AFRL-IF-RS-TR-1999-85
Final Technical Report
April 1999

DIALECT IDENTIFICATION

ITT Aerospace/Communications Division

Alan Higgins, Peter Benson, K. P. Li, and Jack Porter

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

19990726 068

AIR FORCE RESEARCH LABORATORY
INFORMATION DIRECTORATE
ROME RESEARCH SITE
ROME, NEW YORK

DTIC QUALITY INSPECTED 4

This report has been reviewed by the Air Force Research Laboratory, Information Directorate, Public Affairs Office (IFOIPA) and is releasable to the National Technical Information Service (NTIS). At NTIS it will be releasable to the general public, including foreign nations.

AFRL-IF-RS-TR-1999-85 has been reviewed and is approved for publication.

APPROVED:

SHARON M. WALTER
Project Engineer

FOR THE DIRECTOR:

JOHN V. MCNAMARA, Tech Advisor
Information & Intelligence Exploitation Division
Information Directorate

If your address has changed or if you wish to be removed from the Air Force Research Laboratory Rome Research Site mailing list, or if the addressee is no longer employed by your organization, please notify AFRL/IFEC, 32 Brooks Road, Rome, NY 13441-4114. This will assist us in maintaining a current mailing list.

Do not return copies of this report unless contractual obligations or notices on a specific document require that it be returned.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503.

1. AGENCY USE ONLY *(Leave blank)*

2. REPORT DATE
April 1999

3. REPORT TYPE AND DATES COVERED
Final Oct 94 - May 98

4. TITLE AND SUBTITLE
DIALECT IDENTIFICATION

5. FUNDING NUMBERS
C - F30602-94-C-0289
PE - 35885G
PR - 1049
TA - L0
WU - 02

6. AUTHOR(S)
Alan Higgins, Peter Benson, K. P. Li, and Jack Porter

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
ITT Aerospace/Communications Division
ITT Defense & Electronics
100 Kingsland Road
Clifton NJ 07014-1993

8. PERFORMING ORGANIZATION REPORT NUMBER
N/A

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Air Force Research Laboratory/IFEC
32 Brooks Road
Rome NY 13441-4114

10. SPONSORING/MONITORING AGENCY REPORT NUMBER
AFRL-IF-RS-TR-1999-85

11. SUPPLEMENTARY NOTES
Air Force Research Laboratory Project Engineer: Sharon M. Walter/IFEC/(315) 330-7890

12a. DISTRIBUTION AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*
The objective of this effort was to develop and evaluate a capability to automatically (by computer) determine the dialect spoken in samples of recorded speech.

14. SUBJECT TERMS
Dialect Identification, Speech, Language Identification, Dialect, Language

15. NUMBER OF PAGES
96

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UL

Table of Contents

1 Introduction 1
1.1 Additional Contract Information 1
1.1.1 Consultants on Dialectology 1
1.1.2 Database Used 1
1.1.3 Schedule and Period of Performance 2
1.2 Report Contents 2

2 Dialectology 3

3 Literature Survey 5
3.1 Goals 5
3.2 Procedural Differences 5
3.3 Latin American Spanish Dialectology Literature Survey 5
3.3.1 Organization of Latin American Spanish Dialects 7
3.3.1.1 Major Influences on Latin American Spanish 7
3.3.1.2 Imposing Order on Chaos 9
3.3.1.3 An Historic Approach 11
3.3.1.4 Cotton and Sharp's Approach 12
3.3.1.5 Resnick's Approach 13
3.3.1.5.1 Main Phonological Features Used by Resnick 14
3.3.1.5.2 Resnick's Organization of Latin American Dialects 15
3.3.1.5.3 Additional Features 16
3.3.2 Summary of Selected Properties of Cuban and Liman Dialects 16
3.3.2.1 Introduction 16
3.3.2.2 Citations from Cotton and Sharp 16
3.3.2.2.1 Cotton and Sharp on Cuban 17
3.3.2.2.2 Cotton and Sharp on Liman 18
3.3.2.3 Citations from Resnick 20
3.3.2.3.1 Method of Presentation 23
3.3.2.3.2 Comparing Liman Populations 23
3.3.2.3.3 Resnick on Cuban 24
3.3.2.3.4 Resnick on Liman 25
3.3.3 Inferences From the Literature 26
3.3.3.1 Consistency of Subject Matter 28
3.3.3.1.1 Sources and Implications of Differing Subject Matter 29
3.3.3.2 Consistency of Language Properties 30
3.3.3.2.1 Internal Inconsistencies 31
3.3.3.2.2 Cases of Incomparability 32
3.3.3.2.3 Cases of Disagreement 32
3.3.3.2.3.1 Disagreement on /f/ in Cuban Dialects 33
3.3.3.2.3.2 Disagreement on /x/ in Liman Dialects 33
3.3.3.2.3.3 Complexity of Resolving Disagreements 34
3.3.3.2.4 Ambiguous Cases 34
3.3.3.2.5 Cases of Questionable Agreement 36
3.3.3.2.6 Cases of Agreement 36
3.3.3.3 Conclusions About Consistency 36
3.3.3.4 Implied Differences and Similarities of Cuban and Liman Dialect Groups 37
3.3.3.4.1 Differences 37
3.3.3.4.2 /b/ in /lb/ Context 38
3.3.3.4.3 Allophones of /x/ 38
3.3.3.4.4 /s/: Generally 38
3.3.3.4.5 /s/: Word-initial 38
3.3.3.4.6 /r/ and /l/: Generally 39
3.3.3.5 Ambiguous Cases 39
3.3.3.6 Conclusions About Similarities and Differences 39

3.3.4 Consultant's Comments 39
3.3.4.1 General Comments 40
3.3.4.1.1 On Selection of Sources 40
3.3.4.1.2 On the Diversity of Spanish in Lima and in Cuba 40
3.3.4.1.3 On Choosing Subjects 41
3.3.4.1.4 Exceptions from the Literature 42
3.3.4.1.5 Reliable Phonological Characteristics of Cuban and Liman Dialects 43
3.3.4.1.6 On Differences Between Cuban and Liman 44
3.4 Arabic Dialectology Literature Survey 45
3.4.1 An overview of Arabic Dialects 45
3.4.2 Dialect Selection 48
3.4.3 Finer-Grained Analysis of the Selected Dialects 49
3.4.3.1 Understanding what distinguishes dialects of Arabic 50
3.4.3.1.1 Which Dialects share common reflexes of the /[dZ]/ 50
3.4.3.1.2 Which Dialects share common reflexes of the /q/ 51
3.4.3.1.3 Which Dialects share common reflexes of the /k/ 51
3.4.3.1.4 Which Dialects share common reflexes of the interdentals 52
3.4.3.1.5 Which Dialects share common reflexes of the long vowels 52
3.4.3.1.6 Which Dialects share common reflexes of the short vowels 53
3.5 Conclusions of the Literature Survey 54

4 Baseline System Description 56
4.1 Baseline System Block Diagram 56

5 Database 58
5.1 Concentration on Cuban and Liman 58
5.2 Database Segmentation; Definition of Classes 59

6 Baseline System Testing 62
6.1 Database and Recognition Task 62
6.2 Summary of Most Important Results 62
6.2.1 Dialect and Language Identification Compared 62
6.2.2 Effect of Test Sample Duration 63
6.2.3 Effect of SNR 68
6.2.4 Effects of Channel Bandwidth 71
6.2.4.1 Elimination of Highest-Frequency Filter Channels 71
6.2.4.2 Elimination of Lowest-Frequency Filter Channels 74
6.2.4.3 Modified Filter Banks 75
6.2.5 Spectral Tilt 76
6.2.6 Performance Versus Number Reference Speakers per Dialect Class 77

7 DID System Development 80
7.1 Error Analysis 80
7.2 The Highland-Lowland Distinction 80
7.3 Speaker-independent System Tests; Comparison with LID 81
7.4 Experiments Using Syllabic Prosodic Features 81
7.5 Listener-identified Dialect-specific Segments 82

8 Conclusions 84

1. Introduction

The objective of this work was to develop and evaluate a capability to automatically determine the dialect spoken in samples of recorded speech.

Language identification (LID) software automatically determines the language spoken in samples of recorded speech. ITTI had developed several LID programs when the testing phase of this effort began, the two most important being a "speaker dependent" and a "speaker independent" version. Both of these programs served as baseline systems. In broad terms, this effort had as one objective the testing of pre-existing LID algorithms on a dialect identification (DID) task, and subsequent development of a baseline system to improve its DID performance as another objective. The testing was specifically required to assess the effect on DID performance as another objective. The testing was specifically required to assess the effect on DID performance (i.e., accuracy) of operating parameters known to be important in tactical applications of speech-related automatic recognition algorithms, including speech segment duration, signal-to-noise ratio (SNR), bandwidth, amount of available dialect sample data, and spectral tilt variations such as those which are introduced by various communications channels.

At the outset of the effort, ITTI was directed to consider wholly new approaches to DID, apart from the techniques previously found useful for automatic LID and incorporated in the pre-existing LID programs. Priority was to be given to consideration of the findings of academic dialectologists, by examining the technical dialectology literature in hopes of finding additional, new approaches to automatic DID. A literature survey and analysis of the potential of what was found there for automatic DID was therefore also a high priority objective of this effort.

1.1. Additional Contract Information

Some further details about the contract help to fill out an overview of its execution.

1.1.1. Consultants on Dialectology

ITTI enlisted the aid of two experts for reviewing and interpreting the technical literature of Spanish and Arabic dialectology. Professor John Lipski of the Spanish Department of the University of New Mexico, a specialist in Spanish dialectology (and author of a recent book on the subject) helped interpret that subject for us, and provided interesting opinions on the prospects for successful automatic DID. Professor Alan Kaye, an Arabic dialectologist of world renown at California State University, Fullerton provided counsel and direction in assembling and interpreting what material there is available on Arabic dialects.

1.1.2. Database Used

The database used for testing and subsequent development of the baseline DID system was collected under direction of the Air Force Research Laboratory-Rome Research Site (AFRL-RRS). It consists of speech from a set of native Latin American Spanish speakers recorded in Miami, Florida. The total number of speakers interviewed and recorded is 214. Two linguists supervised and participated in the collection of the interview and speech data, and its subsequent documentation and marking.

1.1.3. Schedule and Period of Performance

This contract was executed in three phases. The literature survey was performed and documented between October, 1994 and April, 1995. The baseline algorithm was tested between July, 1996 and May, 1997, at which time algorithm development commenced. The technical research was completed in April, 1998. About two and one-half man years of effort were expended in the process.

1.2. Report Contents

Following this Introduction, Section 2 provides some background information on dialect variabilities within languages. Section 3 describes the Spanish and Arabic dialectology literature surveys and conclusions reached as a result of performing those surveys.

Section 4 presents a very brief description of the baseline (LID) program which ITTI had developed prior to the start of the subject contract work.

Section 5 presents information about the Latin American Spanish dialect database which AFRL-RRS had collected, including its labeling and subsequent subdivision into three classes for automatic DID experimental purposes.

Section 6 is a summary of the detailed testing performed on the baseline program, to determine its robustness or sensitivity to system operating variables known to be important in tactical applications of speech-based recognition systems.

Section 7 describes the research performed in an effort to improve the performance of the baseline system in its application to DID. This research included an attempt to use some of the observations found in the Latin American Spanish dialect literature survey.

Finally, Section 8 contains conclusions reached about various aspects of automatic DID, especially with regard to Latin American Spanish dialects, as indicated by results obtained on the database provided by AFRL-RRS.

2. Dialectology

Dialect diversity is a common fact of variability within languages. Languages are spoken by communities of people who understand one another for the most part. One mechanism for dialect development is related to cultural expansion. As a language spreads over a large geographical area, the details of the language change and the geographical variants of the language take on distinctive identities. The distinctions can have many forms, from accent differences to differences in syntax and semantics. However, dialect distributions are rarely simply geographical, as people emigrate into one geographical location from another so that there may be pockets of speakers of one dialect within the geographic domain of another.

Social and professional dialect differentiation occurs commonly and perhaps universally as well. Differences in the words chosen to express concepts form the core of jargons or professional dialects. Differences in area of origin, previous language experience and education may provide the basis for social dialect differentiation. One talker may participate in several dialect groupings, using the dialect most appropriate to his immediate situation. (This phenomenon is known as "code switching".) Dialect diversity then is a common fact of languages and is correlated with where, with whom and about what, people talk.

Being able to identify the dialect that a person uses is a capability that could serve many functions. Professor Henry Higgins, a character in the play "Pygmalion" and fashioned after the real-life phonetician Henry Sweet, could place a speaker within blocks of his home in London, merely by attending to the phonetics (and no doubt word choice) of the speaker's speech. Knowing the origin of speakers in a military venue might inform one of the chain of command and plausible clues as to the identity of the talker. It is not inconceivable that being able to recognize the dialect of a talker would assist in improving the recognition performance of an automatic speech recognition system. A more precise language or acoustic model, selected on the basis of dialect, might provide a direct boost to recognizer performance.

Creating an automatic system for identifying the dialect of a speaker is a difficult technological challenge, in part because many extraneous factors affect the acoustic signal on which the decision must be based. Variations in the acoustics of speech between talkers are much larger than variations between languages and it is likely that variations between dialects are smaller than variations between languages.

The differences between dialects can be found at all levels of linguistic analysis but current speech processing techniques may be relatively insensitive to some of the differences that human listeners find salient. For example, to be able to capture syntactic and semantic differences between dialects requires a language model of some size and complexity. For example the distinctive use of the ustitive tense in the American dialect called Black English, as in "He be going to the store" is a clear discriminator of this sentence from Standard American. In an automatic system, this discriminator is only going to be useful as a feature if there is a speech recognition system that can recognize these words reliably. Function words such as "be" are very difficult to recognize reliably, even for co-operative speakers in good acoustic channels. It is unlikely that the current state of the art in speech recognition will support such delicate distinctions. If one adds the fact that in a military environment the messages

tend to be short and with a formal phraseology and syntax, the distinction one finds in syntax, semantics, and perhaps even word choice will be limited in their contribution to dialect identification, even if it could be assumed that most words could be recognized reliably.

Accent is a key area for dialect distinctiveness that is amenable to automatic processing. Accent is largely described by the transformations of the phonetics that separate one set of pronunciations from others. For example, American English has dialects which are described as [r]-less* in that they do not have a phonetic [r] sound in places where other dialects have one. [pak] for [park], in the New England dialect of American English shows a transformation wherein the [r] is deleted in circumstances where in other dialects it is not. As a principle concern of dialectology is to catalog and categorize these transformations, the literature of dialectology may provide invaluable guides in developing an automatic dialect recognition capability.

* The orthography of a word, *i.e.*, the dictionary spelling, is usually represented by placing the spelling in double quotes. The phonemic, or what might be called the underlying phonic, spelling of a word is usually represented between slashes, *e.g.*, /pit/ is the phonemic spelling of the name "Pete". The phonetic quality of a sound is represented by the phonetic transcription enclosed in square brackets; thus [p^h] is an aspirated /p/ sound, found in word-initial position in English.

3. Literature Survey

3.1. Goals

The goals of the literature survey were to obtain an understanding of the differences between the dialects of interest as viewed by academic experts in the field, and to assess the potential use of those differences in an automatic DID system. If the differences cited had such potential, developing techniques for using them would become a research topic later in the contract work.

The literature survey covers two dialects of Latin American Spanish and five dialects of vernacular Arabic. These two languages were selected because they were the target languages for the algorithm development portion of the study. These languages were originally the ones for which databases of operational material was to be collected. In lieu of that material, a database collection effort was mounted by AFRL-RRS for Latin American Spanish.

Standard linguistic notation is followed: orthography appears in italics, phoneme strings are surrounded by slashes (/.../), and phonetic spellings appear in brackets ([...]).

3.2. Procedural Differences

The literature surveys of Latin American Spanish and of Arabic were executed in somewhat different ways for several reasons, including; a) different personnel were assigned to work on the two languages; b) different consultants were used for the two languages; c) the richness of material differs for the two languages; and d) the Latin American Spanish dialect database was developed in parallel with the literature survey on Spanish, and helped concentrate interest in specific dialects of that language. For these reasons, the two literature surveys differ in scope and detail and are described separately.

3.3. Latin American Spanish Dialectology Literature Survey

There is almost no general agreement among Spanish dialect experts on a system of organization of Latin American Spanish dialects. To illustrate the diversity of opinions and methods, two different approaches to organizing Latin American Spanish dialects are described; that of M. C. Resnick and that of Cotton and Sharp. As many of the recognized experts in this field have spent major portions of their lifetimes studying the problem, hence are very much more qualified than we to speak to this issue, it would be misleading and inappropriate for us to endorse one method of classification over any other. The approaches of these two authors are therefore presented *only* to illustrate the diversity of expert opinion on the subject of organization.

