattern recognition, multiple classifier systems and neural networks.

His publications are:

- A. Saranlı and M. Demirekler, "A Unified View of Rank-Based Decision Combination," Submitted to ICPR'2000 *International Conference on Pattern Recognition*, November 1999.

- A. Saranlı and M. Demirekler, "A statistical unified framework for rank-based multiple classifier decision combination," To appear in *Pattern Recognition*.

- A. Saranlı and M. Demirekler, "Rank-based multiple classifier decision combination: A theoretical study." Submitted to *International Journal on Advanced Computational Intelligence*, November 1999.

- A. Saranlı and M. Demirekler, "Rank-based multiple classifier decision combination: A theoretical study," in *Proceedings of the IEEE International Workshop on Intelligent Signal Processing*, (Budapest, Hungary), pp. 51–56, September 1999.

- A. Saranlı and M. Demirekler, "The POLYCOST Database and VQ-based closed-set text-independent speaker identification," in *Proceedings of SIU'99 National Confernce on Signal Processing and its Applications*, (Ankara, Turkey), June 1999.

- A. Saranlı and B. Baykal, "Complexity reduction in radial basis function (RBF) networks by using radial B-spline functions," *Neurocomputing*, vol. 18, pp. 183–194, 1998.

- M. Demirekler and A. Saranlı, "A study on the convergence properties of evolution strategies(ES) with a case study on finding the global optimum solution of the multi-pulse excitation problem," *ELEKTRIK, Turkish Journal of Electrical and Electronics Engineering*, vol. 5, pp. 325–246, March 1997.

- M. Demirekler and A. Saranlı, "A study on improving decisions in closed

set speaker identification," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, (Munich, Germany), pp. 1127–1130, April 1997.

- A. Saranlı and B. Baykal, "Real-time prediction of chaotic time-series using radial basis function (RBF) networks," in *Proceedings of SIU'96 National Confernce on Signal Processing and its Applications*, (Antalya, Turkey), pp. 107–113, May 1996.

- A. Saranlı, "Finding the global optium pulse locations in the multi-pulse excitation method using evolutionary strategies(ES)," in *Proceedings of SIU'96 National Confernce on Signal Processing and its Applications*, (Antalya, Turkey), pp. 5–11, May 1996.

- A. Saranlı and B. Baykal, "Chaotic time-series prediction and the Relocating-LMS (RLMS) algorithm for radial basis function networks," in *Proceedings of The European Signal Processing Conference*, (Trieste, Italy), September 1996.

- A. Saranlı, "Investigation of an alternative B-spline basis in adaptive RBF networks, with applications to system identification and time-series prediction," Master's thesis, Imperial College of Science, Technology and Medicine, University of London, 1994.