The second part of this literature review presents specific material from the literature on phonological properties of two dialect groups which are well represented in the new AFRL-RRS Spanish Dialect database. The two groups are Cuban and Liman, *i.e.*, the Spanish of a segment of the population of Lima, Peru. We refer to these as "dialect groups" because, when pressed, dialectologists almost always delight in pointing out that there are consistent differences at some phonological level across certain sub-populations of any larger population; hence one would be unlikely to find any dialect expert who would claim that there is

only one dialect spoken in Cuba, or by any particular group of Limans. There does seem to be some agreement that one should expect greater differences between than among the Cuban and Liman subjects in the database. Even that level of agreement requires stipulation that the Liman subjects in the database are predominantly of middle- or upper-level socio-economic standing and, presumably, education. With this justification, then, the second part of this literature review summarizes two writers' descriptions of the phonology of the speech of Cubans and the Limans of interest.

The specific phonological material in the second part of this review serves several purposes. First, it exemplifies the kind of material found in dialect literature. That is, it shows what experts consider the salient properties of dialects, and how those properties are described. A major interest in this project is an assessment of how useful dialectological data may be, as a basis for an automatic dialect identification system. *How* dialect properties are described by these experts turns out to be very relevant to this question.

The technique almost universally used to describe any Latin American Spanish dialect is by comparison with a "normal", or dominant mode of pronunciation, which the writers assume to be well known to the reader. The reference to the expected form of pronunciation is often mediated by reference to orthography, which is in fact standardized in the Spanish speaking world. An example will help clarify the point. Cubans and Limans are both claimed by Resnick to sometimes drop the /d/ in the frequently occurring suffix *-ado*, producing a sound [áo] or [áw]. To detect that this is happening, it must be known when something which is *normally* pronounced one way is in fact being pronounced a different way. This is easy for fluent Spanish speakers to do, as the circumstances calling for the *-ado* suffix are transparently well known to them, but these same circumstances may be very difficult for an automatic dialect recognizer to detect. One might try detecting sound sequences of the forms [á ð o], [áo] and [áw], and checking their relative frequencies of occurrence. That technique might in fact work, but notice that it uses a dialect property *different* from what the dialect literature cites; the literature doesn't say anything about the relative frequencies of those three sound sequences, except by a remote and in fact uncertain implication. Any such observation would have to take into account the frequency with which the [áo] and [áw] sound sequences may arise naturally in other contexts, which would confound any statistical evidence which might otherwise exist of dropped /d/'s. The common tendency to characterize dialects by comparison to a standard or expected pronunciation thus tends to put the citations found in the literature at a considerable remove from *direct* application in automatic dialect recognition.

When one becomes aware of them, these indirect and sometimes obscure references to an expected form of pronunciation are seen to be ubiquitous in the literature. Cotton and Sharp go so far as to refer to "Standard Spanish" and Resnick also speaks of the "standard language" but some experts object that there really is no such thing. Lipski, for example, points out that "No such thing as universal 'standard Spanish' is recognized in the Spanish-speaking world, in any individual country, or by linguists working with Spanish." Nevertheless, normative standards of pronunciation are so constantly used in the literature that formulating ways to translate phonological observations based on them into automatic procedures may pose the most difficult aspect of using the literature as a guide in developing dialect

identification algorithms for use with free text. This and related issues will be discussed thoroughly, should any attempt be made to apply findings from the literature.

The second purpose of the specific phonological citations from the literature on Cuban and Liman is that they afford an opportunity to assess, in a rough way, how consistent this small sample of the literature is in its prediction of properties of these two dialects. It will also be interesting to evaluate how consistent the two sources are with respect to their choice of subject matter, *i.e.*, which phonological features each evaluates.

The third purpose the collected specific phonological citations from the literature serves is as a basis for illustrating one of the steps which may be necessary in developing algorithms for distinguishing an arbitrary pair of dialect groups. As the material found in the literature is like what is presented here, *i.e.*, it is given in the form of separate and independent descriptions of the dialects, the first step in developing an automatic discriminator may be to find what differences these descriptions imply exist between the chosen dialects. If only the differences were presented in this document, the reader would have no opportunity to observe the process of deriving the differences from the descriptions actually encountered in the literature. A general-purpose dialect identification capability, based on technical dialect literature and intended to distinguish any pair of dialects, is likely to include this difference extracting step, so it is documented here, by way of example.

Finally, direct citations from the literature are necessary to provide traceability of the developed algorithm to its sources in the literature, should such a connection be established.

3.3.1. Organization of Latin American Spanish Dialects

As mentioned above, it is the "proper" organization of Latin American Spanish dialects that the experts can't seem to agree on. There are some tangible reasons for this which should be understood before examining some of the approaches to organization which have been developed in the past or are currently proposed. The critical fact is that Spanish in the Americas has been influenced by several factors which have had strong and roughly equal influence. Adding to that difficult situation is that at least one of those factors is quite complex in and of itself. Chief among these high impact factors is (without any priority intended) are:

a. differences which existed within peninsular Spanish when it was imported to the Americas;

b. a diverse set of indigenous languages pre-existing there; and

c. a persistent correlation of linguistic influences with social stratification.

The account which follows can be found in Cotton and Sharp, and seems to be in general agreement with other authors

3.3.1.1. Major Influences on Latin American Spanish

Peninsular Spanish *i.e.*, Spanish in Spain, had a diverse and strong dialectal structure before it was exported to the West. The strongest division was between the dialect of the *altiplano*, the central plateau including Castile, and the southern regions of the country, including Andalusia. The seat of power was in Castile and it was natural that administrative and

governing officials, primarily from that region, took their Spanish to the New World administrative centers which (except for Lima), were in the inland highlands. The sea-faring people, in contrast, came from the other region of Spain, and from the Canary Islands, which shared the dialect of the south of Spain. They interacted primarily with the people on the coastlines of the area, so there developed a highland-lowland distinction in the New World that paralleled the Castile-Andalusia difference in Spain. That distinction persists as a major dialect division today within many countries, including Mexico, Colombia and Peru.

Lima, the only original administrative center on the coast, was subject from the first to both Castilian and Andalusian influences, so has long been dialectally diverse. Today, massive migration from all parts of the region to Lima has intensified that diversity.

Another, more complex factor affecting dialect is the effect of the indigenous languages on the imported Spanish. The pre-existing language substrate is seen to affect even the Spanish spoken by mono-lingual Latin Americans. This becomes less surprising when it is realized how well and alive several of those languages are, even to this day. Mayan is still more widely spoken than Spanish in parts of the Yucatan and is common in Guatemala; Guarani is widely spoken in the area between and adjacent to Brazil and Argentina and is the dominant, preferred language in Paraguay, and Quechua more than survives throughout the Andean region. In addition to these three indigenous languages, Taino and African import languages are influences in the Caribbean region, Nahuatl in central Mexico, Caribe in the north coast of South America, Aymara in Bolivia and part of Chile, and Querandi, Pampero and Mapuche all influence the Spanish of Chile and southern Argentina. Many of these indigenous languages are very different, as is their influence on Spanish. This brief overview of the language substrates over which Spanish was laid gives some idea of the complexity of their distribution and influence.

The third principal factor contributing to the complex organization of Latin American Spanish dialects is the persistence of a strong socio-linguistic effect. Society all over this region has always tended to be polarized into upper and lower classes, even today. The upper classes, as is usually the case, establish the prestige dialect of a region. For the new Spanish language, that was the dialect of the administrators and governors, *i.e.*, Spanish of the *altiplano*, or Castilian. This form of Spanish naturally became the form taught in the schools, so there developed a reinforcing correlation of dialect with educational level. The lower class seems to have been more influenced by the indigenous languages and dialects of southern Spain, so the socio-linguistic effect was present from the beginning of Spanish acquisition by the Latin Americans, and it is as persistent as the social stratification itself.

All three of the major influences on dialect just described do not respect political divisions. The socio-linguistic dimension is obviously independent of geographical divisions at any level above neighborhood in urban centers, as any larger area is likely to contain several social classes. The language substrate effect doesn't correlate well with political boundaries either. The native region of Guarani sprawls across boundaries of Argentina, Brazil, Uruguay, Paraguay and possibly other countries. Quechua is influential in the highland areas of most of the northern Andes, including territory in at least four countries. The highland-lowland distinction muddles dialects within countries too, as most Latin America countries have both coastal and mountainous regions. Taken all together, these major influences make

it clear why Spanish dialect correlates poorly with political geography, say at the level of country. Cuba and Peru are perhaps at the poles of variability in this respect, in that Cuba is significantly more homogeneous dialectally than most other Latin American countries, and Peru is about as diversified as any of those countries. Lipski says

> "...Peruvian Spanish is really a cover term for dialects which, from a phonetic point of view, are as different from each other as, say, Jersey City, Mobile, Omaha, Sydney, and Port of Spain. There is no widespread agreement of how many dialect zones are found in Peru..."

He goes on to identify important roles for each of the major effects discussed above. The indigenous Quechua causes dialect differentiation even between monolingual Spanish speakers and balanced bilinguals of the inner highlands; the strongest division over the whole country is a highland/lowland disparity, and socio-linguistic variability is a dominant factor within Liman Spanish.

The confused correlation, or lack thereof, between dialect and geographical area might be surprising to (U.S.) Americans, as we are accustomed to the notion of an identifiable southern accent, and, with varying degrees of certainty, Mid-western, New York and Jersey accents, *etc.*. The differences between the distributional characteristics of American English and Latin American Spanish are so great that Resnick devotes the fist page and a half of his book to carefully spelling them out, presumably to disabuse the reader of unwarranted expectations. In Latin American Spanish, one great difference is that one finds very similar dialects spoken in places thousands of miles apart, as well as very dissimilar dialects spoken in places close to one another. To replicate the effect in the United States (with some exaggeration), imagine a random redistribution of cities among the states, so Maine and Texan accents might be found in one state, and the Maine accent might be found in California and Maine.

3.3.1.2. Imposing Order on Chaos

It is not particularly difficult to produce an *arbitrary* scheme of classification of Latin American Spanish dialects. One could, for example, split first along the highland/lowland division, then along the indigenous language influence division, then the socio-linguistic, followed by other criteria to reach whatever level of differentiation might be desired.† Of course it is obvious that the leaves, or terminal nodes, of this tree structure would bear almost no relation to any form of geopolitical subdivision of the region. Resolution of countries, for example, would be lost at the first split. However, the preceding discussion has made it clear that it is impossible to organize Latin American Spanish dialects in a way which simultaneously brings together dialects which are similar phonologically and which also keeps together the dialects of countries. So the failure of this method of organization to preserve national - or other political - boundaries is not what the experts object to. It is the suggestion implicit in such a scheme that the splitting criteria at one level are more important or significant than the criteria at a lower level. Some experts would object to the arbitrary scheme above

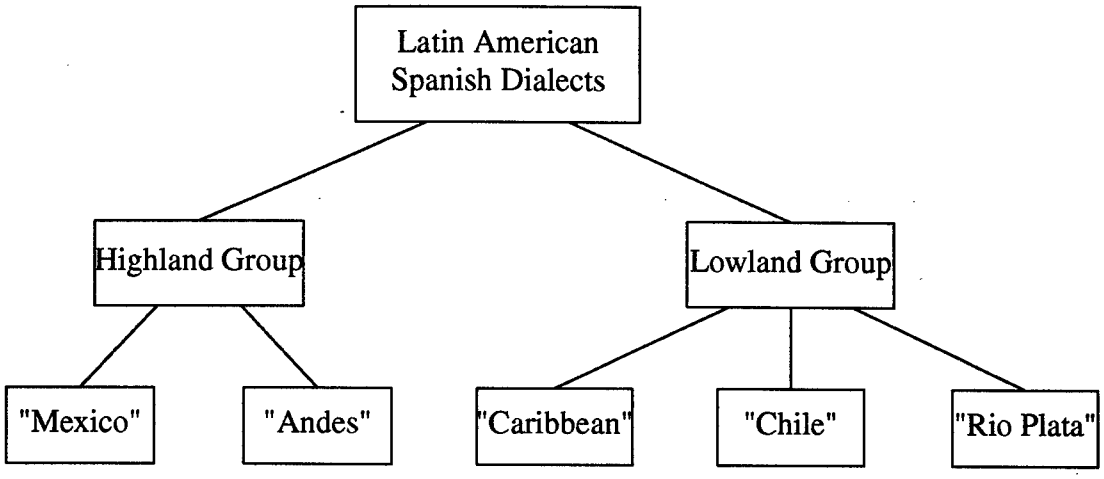
† The suggested splitting criteria would no doubt be replaced by roughly equivalent phonological criteria. For example, the highland-lowland split might be formulated in terms of how precisely consonantal clusters are articulated, or the treatment of syllable-final /s/ and /n/. The main point remains, however, that there is no obviously natural way, hence no widespread agreement, on how to *order* the differentiating criteria.

because, by splitting first on the highland-lowland dimension and then on the language substrate dimension subtly, at least, implies that the former is a more significant division that the latter.

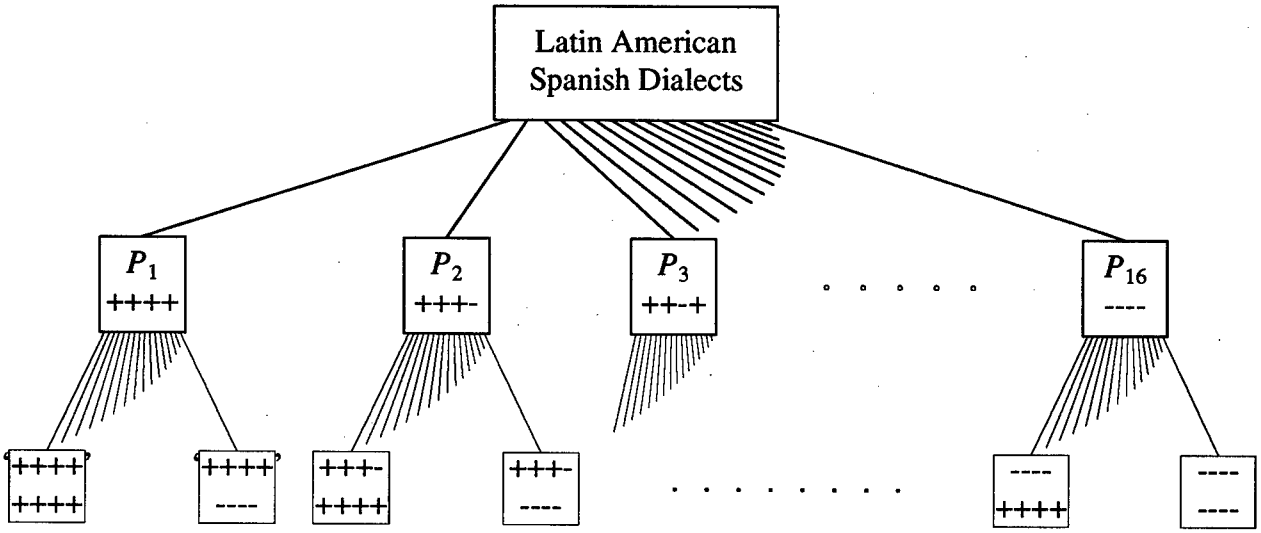
This position is quite reasonable. If the top level split is between highland and lowland varieties of Spanish, one would like to infer that the disparity between within-category and between-category differences is greater than would be achieved had a different splitting criterion been applied first. That kind of concern is always of interest in building classification systems in the biological sciences; taxonomy relies on that principle to justify inferring things about evolutionary development from taxonomic structure of a group of organisms. In modern data structure terms, one would like the organization to be a *bona fide* "dendrogram". This, then is the area of disagreement; the order of importance of dialect discriminators.

We will examine one approach to this problem which is historically important, and then the approach of Cotton and Sharp, and finally the approach taken by Resnick.

Figure 3.1 is a schematic representation of the tree structure suggested by the latter two approaches, which may aid in understanding the discussion.



per **Cotton and Sharp**



per **Resnick**

Figure 3.1: Schematic representations of two approaches to organizing Latin American Spanish dialects.

3.3.1.3. An Historic Approach

In 1921 Pedro Henriquez Urena set up five dialect zones based on the influence of Nahuatl, Caribe, Yarahuaca, Quechua, Aruacano and Guarani substrate languages Urena thus chose one of the three major influences discussed above as primary. His scheme proved useful, especially for studying the effects of the indigenous languages, but also more generally,

so that it is sometimes still used today. However, modern dialect experts generally reject this scheme on two grounds; first, it is inadequate, in the sense that it does not represent the real situation very well; and second, it is methodologically unsound because it is an *a priori* classification, based on what one expects to find in the language due to historical influences, rather than being derived from what is actually observed of the language.

3.3.1.4. Cotton and Sharp's Approach

Cotton and Sharp present a double system of classification of Spanish dialects in Latin America, and a lot of historical and linguistic information to justify their system. The main system is neatly summarized in the form of a map, on page 91 of their book. The five major regions and dialects are labeled "Mexico", "Caribbean", "Andes", "Chile" and "Rio Plata"; labels which unfortunately create the misleading impression that dialect correlates with country better than it does. The weakness of that correlation becomes evident in the description of the geographical regions most closely associated with their dialect categories:

The five regions are described in the following way:

Mexico. Mexico [except the coasts] and the adjacent areas of Southwestern United States and all of the republic of Guatemala.

Caribbean. Includes both coasts of Mexico and both coasts of Colombia, all of Venezuela and the islands of Puerto Rico, Cuba and the half of Hispaniola known as Santo Domingo.

Andes. Extends from the highlands of Colombia through Ecuador, Peru, Bolivia, northwestern Argentina and northern Chile excepting Lima.

Chile. The remainder of Chile not in the *Andes* category.

Rio Plata. All of Argentina not in *Andes*, [plus] Uruguay, and Paraguay.

(Most of Central America is cited as a mixture, or transition zone of these five dialect groups, and is seldom mentioned in their book.)

It is implied and almost stated that the dialect variability across these various categories is greater than within them. The specific example is given of Santa Fe, New Mexico and Guatemala City; although the dialects of these locations are quite different, they are more similar to each other than they are to dialects found in any of the other categories. But it is acknowledged that there is a lot of variability within each of these major categories.

Overlaid upon this principal, five category, classification is another, two category distinction, between highland and lowland or coastal Spanish. "The speech of the Mexican plateau, indeed, bears a closer resemblance to that of the Andes than does either one to those of the other three regions. And the three other regions, in turn, are like one another in many ways." So the second, more inclusive, classification these authors would impress on their five category system is a distinction between highland and lowland varieties. The former comprise the Mexico and Andes categories and the latter the Caribbean, Chile and Rio Plata categories. The lowland dialect regions are collectively known as the *Cono Sur*.

In the schematic representation of Cotton and Sharps' organizational scheme given in Figure 3.1, the highland-lowland division appears as an intermediate level of classification,

which is made possible by the consistency they have chosen to impose between that and the five category, substrate language-based classification scheme. At the lower level, the descriptors, *e.g.* "Mexico", are put in quotation marks to signify the very loose correlation these dialect groups have with the geographical and political entities the descriptors also represent.

The five principal regions defined by Cotton and Sharp correspond generally to regions of influence of substrate languages.† By presenting that form of division first, they seem to be giving precedence to that criterion. But then they reveal that they have also chosen their categories to be quite compatible with the highland-lowland distinction. Perhaps a fair characterization of their scheme of classification is as a joint highland-lowland and substrate language scheme.

Cotton and Sharp do not present any scheme of organization of dialects within any of their five major categories, but they list many, many distinctions within them. If one examines the descriptions of each of the five major dialect groups hoping to gain an understanding of the number and organization of the subdialects within each, one is quickly defeated by a maze of detail. Some useful properties of the larger group are given, but most of the information tends to be an immense welter of detailed *differences* which exist *within* the region. A brief excerpt from the description of the Andes category, found on page 178, illustrates the point.

> "In Colombia and Ecuador, /p/,/t/ and /k/ are frequently voiced, except in Bogota, where Quechua influence is absent, producing *chamba (champa), que dal (que tal)*, and *golumpio (columpio)*. By contrast, in Bolivia, /d/ is replaced by /t/ after a preceding /s/ and *desde* becomes *deste*. As elsewhere in the Hispanic world, the stops /b/, /d/ and /g/ have fricative allophones. In most of the Andean countries the graphemes *b* and *v* are pronounced identically as [β]. In Colombia, however, research now indicates at least 30% of Bogotanos make a phonemic distinction between [b] and [v]. Moreover, Colombians pronounce ..."

3.3.1.5. Resnick's Approach

Resnick finesses the problem of assigning priority to any special dialect discrimination factors in a clever way. After explaining that the correlation between dialect and geographical place is very complex for Latin American Spanish, and emphasizing that very similar dialects can be found thousands of miles apart while very dissimilar dialects are also spoken in one and the same location by different sub-populations, he goes on to present a scheme for inferring the likely origin of a Spanish speaker, *i.e.*, a place (perhaps modified by a socio-economic group or gender or age group label) where spoken Spanish is like that of the speaker in question.

His scheme of identification avoids preference for one of the major dialect-influencing factors over another by giving instead a set of eight phonological speech features to be evaluated in the subject's speech. These are "binary" features, meaning that each is to be given one of just two vaules, by convention a "+" or a "-", according as the feature is present in, or absent from, the speech sample, respectively. He then gives a set of "indices" for finding a place, and perhaps a socio-economic or other group, where the speech has characteristics

† Lipski finds their organization of Spanish dialects identical to that of Henriquez Urea.

implied by the pattern of signs. The binary features, and rules for assigning + or - to each, are presented in two groups of four each, the first group being labeled A1 through A4 and the second labeled B1 through B4. The indices are also arranged so that one can find a dialect's place from the sign pattern of the four "A" features alone, and then refine it or not using the "B" feature signs. These "preliminary" places are indicated symbolically in Figure 3.1 as P_1 through P_{16}.

Two features of Resnick's treatment are unusually interesting. One is that there is hardly any notion of any precedence among his eight features. They must be evaluated simultaneously, or at least in groups of four, to make use of his tables. His discussions makes it clear that he has no interest in evaluating or comparing the linguistic or dialectal significance of any of them; his only concern is that they are convenient and effective when used in connection with the indices he provides.

The other salient feature of Resnick's approach, and the most interesting one, is his avoidance of dialect labels. Notice that the indices he provides lead the user to a *location* whereof the subject's dialect is probably typical, and not a name or other form of dialect label. This stratagem is most interesting and significant. Dialect names and labels are in fact avoided throughout his book, in favor of what may be called the "characteristic place" of an individual's speech - a place, possibly further modified by a social group, gender or age group, of which the speech is typical. By concerning himself only with the correspondence of speech features and characteristic place, and never addressing questions like when the speech of two places is distinct enough to be considered different dialects, Resnick avoids many issues on which there is little or no agreement among experts. For example, one could count the number of characteristic places mentioned in his indices but, as we are not told which have distinct dialects, there is no way to count the number of dialects he treats. To do so one would need to know the correspondence between characteristic place and dialect, something he avoids discussing. A similar artifice was used in attaching "dialect labels" to the Spanish segment of the OGI database. The labels are, in fact or intention, characteristic places for the speakers.[Reka94]

3.3.1.5.1. Main Phonological Features Used by Resnick

The phonological subject matter of Resnick's main features, A1 through B4, is presented in Table 3.1

Feature	Subject Matter
A1	Retention of /s/
A2	*rr* as voiced apical trill
A3	/x/ as [h]
A4	*y* and *ll*
B1	*b* in *lb* context
B2	Word-final /n/
B3	/r/ and /l/
B4	Vowel voicing

Table 3.1: Subject matter of Resnick's main phonological dialect features.

3.3.1.5.2. Resnick's Organization of Latin American Dialects

Resnick seems to argue, at the beginning of his second chapter, that his method "of organization is exactly the opposite of a classification", presumably because he doesn't argue for, or even discuss, the relative importance of the features he uses, or the relation they might have to historical, environmental or social factors. With respect to the relative importance he attaches to his features, it might be argued there is an implicit suggestion that his "A" features are more important than the four "B" features, as the former are adequate for finding an initial estimate of the characteristic place. But he is mute on the subject. Unlike Dr. Losiewicz, I find little indication in Resnick's book that he attaches the greatest importance to his first-described feature, A1, which evaluates if /s/ is deleted, retained or modified in certain contexts[Losi94].† His stated criteria for selecting the features he did lists ease of evaluation first, and "separate out the largest number of socio-political-geographical entities from each other" as the last. So the A1 features may just be the easiest one to evaluate. The features must be used simultaneously, or at least in groups of four, to access his indices, further diluting any indication of special importance of any one.

Whether or not it is a method of classification, the scheme he presents may be described as method of searching a tree structure which relates his features to characteristic place. The tree has sixteen branches from the root to first level nodes (one branch for each possible assignment of signs to the four "A" features), and sixteen branches from each the first nodes to the leaves (one branch from each first level node for each possible assignment of signs to the four "B" features.) There are thus 16*16 = 256 leaves, or possible characteristic places, potentially described by his scheme. Furthermore, since each of these 256 leaves is reached by an evaluation which differs from every other on at least one speech feature, there are at least 256 different types of speech corresponding to those characteristic places. In this sense at least, Resnick differentiates sixteen possible dialects at the first level of discrimination and 256 possible dialects at the second. This method of organization precludes any simple overview or summary of Latin American Spanish dialects.

† Lipski points out that the overwhelming majority of Spanish dialect studies have dealt with /s/ and that whole dissertations have been devoted solely to surveying the literature of this limited subject. In spite of this evidence of a consensus among other authorities about the importance of the /s/ phenomena, I find no mention of its relative importance in Resnick's book.

3.3.1.5.3. Additional Features

Resnick describes, and gives guidance in using, many more than just the eight basic "A" and "B" features. These additional features, he claims,

> "...will provide the investigator with even finer dialectal discrimination and specification when needed than are possible with the 256 potential discriminations provided by [the first eight features]".

He apparently subscribes to the view that, by applying more and more *dialect* discriminators, one will eventually single out idiolects, *i.e.*, the speech of an individual at a particular time and circumstance. There is never any mention of the possibility that some discriminators he presents may distinguish a characteristic too fine to be considered dialectal.

3.3.2. Summary of Selected Properties of Cuban and Liman Dialects

3.3.2.1. Introduction

Descriptive material in the two references was considered relevant if it had some potential use as a basis for automatic dialect identification on speech material of a type stipulated in the Statement of Work for this project. The speech material described in the SOW typically consists of utterances by non-cooperative speakers with durations on the order of a few seconds. By "non-cooperative," we mean that subjects are not to be expected to speak any predictable text, or make any special effort to be understood. This type of speech precludes using "higher order" properties of dialect, which are only revealed in utterances lasting more than a few seconds. Lexical, syntactic, grammatical and semantic properties of dialect were generally considered unusable under the utterance duration constraint, so were treated as irrelevant for the purpose of this project. Attention was focused on the shorter duration phenomena of dialect, which consists of phonological properties which are manifest as acoustic events. Most of the properties collected are segmental, but as suprasegmental properties also have limited potential for use in an automatic dialect detector, some material on those properties was also gathered.

Spanish dialect specialists have long made use of the *vos-tu* distinction, between two very different forms of the second person familiar personal pronoun, and the verb forms they are used with. This otherwise useful property of speech was also ignored, as it is unlikely that familiar usage of any form will appear in the Spanish to which an automatic dialect recognizer might be applied. It would not arise spontaneously in an interview situation where the people talking are not on a first name basis. Therefore, it probably does not occur in the new AFRL-RRS Spanish Dialect database, either.

3.3.2.2. Citations from Cotton and Sharp

Cotton and Sharp cover Latin American dialects in five groups, as mentioned earlier. They put the Cuban dialects in a "Caribbean" group, and say that there is little differentiation within the country. Lima would normally belong to the Andean category, but they cite it explicitly as an exception. In the process of describing the geographical span of the "Andes" dialect group they state:

"... extending from the highlands of Colombia through Ecuador, Peru, Bolivia, northwestern Argentina, and northern Chile, with the exception of Lima, which belongs to this area geographically but not linguistically, ..."

As Lima is not again mentioned, it would appear that properties of Liman Spanish cannot be extracted from Cotton and Sharp. However, on the advice of our consultant, Professor John Lipski, we impute the properties assigned to coastal Peru, as found in Cotton and Sharp, to the Liman dialect group. Lipski warns us that Liman Spanish is diverse, especially on socio-linguistic lines, and is greatly confused by massive migration to Lima in recent years. However, he assures us that, if we can select on origin, we will find that native Liman subjects have coastal dialect characteristics. Without restricting attention to natives, there appears to be no reasonable way to associate dialect characteristics with Limans. We therefore assume that the Liman dialect group to be characterized is that of Lima natives. It must therefore be born in mind, in any application of the findings of this literature search, that those findings are based on the assumption that the Liman subjects are natives. This approach is reasonable because the new AFRL-RRS Spanish Dialect database does have enough information about each subject to assess origins. Further support for this approach may be found in the fact that most of the first twenty Liman subjects examined appear to be natives, and that Dr. Beth Losiewicz, one of the interviewers for the AFRL-RRS collection effort, feels that they are almost exclusively from the middle and upper classes and have dialect characteristics in common.

The number in angled brackets at the end of each citation from Cotton and Sharp is the number of the page from which it is taken. We used the 1988, Georgetown University Press edition of their book. Most of the citations are paraphrased versions of what is found in the book. Direct quotations are used instead whenever there is a possibility that precise interpretation might be difficult or subtle. Our comments, if any, appear in braces. Each citation is given a letter and number designation, for ease in later reference. Whenever specific examples of a phenomenon are given, they are taken directly from the text.

3.3.2.2.1. Cotton and Sharp on Cuban

Cub1. Stops are standard. <203>

Cub2. Most fricatives are standard, but pronounced with less tension and therefore less clarity {note reference to unspecified standard} <203>

Cub3. There is no phonemic distinction between *b* and *v*, and both are realized as [β]. <203>

Cub4. The voiceless counterpart (of *b* and *v*), *f*, is the bilabial fricative [ϕ] for the majority of speakers, but sometimes avoided by the elite. <203>

Cub5. [ð] is relaxed or, often, deleted. <203>

Cub6. "Fricative [ɣ] is the same as elsewhere." <203>

Cub7. [x] is in free variation over two forms: "It can be a very weak velar in which the tongue does not touch the velum, producing a sound similar to [h], or it may be an actual [h], laryngeal or pharyngeal." <204>

Cub8. When /s/ occurs syllable-final either before a pause or another consonant, it reduces to an aspiration or is simply deleted. <204>

Cub9. /s/ has no [z] allophone. "The tongue approaches the point of articulation of a following consonant, giving the air audible friction but allowing it to pass through the oral cavity without voicing." <204>

Cub10. When /s/ is followed by a nasal consonant, "a preceding vowel is nasalized and /s/ becomes the appropriate nasal, but devoiced"; [mismo] -> [mm̥ĩmo]. <204>

Cub11. Lateral /l/ is that of general Spanish. <204>

Cub12. "Implosive /l/ and /r/ are in free variation." <204>

Cub13. Castilian /ʎ/ is non-existent. <204>

Cub14. The vibrants /r/ and /rr/ are pronounced as in general Spanish. <204>

Cub15. "In many areas, syllable-final /r/ disappears before a pause or a following consonant." {It isn't clear that Cuba is such an area.} <204>

Cub16. ".../n/ before a pause, when word-final or when syllable-final before another consonant, becomes the velar [ŋ], especially when its syllable is a prefix that corresponds to an independent preposition like *en* or *con*." The prepositions *en* and *con* are thus regularly pronounced with final [ŋ], "and this practice is continued when *en* and *con* are internal to words; *entrar* -> [eŋtrár], *convenir* -> [koŋbenír]." <204>

Cub17. "In Standard Spanish, /e/, /a/ and /o/ have as allophones [ɛ], [ɑ] and [ɔ] in certain environments: *e.g.* when in contact with [r̄] [r̄éɣla] or when preceding /x/, [káxa] {Should this not be [káxa]?}, or in a syllable closed by any consonant except *m, n, s, d, x*; [sɑl]." <204>

Cub18. When syllable-final /s/ is aspirated or lost, a preceding /e/ -> [ɛ], /a/ -> [ɑ] and /o/ -> [ɔ]. <204>

Cub19. "If /s/ is deleted, the opening of the vowel may serve to supply an absent morpheme or identify a part of speech:
[pjé] '*pie*' [pjé] '*pies*'
[bé] '*ve*' [bé] '*ves*'
[djó] '*dio*' [djɔ] '*Dios*'." <204>

Cub20. "...the lowland dialects in general are delivered at a more rapid pace than in the highlands, and, as in the highlands, they differ from one another substantially" "The tempo of speech in Cuba is faster than in Santo Domingo, where it is relatively slow." <205>

Cub21. "Pitch is higher in Cuba, while in Santo Domingo it is fairly slow, as in Castile." <205>

3.3.2.2.2. Cotton and Sharp on Liman

Lim1. Graphemes *b* and *v* are pronounced identically as [β]. <178>

Lim2. /f/ is interchangeable with [β] and a number of other sounds such as [x] and [h]; *foto -> boto, elefante -> elebante, función -> junción, bufón -> bujón, fumo -> jumo.* <178>

Lim3. "When /f/ is pronounced, it is a bilabial [ɸ] in the popular speech of Peru and Ecuador, as in *juamilia* (*familia*) and *enjuermo* (*enfermo*)." {However} "In the prestige dialect of Peru [ɸ] is avoided." <178>

Lim4. "Fricative [ð]...is lost on the coast, even among the educated, resulting in forms such as *tuavía* (*todavía*), *criao* (*criado*), and *hei bisfío*...at times this sound occurs epenthetically, as in *dentrar* (*entrar*) and *dir* (*ir*)" <178>

Lim5. /s/ in an intervocalic position may become [x], as in *nosotros -> nojotros.* <178>

Lim6. Syllable-final /s/ becomes an aspiration or is lost; *bóscalo -> bójcalo* or *bócalo.* <178>

Lim7. The vowel preceding a deleted [s] is opened. <178>

Lim8. There is an apical [(s .)]. <178>

Lim9. Predorsal [(s sub comma)] becomes interdental in some coastal areas, producing *ceceo* as in southern Spain. {may not apply to Lima} <179>

Lim10. "...coastal /r/ and /rr/ are standard Spanish and regularly contrast." {alternatively,} "Zamora Vicente describes the [r] as similar to that of standard Spanish but preceded by *jota*"; *carro -> cajrro, perro -> pejrro.* <179>

Lim11. Syllable-final /r/ may disappear (as in Andalusia and the Caribbean). <179>

Lim12. "Among un-educated people", word-final [r] and [l] in a word stressed on the last syllable, tends to disappear; *peor -> peó, señor -> señó, trabajar -> trabajá, animal -> animá, papel -> papé.* <179>

Lim13. [r] and [l] are in free variation if they occur syllable-finally in an unaccented syllable or are followed by another consonant; *porque sí -> polque si, por mi madre -> pol mi mare* and *alma -> arma, alguno -> arguno* and *el polvo -> er polvo.* <179>

Lim14. The coast and lowlands are *yeista* {that is, orthographic *ll* is pronounced like English *y*} <180>

Lim15. "..../j/, the descendent of /ʎ/, if intervocalic and and in contact with high front /i/ or /e/ often weakens and disappears"; *novillo -> novio, billete -> biete, silla -> sía, capilla -> capía.* <180>

Lim16. "...the phoneme /ʃ/ is unknown..." <180>

Lim17. "...the fricative *jota*...becomes a mere aspiration [h], as in [dího] instead of [díxo], [muhé] instead of [muxér] and [méhico] instead of [méxico]..." <180>

Lim18. /n/ -> /ŋ/ in word final position, regardless of the environment of the following speech sound:
[uŋ óɣro] 'an ogre'
[coŋ létʃe] 'with milk'
[coŋ náðie] 'with no one'
[eŋ mi tjéra] 'in my land' <181>

Lim19. Velar [ŋ] may even occur within a word, syllable-final; [eŋláse] '*enlace*' [kaŋsáðo], and [óŋra] '*honra*'. <181>

Lim20. "...in the common usage of ... coastal Peru, reduction {of consonant clusters} occurs in words such as
conscripto -> [konskríto]
indigno -> [indíno]
significar -> [sinifikár]
victrola -> [vitróla]" <181>

Lim21. "...speakers who do not know the written forms at times insert an erroneous consonant {in consonantal clusters};
séptimo -> [séktimo]
espontáneo -> [ekspontáneo]
aritmética -> [arismética]
insecto -> [insépto]
alumno -> [alógno]
calumnia -> [kalóbnja]"
{note some of the examples are substitutions, not insertions} <182>

Lim22. In popular speech there is a general tendency towards metathesis (transposition of phonemes or syllables);
nadie -> [nájde]
níquel -> [níkle]
Gabriel -> [grabiél]
pusilánime -> [pusilámine]
murciélago -> [mursiéjalo]
polvareda -> [polvaðéra] <182>

Lim23. Pronunciation of consonants is "diffuse" as opposed to "clear." <182>

Lim24. Pronunciation of vowels is "precise", *i.e.*, not with an "uncertain timbre." <182>

Lim25. "Speakers continually confuse /e/ with /i/ and /o/ with /u/"; *trébol* -> [tríbul]. <183>

Lim26. Pitch is higher than in Castile. <184>

Lim27. Intonation is like that of the Caribbean. <184>

Lim28. The tempo of speech is rapid (comparable to that of Andalusia.) <184>

3.3.2.3. Citations from Resnick

Resnick's book is organized in an entirely different way from Cotton and Sharp's. From one point of view, it is more appropriate for automatic dialect identification, as Resnick's intent is to organize language features so that one can assess some selected properties of a sample of speech, consult tables in the book, and thereby infer a likely origin of the speaker. He presents eight main features to assess, each to be assigned a "+" or a "-", as the feature either is, or is not, found in the speech sampled. These eight signs are used as keys to index into his tables. For ease of entry or compilation, he divides the eight features into two sets of

four, which are labeled A1 through A4 and B1 through B4. He also presents "Indices" to relate the sign patterns to geographical locations, and *vice versa*. Applying his location-to-feature table to Cuba and Lima, Peru, we find the sign pattern of features for these locations is as shown in Table 3.2, below.

	Feature							
Dialect	A1	A2	A3	A4	B1	B2	B3	B4
Liman	±	+	−	−	+	−	+	±
Cuban	−	+	+	−	−	−	−	−

Table 3.2: Occurrence patterns for diagnostic features A1-B4 found in Resnick.

The occurrence of features A1 and B4 in Lima are subject to some qualification, hence the ambiguous ± assignment. Feature A1 relates to retention of /s/ in syllable-final and word-final position. This is true for most talkers, but /s/ is sometimes dropped or modified by the young. Feature B4 relates to regular and consistent voicing of vowels, which is true of most speakers but sometimes women de-voice or aspirate them.

Resnick also includes a large number of phonological features in addition to the eight in the A and B quadruples, but these features are treated differently. One must make a binary decision, a "+" or a "-" for each of the features A1-B4, but the binary choice is not forced for the additional features, at least at first glance. These other features are also discussed in "tables", and are assigned letters C through M. A few of these sets of features have binary values but most allow a wider variety of value assignments.

Table 3.3 summarizes the features Resnick presents in his tables C through M. The general subject matter, the number of features in each table, and those ascribed to Liman and Cuban are shown.

Table	Subject	Features	Liman	Cuban
C	Intervocalic /b,d,g/	C1-C6	C2	C1 through C6
D	/b,d,g/ after /l,r,s,y,w/	D1-D13	D1	(none)
E	/tʃ/	E1-E4	E1	E1,E3,E4
F	/f/	F1-F5	F1	F1
G	*ll* and *y*	G1-G14	G2,G5,G9	G2,G3†,G4,G7-G9,G12
H	final /n/	H1-H7	H1,H2	H2,H5,H7
I	/r/ and /l/	I1-I24	I1	I2-I11,I13-I19,I21,I23,I24
J	*rr*	J1-J7	J1,J4	J1,J5,J7
K	word-initial /s/	K1-K5	K1	K2
L	/x/	L1-L5	L4	L1
M	vowel strength	M1-M3	M1	M1

Table 3.3: Occurrence patterns for additional diagnostic features found in Resnick.

These additional features are considered to be present in a dialect if there is a reference claiming so, but that interpretation complicates drawing inferences from the tables considerably. It would not always be clear, for example, when the *lack* of a reference claiming a feature is to be interpreted as indicating the feature is definitely not present in the dialect. Resnick's explanations of these tables usually does make it clear, however. For example, features in his Table C generally address intervocalic /b/, /d/ and /g/. Features C1 and C2 specifically address the treatment of /d/ in the frequently occurring suffix *-ado*. Feature C1 indicates that the /d/ is retained, and feature C2 indicates that it is deleted, *-ado* being pronounced as [áo] or [áw]. Liman is claimed to exhibit feature C2, according to author number 25, and Cuban is claimed to exhibit both features C1 and C2, according to author 61. There are several different inferences which one might draw from this combination of assignments, especially if one allows for the possibility that author 25 may know the /d/ is often retained but simply neglected to mention the fact in his zealous attention to the phenomenon of its deletion. In Resnick's notes on assigning the C features, however, he explains that when both C1 and C2 are assigned, it indicates that both possibilities are heard, *i.e.*, sometimes the /d/ is retained and sometimes it is deleted, so, by implication, one must assume that assignment of C2 and not C1 implies that the /d/ is consistently deleted in Liman *-ado*. Notice that this interpretation amounts to treating the tables as if feature C1 *were* assigned a value by forced choice; its absence is treated as equivalent to a "-" sign, and its presence as a "+" sign, so that the C2-but-not-C1 marking can be interpreted unambiguously.

We suspect that most of this author's clarifications - in the detailed notes he gives for each table - are based on this same strategy of interpretation, *i.e.*, by treating features as binary choices. The only distinction between the "primary" features A1-B4 and the additional features would then be that the author gives explicit directions for assigning a "+" sign

† Although Resnick gives a reference citing property G3 - the existence of a *ll/y* distinction - in Cuban, his remarks question the validity of the observation. We have assumed it is not a property of Cuban, following the consensus of most authorities; *i.e.*, that Cuban is *yeista*.

and for assigning a "-" sign to the former, but only directions for assigning the feature (a virtual "+" sign) for the latter, leaving conditions for non-assignment (a virtual "-" sign) implied but not stated. Ambiguities are certainly encountered in interpreting these tables. For example, nothing is stated for Liman with regard to features in Resnick's Table D. Does that mean the relevant properties have not been examined for Liman, or that none of the thirteen possibilities in that table, which seem exhaustive, was found to apply when it was examined?

Table 3.3 suggests that there are many properties or features in Resnick's tables C through N which can be used to advantage to differentiate Liman and Cuban. Detailed descriptions of these features and differences are given next, so the reader can assess the extent and character of these differences for himself.

3.3.2.3.1. Method of Presentation

As the overview of Resnick's method of organization shows, it is much easier to identify differences between dialects in his book than it is in Cotton and Sharp. In the latter book, it is in effect necessary to tabulate all properties of each dialect, and then extract the differences as a subsequent analytical step. Using Resnick's tables as we have done above, it would not be necessary to tabulate all the features he mentions for Liman and Cuban in order to find the differences between them that his work predicts. However, it is of some interest to determine how consistent these two works are in their descriptions of the two dialects of interest. To determine that, one must exhibit the whole of Resnick's descriptions of Liman and Cuban for comparison with the whole descriptions given by Cotton and Sharp. To that end, we present below separate descriptions of Liman and Cuban as found in Resnick's work, tabulating all the features he gives for each, shared or not, as was done for Cotton and Sharp.

One difference in the presentation of material from Resnick is replacement of the page number reference used in Cotton and Sharp citations by feature designations in Resnick. They appear in angled brackets at the end of each citation. Features A1 through B4 are followed by a "+" or a "-" sign, according to the assignment made in Resnick. For the additional features, C1-M4, no sign is used, again following his conventions. In many cases these additional features group in natural ways, resulting in a single statement covering several of Resnick's features.

For the interested reader, the pages where these feature assignments are to be found are in the "Country Index", which begins on page 249. The Cuban properties are found on pages 339-344, inclusive. Liman properties are of course indexed in the portion of the Country Index devoted to Peru, which spans pages 395 to 402, inclusive. There is a fairly large section explicitly treating Lima, on pages 400-402 inclusive. A few of the features of coastal Peruvian Spanish, mentioned on pages 395 to 339, have remarks attached which imply or state that they apply also to Lima. Only features explicitly identified as applying to Lima were used.

3.3.2.3.2. Comparing Liman Populations

There are unfortuntely some ambiguities with respect to the "Liman" populations described in Cotton and Sharp and in Resnick. The reader may recall that it was necessary with Cotton and Sharp to ascribe the properties of coastal or lowland Peruvian to the Liman

dialect group. This act was justified on the basis of Lipski's observation that Liman natives do in fact exhibit coastal dialect properties, and limiting attention to natives is perhaps the only way to obtain a sensible association of dialect properties with a subset of the linguistically very diverse Lima population.

Ideally then, to ensure comparability of the two authors, one should restrict attention in Resnick in a similar manner, *i.e.*, use only those dialect properties ascribed to native Limans. Unfortunately there is no opportunities to do so, as he doesn't usually make it clear what segment of the Lima population is being described, or anything about the origins of subjects. It would be necessary to go to the original sources he cites to find that information, if it exists at all. As that is impractical, we must accept the noted ambiguity about the populations being characterized by these two authors.

This problem is almost surely representative of what should be expected when examining more than one source for characteristics of any dialect group. It is therfore relevant to any general plan to base automatic dialect discrimination algorithms on the literature. One should expect it to be more the rule than the exception that uncertainties will exist about the exact correspondence of the linguistic populations which are described in dialect literature and which exist in any database.

3.3.2.3.3. Resnick on Cuban

Each citation is again given a prefix (Cub for Cuban and Lim for Liman) and number designation to ease future reference. Our remarks again appear in braces.

Cub1. /s/ is not regularly a sibilant; it may be lost, or replaced by [h] or another sound. <A1->

Cub2. A voiced apical trill is regularly and consistently pronounced for orthographic *rr* between vowels within a word and for word-initial orthographic *r* following a final orthographic vowel of a preceding word in the same breath group. {Most references cited agree. Canfield, No. 27, says no for some parts of the country.} <A2+>

Cub3. Orthographic *g* before *e* or *i*, and orthographic *j* regularly and consistently are pronounced as a weak pharyngeal (glottal) fricative [h] (similar to English [h]), except possibly word-final. <A3+>

Cub4. Orthographic *ll* and *y* are regularly leveled and share a phone or phones in any position. {Cites Ibasescu as disagreeing, but Resnick rejects that position. Lipski concurs.} <A4->

Cub5. /b/ in orthographic *lb* context is sometimes or regularly occlusive rather than fricative. (Disregard examples of *lb* in the orthographic groups *beu* and *bui*.) <B1->

Cub6. Word-final /n/, before a following vowel or a pause, is *not* regularly and consistently a standard [n]. <B2->

Cub7. /l/ and /r/ are not always distinguished and may be leveled or lost. <B3->

Cub8. Vowels sometimes or consistently devoiced or aspirated or lost after or between voiceless consonants or before a pause. <B4-,M1>

Cub9. Suffix *-ado* pronounced variously as [áðo], [áo] or [áw]. <C1,C2>

Cub10. Labiodental [v] is sometimes heard for normally bilabial /b/ and/or /v/. <C3>

Cub11. /b/, /d/ and /g/ are sometimes occlusive between vowels in normal speech at conversational speed. <C4,C5,C6>

Cub12. /tʃ/ is sometimes heard as: i) a voiceless alveopalatal affricate, [tʃ]; ii) a voiceless palatal affricate; and iii) as a voiceless alveopalatal fricative [ʃ]. <E1,E2,E4>

Cub13. /f/ is heard predominantly as [f] and rarely or occasionally as [ɸ]. <F1,F2>

Cub14. *hie-* is regularly pronounced differently from *ye-*. <G4>

Cub15. /y/ is sometimes or regularly heard as [ŷ] initial after a pause, and frequently between vowels within a word. <G7,G8>

Cub16. /y/ is frequently heard as [y]. <G9>

Cub17. Word-final /n/, when followed by a pause or initial vowel of following word, is sometimes velarized, as [ŋ] <H2>

Cub18. Word-final /n/, when followed by a pause or initial vowel of following word, is sometimes lost and the preceding vowel is nasalized. <H5>

Cub19. /n/ within words, before non-velarized consonants, sometimes velarized, as [ŋ]. <H7>

Cub20. Syllable-final /r/ and /l/ are in free variation. <I2,I5,I6,I15,I16>

Cub21. Syllable-final /r/ and /l/ are sometimes reduced, as [l squiggle]. <I7,I17>

Cub22. Word-final /r/ and /l/ are sometimes deleted. <I8,I18>

Cub23. Within words, /r/ and /l/ may be deleted or completely assimilated. <I9,I19>

Cub24. /tr/ sometimes heard as [tr] or [tɹ]. <I3,I4>

Cub25. Syllable-final /r/ and /l/ may be heard as [h], possibly nasalized, or as voiceless nasal [m̥] or [n̥], or as voiceless sibilant.

Cub26. /rr/ occurs as [r̃], [R] and [r], the last between vowels. <J1,J5,J7>

Cub27. Word-initial /s/ is articulated with the tongue's tip pointed downward, towards the lower teeth, and with the tongue grooved to produce a sibilant rather than a slit fricative. <K2>

Cub28. /x/ is generally heard as [h]. <L1>

3.3.2.3.4. Resnick on Liman

Lim1. A voiced apical trill is regularly and consistently pronounced for orthographic *rr* between vowels within a word and for word-initial orthographic *r* following a final orthographic vowel of a preceding word in the same breath group. {Most references cited agree. Canfield, No. 27, says no in some parts of the country.} <A2+>

Lim2. Orthographic *g* before *e* or *i*, and orthographic *j* are not regularly and consistently produced as [h]. <A3->

Lim3. Orthographic *ll* and *y* are regularly leveled and share a phone or phones in any position. {Cites Ibasescu as disagreeing, but Resnick rejects that position.} <A4->

Lim4. /b/ in orthographic *lb* is regularly and consistently the voiced labial fricative [β]. (Disregard examples of *lb* in orthographic groups *beu* and *bui*.) <B1+>

Lim5. Word-final /n/, before a following vowel or a pause is *not* regularly and consistently a standard [n]. <B2->

Lim6. Orthographic *l* and *r* are regularly distinguished. <B3+>

Lim7. For most talkers /s/ is retained in normal form. It is sometimes dropped or modified by the young. <A1±>

Lim8. For most talkers vowels are regularly voiced; women sometimes de-voice or aspirate them. <B4±>

Lim9. /d/ is generally lost in the suffix *-ado*, becoming [áo] or [áw]. <C2>

Lim10. /b/, /d/ and /g/ are all generally standard after /l/, /r/, /s/, /y/ and /w/. <D1>

Lim11. Affricate /tʃ/ is generally produced as [tʃ]. <E1>

Lim12. /f/ is heard predominantly as [f] and rarely or occasionally as [ɸ]. <F1,F2>

Lim13. *ll* and *y* are leveled and are produced as [ʒ and [y]. <G2,G5,G9>

Lim14. Word-final /n/, followed by a pause or a word-initial vowel, is generally velarized to [ŋ] but sometimes normal. <H1,H2>

Lim15. /r/ and /l/ are generally distinguished in all positions. <I1>

Lim16. /rr/ is produced variously as [r̃] and [ɹ]. <J1,J4>

Lim17. Word-initial /s/ is articulated with the tongue tip pointed upward, towards the upper teeth, and the tongue is grooved to produce a sibilant rather than a slit fricative. <K1>

Lim18. /x/ is sometimes or generally produced as [x], before both front and back vowels with no significant fronting. <L4>

3.3.3. Inferences From the Literature

In this Section we use the citations from Resnick and Cotton and Sharp presented above to determine the degree of consistency across sources and to infer the differences each source implies exist between the Cuban and Liman dialect groups. Table 3.4 correlates the citations from both sources with subject matter.

Subject Matter	Source and Citation			
	Cotton and Sharp		Resnick	
	Cuban	Liman	Cuban	Liman
metathesis		Lim22		
intonation; generally		Lim27		
intonation; pace	Cub20	Lim28		
intonation; pitch	Cub21	Lim26		
vowels; generally		Lim24,25	Cub8	Lim8
vowels; in specific contexts	Cub17,18,19	Lim7		
consonants; generally		Lim23		
consonants; in clusters		Lim20,21		
stops; generally	Cub1			
stops; in specific contexts			Cub5,11	Lim4,10
fricatives; generally	Cub2			
fricatives; other than [x], /s/, /f/, [ð]	Cub6	Lim16		
allophones of /x/	Cub7	Lim17	Cub3,28	Lim2,18
/s/; generally	Cub9	Lim8,9	Cub1	Lim7
/s/; intervocalic		Lim5		
/s/; syllable-final	Cub8	Lim6		
/s/; word-initial			Cub27	Lim17
/s/; before a nasal	Cub10			
/f/	Cub4	Lim2,3	Cub13	Lim12
[ð]	Cub5	Lim4	Cub9	Lim9
affricates			Cub12	Lim11
b and *v*	Cub3	Lim1	Cub10	
/r/ and /rr/; generally	Cub14	Lim10		
/r/ and /rr/; leveling		Lim10		
/r/ and /rr/; syllable-final	Cub15	Lim11		
/r/ and /rr/; allophones			Cub2,26	Lim1,16
/r/ and /l/; generally		Lim12	Cub7,23	Lim6,15
/r/ and /l/; syllable-final		Lim13	Cub20,21,22,25	
/r/ and /l/; implosive	Cub12			
lateral /l/	Cub11			
[n] and [ŋ]	Cub16	Lim18,19	Cub17,19	Lim14
/n/ otherwise			Cub6,18	Lim5
y and *ll*; leveling		Lim14	Cub4	Lim3,13
y and *ll*; allophones		Lim14	Cub15,16	Lim13
y and *ll*; existence of [ʎ]	Cub13			
/j/		Lim15		
/tr/			Cub24	
hie- and *ye-*			Cub14	

Table 3.4: Citations from Resnick, and Cotton and Sharp, on properties of Cuban and Liman dialects, arranged by subject matter.

3.3.3.1. Consistency of Subject Matter

One indication of the independence of two authors' treatments of a language's phonology is that they discuss different phonological topics. Conversely, if they discuss only the same topics, they are at least consistent in their views of what the salient properties of the language are, irrespective of how consistently they evaluate those individual topics. Table 3.4 above indicates an unexpected degree of independence of topic matter in the two sources, as detailed below. Table 3.4 was constructed by making a list of all the phonological topics cited by from both Resnick and Cotton and Sharp. In all, 38 topics were found.

For the Cuban dialects, Cotton and Sharp addressed nineteen phonological topics and Resnick addressed seventeen. However, only six topics are addressed by both authors, and incomparable aspects of one of those topics proved, on close examination, to be addressed by these authors. It is their treatments of /s/ generally: Cotton and Sharp's observation about Cuban dialects (Cub9) states that there it has no [z] allophone. Resnick's observation on /s/ in general (Cub1) is that it is not regularly a sibilant, as it may be lost, rendered as [h], *etc.* As Cotton and Sharp tell us that /s/ is *not* one thing and Resnick tells us that it *is* some other things, these observations are not comparable for agreement.

There thus remain five topics on which the authors' descriptions of Cuban could be compared for agreement:

i) allophones of /x/

ii) /f/

iii) [ð]

iv) *b* and *v*

v) [n] and [ŋ]

The commonality of topic matter was nearly the same with respect to Liman dialects. Cotton and Sharp addressed 23 of the 38 topics, and Resnick only fourteen. Of the topics addressed, eight were nominally common topics, but on closer examination different, incomparable aspects of one topic - general properties of vowel production - was presented. Cotton and Sharp note (Lim24,25) that vowels are pronounced "precisely" on the coast, but that speakers "continually confuse /e/ with /i/ and /o/ with /u/". Resnick's observation on Liman vowel production (Lim8) is that for most talkers they are regularly voiced, but that women sometimes devoice or aspirate them. We chose to consider vowel "precision", confusion and voicing all independent properties of speech and hence observations about these aspects of speech are incomparable.

That leaves seven topics on which the authors' observations on Liman Spanish can be compared for agreement:

i) allophones of /x/

ii) /f/; generally

iii) /f/

iv) [ð]

v) /r/ and /l/; generally

vi) [n] and [ŋ]

vii) *y* and *ll*

On average, each author presents data on about eighteen phonological topics, and on average six of the eighteen topics, one third of those presented, are treated in similar enough manner to be compared for agreement. That shows these authors' treatments of these dialects are largely independent, as they chose to emphasize different phonological aspects in both cases. Hazarding to generalize after examining just two authors, the observed independence suggests that the phonological properties of a dialect group which one is likely to find in the literature will depend markedly on the author consulted. It also suggests that a list of distinct phonological properties of a dialect group will at first grow significantly as new sources are examined, and new materials will be continue to be found until many sources have been examined. That is, several sources may have to be examined to even approach a comprehensive compilation of what is reported in the literature.

This lack of overlap was not anticipated. As stated in the introductory paragraph of this Section, since both of the examined books are compendia of basic source material, some of it shared, much greater overlap of topic matter was expected than was observed.

3.3.3.1.1. Sources and Implications of Differing Subject Matter

One factor which can contribute to different authors' describing the same dialect group in different terms, using different phonological properties of each, is an apparent difference in the dialects they are describing, arising from accidental sampling errors. That is, if one or both authors draw conclusions using a non-representative group of informants, it is more likely that they will not notice or record the same phonological properties as salient. In view of the significant differences noted in the topic matter chosen by our authors, this possibility should be considered.

Fortunately, there are two reasons to reject the hypothesis that the noted difference reflects non-representative sampling. One is that there is substantial agreement between the two authors on the topic matter they both address in a similar enough manner to allow assessment of agreement, as will be shown in the next Section. A second reason is that the difference in topic matter can very plausibly be attributed to a different cause: different goals of the two authors.

A factor which was not appreciated before this examination was carried out, but which probably contributes to the difference in phonological topic matter chosen by the two authors, is the great difference in their approaches to their topic. Resnick's intent was to develop a set of phonological features to be applied uniformly to any sample of Latin American Spanish, and a set of indices for finding a characteristic place using those features. A selection criterion on his features was therefore general utility across a wide spectrum of dialects. Cotton and Sharp made no such attempt. They discuss dialectal variations within each of their five major groups entirely independently. It is only reasonable to expect that the features found useful within one of these discussions may be different from the features found useful in another, and that the total number of features used by Cotton and Sharp to discuss several dialect groups would be larger than in Resnick. Conversely, because of his special approach and interest, one should expect much more commonality of topic matter across any pair of dialects in Resnick than in Cotton and Sharp, and the data for the present case realize that expectation.

There is statistical evidence of both effects in Table 3.4. One finds there that Cotton and Sharp address a total of 31 phonological topics in describing Cuban and Liman dialects independently. Of those 31 topics, thirteen, or about 42 percent, appear in both discussions.

Resnick addresses a total of eighteen phonological topics in describing both dialects, fourteen, or 78 percent, of which are used in both discussions.

It follows that the wide difference in topic matter noted for these two authors may be due primarily to Resnick's unusual agenda to develop or find features of wide utility in Latin American Spanish dialects, and this same disparity would not be found among other authors, like Cotton and Sharp, with a simpler, merely descriptive agenda. If so, the inference hazarded above, to the effect that many sources would have to be examined to even approach a comprehensive compilation of what is reported in the literature, is unjustified and may be wrong.

3.3.3.2. Consistency of Language Properties

Table 3.5 summarizes the results of comparing Resnick's and Cotton and Sharps' phonological descriptions of Cuban and Liman dialect groups. It should be recalled that Cotton and Sharp do not describe Liman dialects *per se*, and only point out that "Lima belongs to [the Andean region] geographically but not linguistically", and that we have chosen to interpret their description of coastal Peruvian dialects as applicable to our Liman sample. Resnick's book does describe Liman specifically. It is therefore possible that any disagreement noted between Resnick and Cotton and Sharp may be due to our inappropriately assigning a coastal Peruvian dialect property to Limans.

Subject Matter	Comparison and Result			
	Cuban		Liman	
	Properties Used (C&S - Resnick)	Result	Properties Used (C&S - Resnick)	Result
vowels; generally			L24,25 - L8	Incomparable
allophones of /x/	Cub7 - Cub3,28	Agree	Lim17 - Lim2,18	Disagree
/s/; generally	Cub9 - Cub1	Incomparable	Lim8,9 - Lim7	?
/f/	Cub4 - Cub13	Disagree	Lim2,3 - Lim12	Agree
[ð]	Cub5 - Cub9	Agree	Lim4 - Lim9	Agree
b and *v*	Cub3 - Cub10	Agree(?)		
/r/ and /l/; generally			Lim12 - Lim6,15	?
[n] and [ŋ]	Cub16 - Cub17,19	Agree	Lim18,19 - Lim14	Agree
y and *ll*			Lim14 - Lim3,13	Agree

Table 3.5: Results of comparing two authors' phonological descriptions of Cuban and Liman dialect groups.

3.3.3.2.1. Internal Inconsistencies

As each author has more than one citation addressing some phonological topics, there is a possibility of inconsistency among the citations of each author. Such inconsistencies do in fact occur in Resnick's book, in the form of conflicting data in the basic sources he references. However, Resnick attempts to address and resolve these conflicts as they arise, and we have followed his recommendations in order to minimize the internal inconsistency of what we extracted from his work. An example of this kind of selection can be found in citation Cub4 from Resnick on Cuban, addressing the treatment of *y* and *ll* in that dialect group. He references Ibasescu as claiming there is some evidence of distinguishing these graphemes, *i.e.*, of *lleista*, but questions her claim. We ignored her observations for that reason.

Two instances of apparent inconsistency within an individual author were noted which could not be resolved. One is in Cotton and Sharps' description of realizations of the phonemic /f/, and shows up as a conflict between citations Lim2 and Lim3. The source of both of these citations is the following paragraph in Cotton and Sharp (pg. 178):

"In ..., Peru, ..., /f/, which did not exist in Quechua, is interchangeable with [β] and a number of other sounds such as [x] and [h]. Thus *foto* becomes *boto*; *elefante*

becomes *elebante* ... When /f/ is pronounced, it is it is a bilabial [ϕ] in the popular speech of Peru and Ecuador ..."

First we are told that the phoneme /f/ has a variety of modes of pronunciation, and then that it has one. It *may be* that the second part of the paragraph means that when /f/ is pronounced as an *f*-like sound, rather than as [x] or [h] *etc.*, which are not so "*f*-like", the particular *f*-like sound produced is the bilabial [ϕ]. If so, a much simpler and far clearer presentation would have noted that /f/ may be pronounced as [ϕ], [x], [h] or other sounds. We inadvertently incorporated this inconsistency by interpreting the first and second parts of this difficult paragraph separately, as seemed natural until the incompatibility was noticed.

The other case of internal inconsistency found is in Resnick's descriptions of the allophones of /r/ and /rr/ in both Cuban and Liman dialect groups. The disagreement is related to pronunciation of /rr/ between vowels. In the case of Cuban dialects, this inconsistency appears in our citations from Resnick as incompatibility of Cub2 and Cub26. Examining Resnick's Country Index in detail makes it clear that he is accurately reflecting incompatibility among the sources he used. One source claims only the voiced apical trill is heard in the intervocalic context. Another says it may be that or [R], a "Voiced or voiceless uvular or velar fricative or trill." A third says it may be the voiced apical trill or, "SOMETIMES", the /rr/ may be reduced to the un-trilled /r/ in the intervocalic context. One gets the impression that this is entirely a question of how much attention to give occasional deviations from a generally heard voiced apical trill for intervocalic /rr/ and word-initial /r/.

For the Liman dialects, the argument over occasional lapses from the standard /rr/ appears as incompatibility between citations Lim1 and Lim16. In this case two sources were willing to claim the standard pronunciation was regularly and consistently heard, the basis for Lim1. One dissenting source also recorded a "Voiced alveolar or prepalatal assibilated or *rehilante* fricative, similar to, but not the same as, English [ʒ]in *lesion*"; yet another sound. Again it is clear that the standard /rr/ is much more common than the variants, but Resnick chose to pass on the dissenting author's view.

Since the situation is the same or very similar for both groups of dialects, in that the standard /rr/ dominates, infrequent lapses from that standard need not concern us very much, as it is doubtful they could be used in automatic discrimination of the dialect groups.

3.3.3.2.2. Cases of Incomparability

The cases marked "Incomparable" in Table 3.5 have been discussed above in assessing commonality of topic matter.

3.3.3.2.3. Cases of Disagreement

Two cases of clear disagreement was noted. The two authors offer conflicting data on the manifestation of /f/ in Cuban dialects, and on the manifestation of /x/ in Liman dialects. These are discussed in the two following Secitions.

3.3.3.2.3.1. Disagreement on /f/ in Cuban Dialects

Cotton and Sharp claim "... *f* is the bilabial fricative for the majority of speakers ..."(Cub4), and cite Zamora Vicente as authority for the claim. This observation appears in discussing dialects of the authors' "Caribbean" group, which includes those spoken in Cuba.

Resnick offers a more complex observation on the topic, possibly depending on which part of Cuba is to be covered. His main citation, on page 340 of his Country Index, assigns his property F1 to the entire country and cites his author number 61, with the qualifying comment "(But 'INICIAL ... DESPARECE ALGUNAS VECES')"(exact quote). Delving into his table of features, his F1 is explained as indicating that /f/ is produced as the common [f]; a "voiceless labiodental fricative". The voiceless bilabial fricative pronunciation would be encoded as F2, hence is rejected. The cautionary remark from author 61, who turns out to be Christina Ibasescu, indicates only that initial /f/ sometimes disappears. Overall, this citation seems solidly behind /f/ being produced as labiodental [f]. However, on page 345 the observation is modified slightly. This entry, for both the province and the city of Havana, assigns feature F1 with the remark "PREDOMINANTLY", and feature F2 with the remark "OCCASIONALLY", both claims being attributed to author 14, who turns out to be Lillian Bertot. So we are left with the picture that /f/ is labiodental [f] in the entire country, but may occasionally be heard as the bilabial [ɸ] in the city and province of Havana. This clearly disagrees with Cotton and Sharps' ascription of the bilabial [ɸ] to the majority of speakers.

Pursuing the subject further, one finds that the occasional bilabial fricative was observed by Lillian Bertot among Cuban immigrants living in Miami, Florida less than one month, and appears in her unpublished Florida Atlantic University M.A. thesis. As the speakers in the AFRL-RRS Spanish Dialect database also are Miami residents, perhaps special heed should be given to this occasional phenomenon. Offsetting that, one may question what weight should be given to unpublished observations by a newcomer to the field. But if sources must be questioned at that level, it should also be noted that Resnick (pg. 450) feels compelled to "question the validity of several of Mrs. Ibasescu's statements", which puts the general observation of /f/ as labiodental [f] over the entire country of Cuba in question. It appears that Mrs. Ibasescu also gathered her data among Cuban expatriates, but in Eastern Europe.

3.3.3.2.3.2. Disagreement on /x/ in Liman Dialects

Cotton and Sharp (Lim17) note that *jota*, the Spanish grapheme *j*, which is one of the orthographic forms of the phoneme /x/, is a "mere aspiration [h]"(page 180). Curiously enough, and making the disagreement with Resnick even clearer, Cotton and Sharp include in the three examples of this observation the transformation of orthographic *x*; "... and [méhico] instead of [méxico]."

In stark contrast Resnick notes (Lim2) that *j* is "*not* regularly and consistently produced as [h]"(my italics), and (Lim18) that /x/ is sometimes or generally produced as the harsher [x].

This may be a case of improperly ascribing Cotton and Sharps' observations to the Liman dialects. The context of their remarks about /x/ make it particularly unclear as to what geographical area is being covered. They of course are discussing their "Andean" dialect group, which covers highland Colombia and south thereof, including coastal areas, to

northern Chile. But the exact wording makes one uneasy about applying this remark to coastal Peru, hence, by our convention, to Lima. The wording is:

> "In both the Colombian and Ecuadoran highlands, the fricative *jota* is less scrapy[*sic*] and more relaxed than in Castile ... On the coast it becomes a mere aspiration [h], as in ..."

If the coastal area indicated includes the coast of Peru - as we assumed - then this feature is to be included in our Liman characterization, as it follows the adopted convention of associating Cotton and Sharps' remarks about coastal Peru to Lima. It could be, however, that they mean to include only the Colombian and Ecuadoran coastal areas. Even that interpretation has a difficulty, however, as the Colombian coastal areas are considered by them to belong to the "Caribbean" dialect area, and would not be expected to be discussed in the "Andean" chapter of their book.

The combination of the self-consistency of Resnick's two observations, their strong disagreement with Cotton and Sharps' Lim17, and the unclear geographical reference of the latter, make it reasonable to disregard Lim17 as a characteristic of Liman. We will do so when deriving differences the two authors observations imply about the two dialect groups being studied.

3.3.3.2.3.3. Complexity of Resolving Disagreements

The reader may be surprised by the complexity of the task of extracting phonological properties of a dialect group from the literature, especially if one attempts to resolve or understand apparent inconsistencies among authorities. In the particular instances just described, a contributing cause may be that Cotton and Sharp give the general tendency over their Caribbean group of dialects, whereas Resnick, through Ibasescu and Bertot, summarizes data from a much narrower collection of Cubans. In that sense, the two authors are actually describing different populations. Non-experts may find it very difficult to make sense, or use, of these data. We see no reason to expect that this task would have been any easier for any other pair of Latin American dialects groups.

It should also be noted that extracting phonological properties of Liman and Cuban would have been still more complex had we used primary sources rather than the summarizations provided by the selected authors. These authors no doubt devoted many hours and experienced judgment to reconcile inconsistencies among their sources. This is made very clear in Resnick, as demonstrated by his remarks about Ibasescu's work, quoted above.

Latin American Spanish has at least been studied for many years, so there are many primary sources and several summarizations available. If one had to deal with dialects or dialect groups, in some other language perhaps, which are not so well documented, extracting phonological properties from the literature would at best require an expert in the language and, at worst, be impossible for lack of data.

3.3.3.2.4. Ambiguous Cases

The two cases marked with a "?" in the table are instances of incompatible phonological properties assigned by the two authors, but to groups of speakers which may or may not

belong to the dialect groups being characterized. The first instance of this situation arises in connection with the general treatment of /s/ in the Liman dialect group. Resnick (Lim7) indicates that /s/ is retained in normal form for most talkers, but sometimes dropped or modified by the young. Cotton and Sharp record (Lim8) that "there is an apical s", which may plausibly be regarded as Resnick's "normal" form. However, Cotton and Sharp also note (Lim9) that "In some coastal areas the predorsal s becomes interdental, producing *ceceo*, as in southern Spain." This is clearly not a normal /s/ and would be incompatible with Resnick if it were known that Lima is one of Cotton and Sharps' "some coastal areas."

The other questionable case noted in the table, in connection with the treatment of /r/ and /l/ in the Liman group, hangs on whether or not the speech properties of one particular group of people should be considered in characterizing the Liman dialect group. Cotton and Sharp (Lim12) state that "Among uneducated people on the coast, [r] and [l], when final in a word stressed on the last syllable, tend to disappear."

Uncovering Resnick's position on /r/ and /l/ for Limans is tedious but, when accomplished, leads to an unambiguous result. He gives no data specifically about "uneducated people", and that is a plausible explanation for the discrepancy noted; *i.e.*, it may be that his description applies more generally to the Limans and ignores the uneducated sub-population Cotton and Sharp mention.

For the record, and to illustrate the process necessary to proceed from characteristic place to phonological properties of a dialect using Resnick's organization of his data, we document here how it can be shown that Resnick's claim about /r/ and /l/, for whatever reason, is inconsistent with Cotton and Sharps'.

How /r/ and /l/ are treated in a dialect is the subject of Resnick's third "B" feature, B3. Quoting from his instructions for assigning a sign to this feature,

> "A plus will be assigned to this category for a given speech if these phonemes, which correspond to orthographic *l* and *r* respectively, are regularly and consistently distinguished in all positions as in the standard language and are not leveled."

Being one of his binary features, Resnick's B3 cannot distinguish among possible deviations from the above treatment of *r* and *l*. To capture nuances of such deviations, Resnick also incorporates non-binary features of type I, which also address *r* and *l*. He distinguishes twenty-four different assignments to I, denoted I1 through I24, and I1 is made equivalent to a plus assignment to B3. All other values of I indicate some form of deviation from the *r* and *l* treatment described in the paragraph above.

Turning to Resnick's indices for data about Liman, one finds on page 400 of the Country Index, that Lima is assigned B feature sign pattern codes 69, 70, 197 and 198. These codes signify B feature sign patterns "+-++", "+-+-", "+-++" and "+-+-", respectively(see his page 15). In each case, the third B feature, B3, is seen to be assigned a "+". Also, on page 401 of the same index, one finds that feature I1 has been assigned to Liman, also signifying the "+" value for feature B3. All the data in Resnick are thus seen to indicate that *r* and *l* are "regularly and consistently distinguished *in all positions*"(my italics), which is incompatible with their being leveled by elision, as Cotton and Sharp claim of the uneducated.

3.3.3.2.5. Cases of Questionable Agreement

Many features of dialects can occur with varying degrees of regularity, and different authors may emphasize different aspects of the same behavior. One may say that a particular phoneme is sometimes, or most of the time, or occasionally produced in one way, and another author may say the same phoneme is sometimes produced in another way. To what extent can they be said to agree or disagree? As long as their claims are not categorical, *i.e.*, exclude any alternative, they are consistent, at least under a strict interpretation of their words. However, categorical statements are common when reciting general properties of the speech of a population, even though they are usually expected to be be interpreted non-categorically, just as one might say "The Romans were short" without intending to convey that every individual Roman was short. In this way and others, the wording sometimes raises doubts about how well authors really do agree. Questions of this general type arose quite frequently in comparing Resnick with Cotton and Sharp. In only one case was the situation difficult enough to mention here. It is indicated in the table above by the entry "Agree(?)", with respect to *b* and *v* among Cubans.

Cotton and Sharp make the categorical statement: "There is no phonemic distinction between *b* and *v* ..., and both are realized as [β]."(Cub3) Is this to brook *no* exception? In Resnick one finds that /b/ is, on some occasions, heard as the labiodental [v], which is a little different acoustically from the bilabial [β]. The reference Resnick gives for the evidence of /b/ as [v] is again Lillian Bertot, with the remarks "SCHOOL INFL" and "{ALSO SPORADIC IN RAPID SPEECH}." These remarks suggest that the labiodental [v] for /f/ is a relatively rare event and it becomes a matter of judgment if it occurs often enough to contradict Cotton and Sharps' observation, which was probably not meant to be interpreted strictly categorically. Our choice to regard it as questionable agreement is based in part on the close acoustic similarity of [β] and [v].

3.3.3.2.6. Cases of Agreement

The two authors were found in substantial agreement on four phonological assessments of each dialect group. In the cases where comparison for agreement was clearly feasible, there are seven cases of agreement, one case of questionable agreement and two cases of disagreement. Thus, considerably more agreement than disagreement was found when direct comparison was appropriate.

As the preponderance of agreement is the same for Liman and Cuban, there is no clear evidence that we erred in associating Cotton and Sharps' claims about coastal Peruvian to Liman.

3.3.3.3. Conclusions About Consistency

The major facts which emerged in comparing Resnick with Cotton and Sharp in an effort to evaluate their degree of agreement are these:

a) These two authors seldom address subject matter which is similar enough to allow comparison, and

b) When possible at all, determining if the authors agree often depends on a detailed examination of their work and sources, and sometimes ends being a question of judgment.

c) When comparison is clearly possible, these two authors agreed four times out of five.

d) These conclusions apply equally well to the Cuban and Liman dialect groups.

e) There is no clear evidence that Cotton and Sharps' claims about coastal Peruvian dialects should not be applied to Liman dialects.

3.3.3.4. Implied Differences and Similarities of Cuban and Liman Dialect Groups

Twenty-two of the phonological topics listed in Table 3.4 have entries for both Cuban and Liman dialect groups, from at least one of the authors. Each of these topics potentially reveals a similarity or a difference between the dialect groups, which can be assessed by analyzing the phonological properties cited for each dialect group. Table 3.6 briefly summarizes the result of that analysis. (There are twenty-one entries because two aspects of *y* and *ll* are treated simultaneously.)

Subject Matter	Citations		Common Property	Different Properties
	Cotton & Sharp	Resnick		
Intonation; pace	Cub20-Lim28		pace is rapid.	
Intonation; pitch	Cub21-Lim16		pitch is high.	
vowels; generally	Lim24,25	Cub8-Lim8	vowels sometimes devoiced	
vowels; specific contexts	Cub17,18,19-Lim7		vowels open in similar contexts	
stops; specific contexts		Cub5,11,Lim4,10		stops more occlusive in Cuban
fricatives; except as below	Cub6-Lim16		none noted	
allophones of /x/	Cub7-Lim17	Cub3,28-L2,18		[h] in Cuban, harsher in Liman
/s/; generally	Cub9-Lim8,9	Cub1-Lim7		modified in Cuban, normal in Liman
/s/; syllable-final	Cub8-Lim6		may be aspirated or lost	
/s/; word-initial		Cub27-Lim17		tongue tip direction
/f/	Cub4-Lim2,3	Cub13-Lim12	?	?
[ð]	Cub5-Lim4	Cub9-Lim9	relaxed or lost	
affricates		Cub12-Lim11	/tʃ/ is [tʃ]	
b and *v*	Cub3-Lim1	Cub10	both [β]	
/r/ and /rr/; generally	Cub14-Lim10		as in standard Spanish	
/r/; syllable-final	Cub15-Lim11		may disappear	
/rr/; allophones		Cub2,26-Lim1,16	apical trill dominates	
/r/ and /l/; generally	Lim12	Cub7,23-Lim6,15		leveled or lost in Cuban; distinguished in Liman
[n] velarized to [engma]	Cub16-Lim18,19	Cub17,19-Lim14	frequently	
/n/; otherwise		Cub6,18-Lim5	lost or reduced word-final	
y and *ll*	Lim14	Cub4-Lim3,13	leveled; *yeista* dominates	

Table 3.6: Similarities and differences between Liman and Cuban dialect groups as implied by two sources.

3.3.3.4.1. Differences

Here we discuss the discriminating phonological properties these two authors imply exist between the Liman and Cuban dialect groups.

3.3.3.4.2. /b/ in /lb/ Context

There is apparently a weak, statistical tendency for the stop /b/, and perhaps /d/ and /g/ as well, to be occlusive - as opposed to fricative - more often in Cuban than in Liman. The rendition of /b/ in /lb/ context is the subject of Resnick's feature B1, and it is to be given a "+" if the /b/ is "regularly and consistently" the normal, voiced labial fricative. His two sources for this feature agree that the regular form predominates, but that a more occlusive form occurs too, but infrequently. One source assigns a "-" sign for the entire country, with the note "SOMETIMES". The other source is quoted four times; twice with the "+" and with the note "PREDOMINANT", and twice with the "-" and note "OCCASIONALLY". Unfortunately perhaps, his only sources for observations on these phenomena are Bertot and Ibasescu. Cotton and Sharp are silent on the phenomenon.

3.3.3.4.3. Allophones of /x/

As noted above, Cotton and Sharps' remark to the effect that /x/ becomes a "mere aspiration [h] on the coast" of Peru (Lim17) is inconsistent with Resnick's two comments on the subject (Lim2,18), and perhaps should not be ascribed to Limans anyway. Disregarding that remark, it appears that the /x/ is reliably harsher in the Liman dialects than in the Cuban dialects. In the former it tends to be pronounced as [x] and in the latter as [h].

3.3.3.4.4. /s/: Generally

The case of general pronunciation of /s/ is slightly clouded by Cotton and Sharps' noting that Liman /s/ sometimes becomes interdental (Lim9), while Resnick says most Liman talkers pronounce /s/ in normal fashion (Lim7). This problem was mentioned above as one of the ambiguous cases, denoted by "?" in Table 3.5, where it was noted as another case where it is not clear that Cotton and Sharps' comments are properly ascribed to Limans. Disregarding their Lim17, one notes a clear distinction within Resnick (Cub1,Lim7), to the effect that /s/ is not regularly a sibilant in Cuban dialects, but tends to be normal in Liman dialects.

3.3.3.4.5. /s/: Word-initial

Resnick notes several sources on word-initial /s/ effects, and fortunately they are fairly consistent. The choices are between his features K1 and K2, which are described thus:

> "K1 is assigned if in the articulation of initial /s/, the informant's tongue tip is pointed up towards the upper teeth, alveolar ridge, prepalatal region, or palate, and if the tongue is grooved ... to produce a sibilant rather than a slit fricative",

and

> "K2 is assigned if in the articulation of initial /s/, the informant's tongue tip is pointed down, toward the lower teeth, and if the tongue is grooved to produce a sibilant rather than a slit fricative".

The latter description applies to Cuban and the former to Liman. The notes attached to the sources indicate "APICODENTAL" for Liman and, "DORSOALVEOLAR CONVEX" and "PREDORSAL" for Cuban.

3.3.3.4.6. /r/ and /l/: Generally

Resnick consistently indicates that /r/ and /l/ are regularly distinguished in all positions in Liman dialects (Lim6,15). As mentioned earlier, Cotton and Sharp claim that these two phonemes are sometimes dropped when word-final, among the uneducated in coastal Peruvian dialects, which we find inconsistent with Resnick's data (Lim12). However, in view of the uncertain justifiability of ascribing speech characteristics of the "uneducated" to the middle- and upper-class native Limans of our presumed sample, we opt to disregard Cotton and Sharp in this case. That leaves a clear distinction between Resnick's remarks about Cuban (Cub7,23) and Liman (Lim6,15), to the effect that /r/ and /l/ are regularly distinguished in Liman dialects, but are often leveled or lost in Cuban dialects.

3.3.3.5. Ambiguous Cases

As noted earlier, there is definite disagreement between the two authors on /f/ in Cuban (Cub4 of Cotton and Sharp *vs* Cub13 of Resnick), and within Cotton and Sharp on Liman (their Lim2 *vs* Lim3). This makes it impossible to infer any consistent difference between the two dialect groups for this phonological property.

3.3.3.6. Conclusions About Similarities and Differences

A glance at Table 3.6 shows that the two authors examined here identify many more shared phonological properties than differentiating ones for the Liman and Cuban dialect groups. It is likely that most of the similarities can be related to the fact that they are both forms of the lowland varieties of Spanish, and that more and greater differences would be noted between any combination of a highland and a lowland dialect or group of dialects.

The very brief descriptions of the shared properties mentioned in the table can be elucidated by examining and combining the various citations for each author.

3.3.4. Consultant's Comments

Professor Lipski was consulted frequently over the course of the literature survey. He supplied numerous bibliographies of original sources and summaries, in both English and Spanish, for various dialect groups before and after the Cuban and Liman dialect groups were selected for detailed study. He also supplied specific phonological data about Spanish dialects. As stated earlier, it was on his recommendation that we ascribed properties of coastal Peruvian dialects given by Cotton and Sharp to the natives of Lima. Equally valuable, as it turned out, were his many cogent observations on all aspects of the literature survey and of the dialect identification task in general. His observations have influenced the Spanish part of this report at several points, not always with acknowledgedment. While the present authors retain responsibility for any errors which may appear in the report, we thank to Professor Lipski for his many useful contributions to the effort.

In the remainder of this section, we have attempted to arrange some of Lipski's observations in a way which loosely parallels the organization of foregoing parts of the report. His comments were not completely integrated with earlier parts of the report because the primary purpose of that section is to present what is found in the literature in order to illustrate by example what happens when a non-expert attempts to extract dialect characteristics from it.

His comments, which may be regarded as more authoritative than the literature we reviewed, are so numerous and modify the information found there in such important ways that they would obscure the literature search *per se* had they been completely integrated.

3.3.4.1. General Comments

Lipski's general comments are most interesting, as they reflect his broad knowledge of Latin American Spanish.

3.3.4.1.1. On Selection of Sources

Our choice of sources for detailed examination was based on availability and apparent comprehensiveness of their treatment of Latin American Spanish dialects. It may well not have been the best possible choice. Perhaps a judicious choice of primary sources may have yielded more concrete information, even possibly some statistical and acoustic data. However, the number of primary sources is far to large to have been surveyed in this project, and they pose problems of their own for the non-expert, as noted later. Even in retrospect our selection is at least defensible. However, Lipski had this to say:

> "... the main problem with both Cotton & Sharp and Resnick is that they introduce too fine a level of detail (e.g. quibbling over occasional realizations of /rr/, /f/, /b/, etc.), while giving no empirically verifiable data on the true extent of variation vs. consistency of each feature. Moreover, while Resnick is personally familiar with the Cuban dialect (Cotton and Sharp evidently are not), none of the authors gives evidence of personal familiarity with Peruvian Spanish, thus reducing all their detailed analyses to a battle of second-hand sources, and reducing comparison between their studies to shadow boxing ..."

3.3.4.1.2. On the Diversity of Spanish in Lima and in Cuba

We have already mentioned the diversity of both Peruvian Spanish and the Spanish spoken in Lima, and the difficulty that variability creates for ascribing dialect characteristics to "Liman". Lipski made us aware of the necessity to restrict attention to native Limans of a limited range of socio-economic conditions in order to deal with even an approximately dialectally homogeneous group.

> "As for Lima, it is obviously a coastal city. However, it is unique among major mainland Latin American cities in having been both the seat of a major colonial division (a Viceroyalty) and a major port of trade. All the other administrative capitals were located in inland regions (Mexico City, Guatemala, Quito, Bogota, etc.). This means that Lima was exposed both to consonant-strong dialects from central/northern Spain (the speech of government and ecclesiastical officials) and the consonant-weak dialects of southern Spain and the Canary Islands (which formed the *lingua franca* of all Latin American coastal areas and ports). As a result, Lima Spanish has much more **vertical** stratification than the rest of Peru. This means that social class differences as regards pronunciation of final consonants are quite large. An educated upper-class Lima native in a semi-formal interview situation may retain a high number of final /s/, and might actually sound like a Colombian

or Mexican for a while. An illiterate working class subject from the same city might reduce final consonants almost as much as a Cuban. As if this weren't enough, Lima has been the scene of massive emigration from all over the country, and a majority of the city's current population (concentrated in the poorer neighborhoods) is not native to the city, or even to the coast."

With respect to the variability of the Spanish of Cuba, Lipski finds it limited and hazards a guess that it might be so small as to defeat any attempt to differentiate among its varieties by automatic means:

> "... from the phonetic point of view, Cuban Spanish is quite homogeneous, and such geographical and social variation as might exist could probably not be feasibly extracted by automatic speech-recognition techniques."

3.3.4.1.3. On Choosing Subjects

Although there is little opportunity to choose subjects in the data collection effort, some selection *among* subjects must be made to form dialectally homogeneous samples for algorithm development and, even more importantly, for algorithm testing. Lipski's comments on subject selection bear on that issue, and also on what information might be solicited from the subjects to help classify their dialect. Class and native place are the crucial issues for Limans.

> " If you have the luxury of pre-selecting your informants, I'd suggest taking ... some upper-middle class neighborhood of Lima as your coastal dialect (in my experience, most expatriate professional Peruvians in this country tend to fit the latter group). Better yet, just get them all from, say, Miraflores (a nice Lima neighborhood where everybody goes to the same private schools.) If you have to take what you can get, just ask the simple question (in Spanish): "are you from the highlands (*la Sierra*) or the coast (*la Costa*)?". All Peruvians will instantly know what this means (and what it implies), and this quick and dirty self-identification will correlate very well with observed phonetic differences (unless your subjects are lying, or have spent many years living in another region). In other words, the division is as much cultural as geographical, but the linguistic correlations are very close. If the speakers are of comparable social class and educational level, this will produce a reasonably homogeneous sample, for purposes of gross phonetic detail. This method works surprisingly well."

Lipski also has a suggestion about how to separate Colombian subjects:

> "... the same question will work for Colombians (who also classify themselves on a two-item scale), replacing *la Sierra* by *el Interior* (the interior). The fact that Colombia has two coasts is irrelevant; Colombians assume that anyone living on either coast speaks the same, and from a phonetic point of view, this isn't entirely false. There are plenty of subtle differences among coastal and inland dialects, but they are more variable and less susceptible to automatic identification."

And on Argentinian, he says

" In Argentina, an increasingly large area of the country is adopting Buenos Aires pronunciation. Unless you get speakers from along the Paraguayan or Bolivian borders, chances are you will get phonetic variants that coincide in large measure with Buenos Aires."

The last comment suggests the possibility that the Argentinian subjects in the AFRL-RRS Spanish dialect database may be dialectally quite homogeneous. (Unfortunately, Buenos Aires is another lowland dialect, like the Cuban and Liman groups.)

3.3.4.1.4. Exceptions from the Literature

As Lipski's general comment about the chosen sources might suggest, he disagrees with several of the phonological properties claimed therein for Cuban and Liman. Specific points on which he differs are listed below.

Cotton and Sharp claim (Cub19) that in Cuban Spanish deletion of syllable-final /s/ is accompanied by an opening of the preceding vowel, but Lipski notes

"... with respect to Cotton & Sharp on compensatory vowel opening coupled with loss of final /s/ in Cuban Spanish. This phenomenon is well-documented in eastern Andalusian Spanish, and many have felt that it also occurs in Caribbean dialects. However, numerous careful studies (... almost none of which are cited by C & S) have shown conclusively that this does **not** occur in Cuban or other Latin American dialects."

There is also a problem with an example of /f/ pronunciation in Liman Spanish according to Cotton and Sharp (Lim3). This citation notes that /f/ is pronounced as a bilabial, but the first example, supposedly of this phenomenon, is pronunciation of *familia* as if it were spelled *juamilia*. Lipski notes

"Peruvian Spanish /f/ may indeed be bilabial (as it is throughout most of Latin America), but the pronunciation of /f/ as [hw] as in *familia* pronounced as *juamilia* is definitely **not** a trait of any lowland area. This pronunciation occurs only in contact with indigenous languages which have no bilabial fricative; they turn a single segment into a sequence of [h] plus semivowel [w]; in effect the first element of the sequence carries the "fricative" information, while the second part carries the "labial" information. No native of Lima, however humble the origins, uses this pronunciation."

One of the great difficulties a non-expert will have in interpreting primary source dialect literature is that writers of that material tend to describe everything that happens in the language of an area, irrespective of any value the observation may have as a dialect differentiator. Lipski finds a typical case of this kind in Cotton and Sharp's claim of lost consonants in consonant clusters(Lim20):

"... reduction of consonant clusters and hypercorrect insertion of syllable-final consonants is found colloquially in **all** Spanish dialects, and has no regional correlation. It's just that some monographs on regional dialects have described all popular speech phenomena as though they were somehow exclusive to the dialect being described. This creates a very misleading impression, since, *e.g.*

pronunciation of *septimo* as *sektimo* can be heard from Bilbao to Buenos Aires."

He also rejects Cotton and Sharp's observation of the interdental /s/ as having any significance (Lim9):

> "Interdental /s/ is occasionally heard almost everywhere among illiterate rural speakers, but is not characteristic of any geographical region outside of western Andalusia. This would not be a useful feature for any area of Peru."

The only exception Lipski noted with the phonological claims of Resnick is to clarify the latter's statement about /s/ not regularly a sibilant. Resnick assigns "-" to his feature A1 whenever this happens, irrespective of where or how the non-sibilant phenomena may occur (Cub1). Lipski notes that, for Cuban, it happens when the /s/ is syllable- or word-final, but that /s/ is regularly a sibilant in syllable-initial contexts in both Peruvian and Cuban Spanish.

3.3.4.1.5. Reliable Phonological Characteristics of Cuban and Liman Dialects

Lipski has a short list of reliable phonological characteristics for the two dialect groups of interest. They differ from Cotton and Sharp's lists primarily in that they are shorter. They are both shorter and different from Resnick's, due to the latter's use of special features selected for a different, more general purpose. Lipski suggests the following for Cuban:

> Cub1. Preconsonantal and word-final prevocalic /s/ is almost always aspirated [h]; phrase-final /s/ most often disappears, except in formal, highly monitored speech.
>
> Cub2. Phrase-final and word-final prevocalic /n/ is velar.
>
> Cub3. Intervocalic /y/ is an approximant or palatal fricative, perceptibly stronger than in Peru. In never disappears.
>
> Cub4. The posterior fricative /x/ is usually a weak aspiration [h].
>
> Cub5. Among the lower classes, and with considerable geographic variation, there is some neutralization of preconsonantal /l/ and /r/. The end results are, however, too variable to be useful for automatic identification purposes.

and for Liman

> Lim1. Rather frequent velarization of phrase-final /n/, and word-final prevocalic /n/.
>
> Lim2. Aspiration of preconsonantal /s/ to [h], while word- final prevocalic /s/ (as in *los amigos*) more frequently is retained as [s], particularly among the more educated classes. Phrase-finally, [s] also predominates in more careful speech, while in lower class or highly colloquial speech, loss of phrase-final /s/ is common.
>
> Lim3. Intervocalic /y/ is weak, has almost no fricative/approximant characteristics, and may disappear in contact with front vowels, as in *silla* (could be pronounced [sia]), *gallina* [gaina], etc.).
>
> Lim4. Trill /rr/ is also given a multiple trill pronunciation, never realized as a groove fricative.
>
> Lim5. All vowels are fully pronounced; there is no unstressed vowel reduction.
>
> Lim6. The posterior fricative /x/ is usually a weak aspiration [h].

3.3.4.1.6. On Differences Between Cuban and Liman

The lists of salient Cuban and Liman dialect group phonological characteristics are unfortunately very similar. The scarcity of differentiating characteristics inferred from the literature is thus supported by Lipski's phonological details and by his comments on what differences should be expected of these two groups, given below.

> "The differences between coastal Peruvian Spanish and Cuban Spanish are usually only differences of degree; between the lowest classes of Lima Spanish and most Cuban varieties, there is considerable overlap for most features, and accurate phonetic distinguishing may not be possible."

> "Between Lima and Cuban Spanish, there are almost no good binary oppositions. Aspiration of /s/ is much more frequent in Cuban Spanish than in middle-class Lima Spanish, but in lower working class Lima/Callao Spanish, rates may be similar. Also, velarization of /n/ occurs at similar rates in both dialects. Intervocalic /y/ is stronger in Cuba than in Peru, but it is not clear to me (after having looked at spectrograms) that this differences is marked enough nor consistent enough to lend itself to automatic identification. Coastal Peruvians do not neutralize preconsonantal /l/ and /r/ with any regularity (although phrase-finally, both may disappear in vernacular speech), but once more, these are highly variable phenomena in Cuban Spanish, and are not prominent in middle-class speech. I am therefore initially pessimistic about finding clear binary differentiators between Lima and Cuban Spanish."

> "To summarize, highland vs. lowland Peruvian dialects could conceivably be differentiated by binary criteria. Coastal Peruvian vs. Cuban Spanish are differentiated by scalar values of the same variables, with considerable overlap among the lower sociolects of Lima Spanish and general Cuban Spanish."

> "Automatically distinguishing between Cuban and coastal Peruvian gets harder as one descends the social scale in Peru and/or gets into small coastal towns (whose residents, however, are not well represented in the United States). Both groups aspirate syllable-final /s/ (more so in Cuba, but at comparable levels among lower-class coastal Peruvians), both groups velarize final /n/, both groups use a weak aspiration [h] for the posterior fricative /x/; /y/ is somewhat stronger in Peru than in Cuba, but I doubt whether this could be detected accurately."

In particular, he is not at all enthusiastic about the few differences we did find by analyzing Resnick and Cotton and Sharp:

> "... Similarly, in my experience, /x/ is not "different" enough between the two dialects to be systematically distinguished by either human ears or machines. As for /s/, I really find no justification for the notion that /s/ in general is "modified in Cuban," nor that "tongue tip direction" is in any way different between the two dialects. I know that you derived this from your bibliographical survey, but in

practice this claim just doesn't hold up. Once more, in citing Resnick in section 3.5.4.4. (p. 42), you note that '.../s/ is not regularly a sibilant in Cuban dialects ...' Again, this is true only in syllable- and word-final contexts. In other positions, Cuban and Liman /s/ are identical."

Yet he is not quite willing to say automatic differentiation of Cuban and Liman is impossible. He holds out hope for two kinds of features. First, he is open to a slight possibility that intonational differences might somehow be used, if appropriate methods of characterizing intonation could be found;

> "There are definite intonational differences between Cuban and Lima Spanish, but to date there exists no consensus among linguists, nor any empirical descriptive framework for accurately characterizing differences among dialects."

And second, a little hope for the /s/ distinctions:

> "The real issue with respect to /s/ is the much greater frequency of aspiration/deletion of syllable-final /s/ among all social classes in Cuba, as opposed to the higher level of retention of sibilant syllable-final [s] among more educated Limans. This difference cannot be reduced to a binary tabular feature, and once lower-class Lima/Callao speakers are taken into account, the difference may disappear altogether."

Finally, while admitting the prospect of separating these two dialects appears dismal from the point of view of conventional dialectology, he can't deny there's some hope for automatic separation.

> " Which leaves, basically, nothing in the way of systematic differences between the two dialects, assuming comparable sociolects. And yet, I'm willing to bet that if your sample contains educated Lima speech (as opposed to Cuban Spanish of almost any kind), some sort of automatic recognition with a respectable level of accuracy could be possible, based on variability of /s/ and maybe other collateral factors."

Perhaps its because he can so easily hear differences between these two dialect groups, in spite of the difficulty of capturing those differences in phonological terms.

3.4. Arabic Dialectology Literature Survey

Arabic dialectology is a difficult arena for a non-Arabic speaking researcher. No generally-recognized compendium of dialect traits exists in book form and journal articles are relatively hard to obtain and harder yet to synopsize. To cut through these problems, we engaged the services of a world renowned Arabic dialectologist, Dr. Alan Kaye of California State University, Fullerton. He agreed to provide us with a pre-publication copy of a chapter in a book he is writing on Arabic dialects. The chapter is a survey of dialect differences.

3.4.1. An overview of Arabic Dialects

In many ways, the diversity of Arabic dialects is similar to the diversity of Romance languages. Many Arabic speakers exhibit *diglossia*, using their native vernacular and Modern Standard Arabic (MSA) as the situation demands. Today's spoken dialects are the current

descendants from the vernaculars spoken by the Arabic invaders during the era of Islamic expansion. MSA is an artificial language, in that it is no one native language. It is an accepted standard among educated Arabic speakers throughout the Arab world. The pronunciation and lexical choice in MSA can be modified significantly by admixture with the local vernacular. The mixing is a socio-linguistic phenomenon in that the amount of vernacular influence depends upon the social situation. The use of two languages such as the vernacular and MSA is termed diglossia. Arabic diglossia parallels the diglossia in Europe when the local vernaculars co-existed with Latin. How mutually intelligible the vernacular and MSA are is an empirical question which has not been addressed in the literature, as far as we can determine. The Arabic spoken in Uzbekistan and the language erroneously called Maltese Arabic on the island of Malta are clearly different languages, as defined by the criterion of mutual intelligibility. There may also be a language difference between the vernacular Arabic dialects spoken east of Libya and those spoken from Libya west. Thus, Morrocan Arabic and Egyptian Arabic are really different languages despite many similarities in grammar, phonology, and lexicon.

The extent of the details in Arabic dialectology is vast. Some of the more divergent dialects are away from the central core of Arabic speakers, e.g., in the Sahel from Somalia to West Africa, and in the Turkic speaking crescent which stretches from Central Asia to Anatolia. Capturing all this diversity is well beyond the amount of time scheduled for this effort, yet it is precisely these areas, Somalia, Chechnya, Afghanistan, Nigeria, etc. in which the United States has shown evidence of strategic interest.

While there are many other dialects in Asia, and in Saharan and sub-Saharan Africa, within the central, vernacular dialects, Kaye asserts that there are the following major dialects: Cairene spoken in Cairo and Lower Egypt, Syrian-Lebanese spoken in the Levant, Iraqi spoken in Mesopotamian, dialects of the Arabian peninsula, and North African. The major dialects are typically subdivided into East and West. Table 3.7 groups these dialects and shows their constituent populations.

Group	Dialect	Areas
East	Peninsular	Saudi Arabia, Yemen*, Kuwait, Oman, the United Arab Emirates
	Mesopotamia	Iraq
	Syrian-Lebanese	Syria, Jordan, Israel, Lebanon, Palestine
	Egyptian	Lower Egypt
	Egyptian, Sudani Chadian,Cameroonian, Nigerian Afghani Uzbeki	Upper Egypt, Sudan Chad, Cameroun, Nigeria Afghanistan Uzbekistan
West	North African	Libya, Tunisia, Morocco, Mauritania
	Maltese †	Malta

Table 3.7: The major divisions of Arabic dialects

* Many rural Yemenite dialects are archaic and preserve features of Classical Arabic

† Maltese Arabic has some Eastern linguistic traits, such as the glottal stop reflex of Classical Arabic /qaf/.

The two principal phonetic/phonemic differences between the East and West dialects are that the Western dialects have, generally speaking, final stress and have lost many short vowels and reduced many long vowels. Eastern dialects preserve have penultimate or ante-penultimate stress and tend to preserve the Classical vowels. There are also morphological, syntactic and lexical differences as well, which will not be covered in this summary.

There are also differences between the sedentary and the nomadic dialects. These dialects are also referred to as urban and bedouin, respectively. The dialect differences pre-date the emergence of Arabic from the Arabian peninsula and complicate the description of Arabic from a geopolitical perspective. The basic differences between sedentary and nomadic Arabic is that the nomadic dialects tend to voice and front the Old Arabic /q/, a voiceless uvular stop, to either /ɢ/ (voiced), /dʒ ,(voiced, fronted, and affricated/ or /dz/ (voiced, fronted, and assibilated). The syllable structure is also different.

There are also communal dialects that cut across the areal divisions. Religious dialects exist for the three major religions and there are some differences that arise from the subsequent schisms of these major groups. These communal differences can be minor lexical and prosodic differences, moderately consistent phonetic differences, or major dialectal

cleavages. The latter is exemplified by the Jewish and Christian communities in Baghdad which speak a sedentary form of Arabic, while the Muslims speak a dialect based more on a Bedouin prototype.

3.4.2. Dialect Selection

It is necessary to delimit the dialects of Arabic we wish to account for. There are too many very different and distinct dialects of Arabic to treat within this document. We must also limit ourselves to certain social, ethnic and religious groups since each of these factors introduces new and sometimes seemingly non-systematic changes. That is, what characterizes a Christian in Baghdad may not necessarily characterize Christians in Algiers. The effect of one's ancestors being Bedouin in Baghdad may not have the same effect in Damascus.

Initially, we sought guidance from the Government on delimitation criteria or even selection. The experts at AIA, who might have been in a position to help us, were transferred before they could give recommendations. Therefore, P. Benson in consultation with Prof. Kaye made a selection of the five regional dialects from the central or core Middle Eastern area. The set of dialects may have some operational significance, given world events. These dialects are also distinct one from the other, and there is relative homogeneity within each dialect area. The dialects are given in Table 3.8

- Iraqi
- Gulf (from Kuwait to Oman, including Saudi Arabia)
- the Levant (including Syria, Palestine, Lebanon and Jordan)
- Lower Egypt
- Morocco

Table 3.8: A preliminary breakdown of Arabic dialects

Homogeneity within these dialects can only be achieved by limiting the population over which the dialect is defined to be city dwellers who are typical of the culture. This limitation excludes the rural and bedouin population to a great degree, as well as the ethnic and religious minorities. There are dialect differences that span political dialect boundaries and aggregations, such as Iraq or the Gulf States. The bedouin, for example, often share some common linguistic as well as cultural features across the political dialect boundaries.

This delimitation is not meant to circumscribe the set of dialects for which automatic dialect identification is possible. To the contrary, automatic dialect identification needs to be able to discriminate arbitrary dialects when operational. Rather the delimitation is meant to circumscribe the set of Arabic dialects that might be used for the initial development of algorithms. It is hoped that an automatic dialect identification scheme for these dialects might have some immediate use. In the long run, it will be necessary to instruct and/or train the algorithms with materials bearing directly on the dialect distinctions called for operationally.

3.4.3. Finer-Grained Analysis of the Selected Dialects

In the chapter provided by Kaye a finer grained analysis of the consonants in the various dialects can be found. This analysis is not nearly as detailed as the one given for Cuban and Liman Spanish. In part, this difference in detail is because the literature is lacking; in part this difference is because the source we have used, Prof. Kaye, can characterize the major and consistent differences of the particular dialects we have selected. The detailed variation found in the Latin American Spanish literature may reflect some class and socio-economic differences which we directed Prof. Kaye to ignore.

In the following paragraphs the major characteristics of the five dialects are tabulated. Almost all Arabic dialectology discusses the differences between dialects in terms of how the current vernacular reflects or preserves phonemes from Classical Arabic. The major phonemes that change are /q/,/k/,/dʒ/, interdentals and vowels.

Gulf Arabic is meant to include the Gulf States from Kuwait up to but not including Oman and Saudi Arabia. This dialect area is somewhat smaller than the comparable one listed in Table 3.8 because inclusion of the southern dialects in the larger Gulf region increases the heterogeneity unacceptably. In Gulf Arabic, we find

[j] for [dʒ],
[ɢ] for [q],
[k] and [tʃ] or [ts] for [k] { affrication is conditioned by front vowels}
[θ ð ð̣] remain the same
5 long vowels [a: e: i: o: u:]
3 short vowels [i a u]

Iraqi Arabic includes the Muslim Arabs of Iraq and excludes the Kurds, Christians and Jews.

[dʒ] for [dʒ]
[q] and [ɢ] for [q]
[k] and [tʃ] for [k]
[θ ð ð̣.] remain the same
5 long vowels [a: e: i: o: u:]
3 short vowels [i a u]

Levantine includes Syria, Lebanon, Palestine and Jordan. Damascus seems to be somewhat different.

[dʒ] for [dʒ], but [ʒ] in Damascus
[q] for [q], ? = glottal stop
[k] for [k]
[θ ð ð̣.] replaced with [t d ḍ], except they remain in
the speech of rural and bedouin speakers

5 long vowels [a: e: i: o: u:]
short [i] [u] become ə
short [a] fronts to [æ], or even [e] or [i].

Lower Egypt includes the Nile delta, Cairo and Alexandria

[ɢ] for [dʒ]
[q] for [q]
[k] for [k]
[θ ð ð̣.] replaced with [t d ḍ]
5 long vowels [a: e: i: o: u:]
3 short vowels [i a u]
short [a] fronts to [æ]

Morroco is meant to include metropolitan Morroco and not the Berber-substrate hinterlands.

[ʒ] for [dʒ] but dissimilation in roots containing sibilants
[q] for [q]
[k] for [k]
interdentals replaced with homorganic stops
no long/short distinction, only 6 vowels (3 stable and 3 variable)
vowel reduction to ə and consonant cluster formation

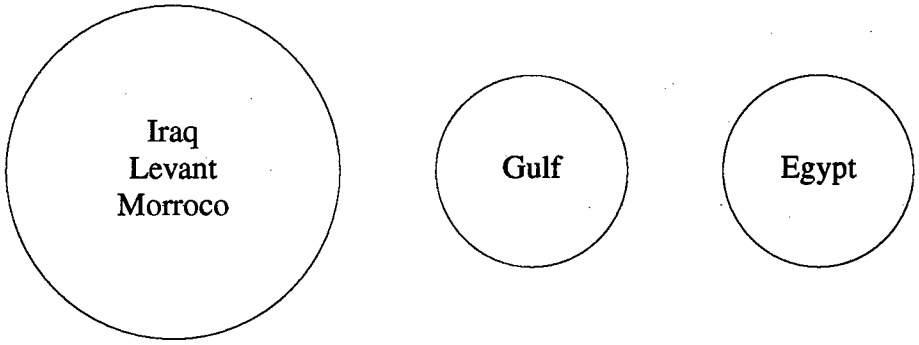
3.4.3.1. Understanding what distinguishes dialects of Arabic

Given that different dialects use different reflexes of the Classical phonemic inventory, it might be possible to separate them based on those differences. The dialects are not wholly different in the reflexes used and the similarities do not lend themselves to a tree structure. For example, Iraq and Morroco share a common reflex of /dʒ/, while they differ on vowels. Iraq and the Gulf share in their treatment of vowels and differ on /dʒ/. In the sections below, the common reflexes are gathered together by regional dialect. Then a pictorial representation of the groupings that share common features are shown.

Alan Kaye has pointed out that the differences discussed above hold true for the pure vernacular dialects. In conversation that moves up and down the continuum from vernacular to MSA, pronunciation may change, e.g., /k/ could easily shift to a /q/.

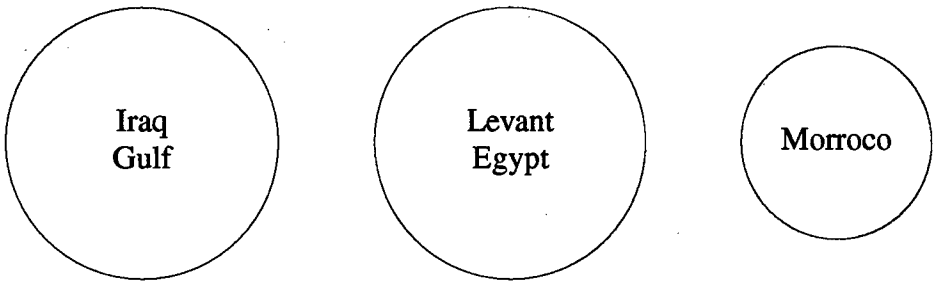
3.4.3.1.1. Which Dialects share common reflexes of the /dʒ/

1. Iraqi, Damascus Levantine and Morroco /dʒ/
2. Gulf /j/
3. Lower Egypt /g/



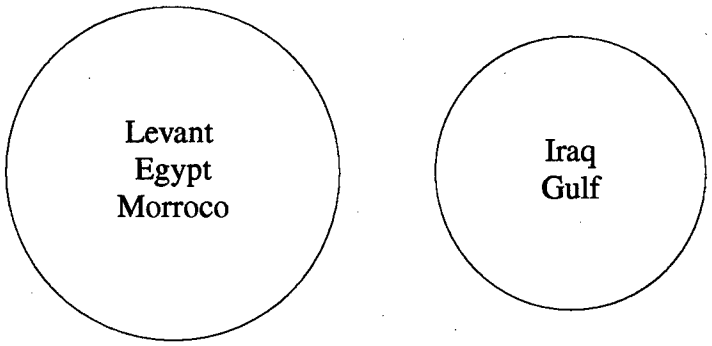
3.4.3.1.2. Which Dialects share common reflexes of the /q/

1. Iraqi /g/ Gulf (although some bedouin use /ɢ/)
2. Morroco /q/
3. Levantine and Lower Egypt have /q/ (glottal stop)



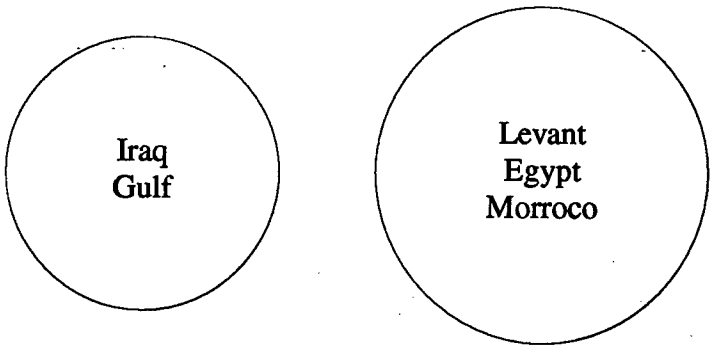
3.4.3.1.3. Which Dialects share common reflexes of the /k/

1. Levantine, Lower Egypt and Morroco retain /k/
2. Iraq and Gulf make /k/ an affricate; in the Gulf affrication happens only before front vowels



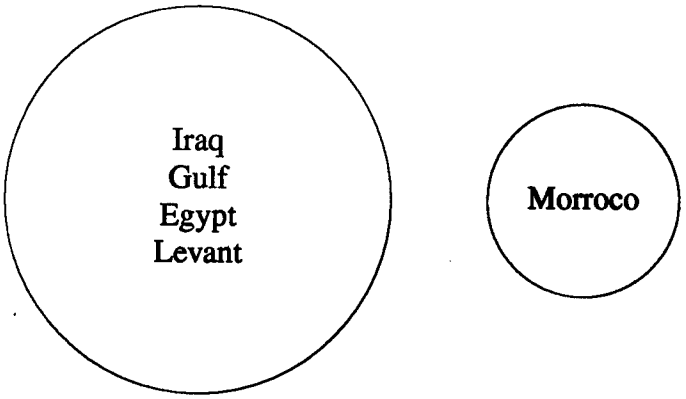
3.4.3.1.4. Which Dialects share common reflexes of the interdentals

1. Iraqi and Gulf retain the interdentals
2. Levantine, Lower Egypt and Morroco replace them with homorganic stops



3.4.3.1.5. Which Dialects share common reflexes of the long vowels

1. Gulf, Iraq, Levantine and Lower Egypt retain long vowels
2. Morroco has no long short distinction, only 6 vowels (3 stable and 3 variable) and vowel reduction[Harr62]



3.4.3.1.6. Which Dialects share common reflexes of the short vowels

1. Gulf retains three short vowels /i a u/[Char64]
2. Damascene Levantine /i u/ becomes ə, retains short /a/

2a. The rest of the Levant retains the vowels

3. Lower Egypt has shifted /a/ for /i/

4 Iraq has developed an /o/

5. Morroco has a new vowel structure

3.5. Conclusions of the Literature Survey

In order to exploit dialect differences to separate dialects automatically, the differences must be reliably present in the material to be sorted and the differences must be amenable to automatic detection. The differences that we have abstracted from the dialectological literature are not guaranteed to fulfill either of these conditions.

First, the dialect differences described in the dialectology literature are those found by human dialectologists. The procedures that they use are fundamentally different from the procedures used by signal processing algorithms. Dialectologists have different goals. For example, they often are searching for conservative versions of the dialect as a kind of window on the past, allowing them to find evidence for the historical version of the language. In this effort, they seek out the oldest and often least typical speakers from whom to acquire data. The data acquired in this manner does not bear upon reliable characteristics of the local dialect that can be found commonly among the speech of the people.

The dialect work from Spanish and from Arabic is often couched in terms of a comparison with the standard or the Classical language. That is, the properties that define a dialect are the differences from the Standard. If we are trying to identify a dialect we would need to determine not what is in the signal but how it differs from the standard. Where these differences are word or context specific, detection of dialect differences requires that we detect the word or context and then judge the degree and direction of difference from the Standard.

Dialectologists, in general, do not acquire information about the relative frequency of dialectal phenomena. Although this practice may be changing, the core of the dialectology material pre-dates these changes and the material about Spanish and Arabic tends to be from an earlier period. Thus, it is often not known whether some feature that describes a particular dialect is a common feature or an infrequently heard one. An example might be a widespread stylistic variation among women in many languages to de-voice words. This de-voicing is never frequent enough to be useful in distinguishing dialects but it is a fact about the dialect. Thus, it is incorporated in the description, but will not assist the automatic procedures.

Without quantitative information concerning the relative occurrence of dialect features, it is not possible to know whether they can be exploited. One could try all features that all dialectologists agree upon and cast out those which do not serve.

Reliability and consistency are required for any non-expert to make sense of a dialect difference cited in the literature. A major effort was mounted in this work to determine whether the dialectologists' reports were consistent both internally and with one another. In Spanish, it was found that they were not. This finding is colored to some degree by the fact that there is neither an accepted classification scheme nor an accepted descriptive framework. Even the framework of the IPA is not adhered to and the record of dialect characteristics are often not comparable.

The value of IPA transcriptions can themselves be faulted. Transcriptions are always going to be impressionistic. Distinctions that are clear to one listener may be inaudible or non-existent to another. The human ear is a marvelous signal processing device and it is approximated in modern digital signal processing programs only in the crudest of fashions.

The relation then between a consistent IPA description of a dialect and its signal processing alternate may be weak. Knowing that one dialect always uses an [ae] where another uses an [a] may not be of direct value.

Perhaps, the most common problem in exploiting the dialectologist's experience in the automatic signal processing domain might be referred to as the missing example problem. When a description is phrased in terms of the differences between a dialect and the Standard, the difference often turns out to be that the dialect no longer retains, say, an [x] word finally. There is no positive way one can separate two dialects that differ in this fashion. If one looks for the [x] to find the dialect that retains the [x], there is a non-zero probability that the particular speech segment from the dialect that retains the [x] will not contain an [x]. This problem is exacerbated when the segments are short. Short segments diminish the likelihood that any given phoneme will appear.

With rare exception, there is no acoustic phonetic research distinguishing dialects of Latin American Spanish in which we have interest. Nor are there acoustic phonetic papers on separating Arabic dialects. In some sense, the work we are doing breaks new ground on many fronts. If we are to use published dialectology, we must reply upon the various transcriptions to give some indication as to what the acoustics are. These descriptions, especially as revealed in the Spanish study, are often impressionistic. That is, the phonetic comments will label a segment as "slightly more open" or with "rougher aspiration." These descriptions refer only the impression that the sounds left upon the hearer and are not usable in signal processing.

4. Baseline System Description

Following is a very brief description of ITTI's baseline algorithm for DID, included here as an aid in understanding the remainder of this report. Much more detailed descriptions of the baseline algorithm and its component parts are available in the technical literature

This algorithm was developed over a period of four years for Language Identification, and represents the state of the art in that discipline at the start of this contract. It is a natural starting point for development of a DID capability because many of the same features of language serve to distinguish dialects as serve to distinguish languages; in fact, some linguistic groups are considered distinct dialects by some some authorities and distinct languages by other authorities.

4.1. Baseline System Block Diagram

The language identification algorithm consists of the components shown in Figure 4.1. The main elements include:

(a) Preprocessing: to remove bias and to condition the speech signal into a specific dynamic range.

(b) Parameterization and normalization: to obtain acoustic-phonetic parameters at a 20 msec frame rate. This process includes blind deconvolution and obtaining relative amplitudes.

(c) Trained artificial neural network (ANN) marking: to mark each specific segmental phonetic event, such as a syllabic nucleus.

(d) Syllabic feature extraction: to encode the syllabic on-set, coda, or intra-syllable phonetic events at each marking. These features constitute the principal (and original) language sample representation. The feature space has been extended to include syllabic prosodic features which are encoded from amplitude contours, pitch contours and timing information. An eigenvector reduction process is also used for dimensionality reduction to improve performance and reduce processing time. The individual feature extraction processes can be selectively activated or deactivated.

(e) Matching processes: to measure message distances and evaluate a number of possible scoring schemes, to provide scores for identifying languages. The user can also specify a particular scoring technique to obtain individual test sample scores and language decisions.

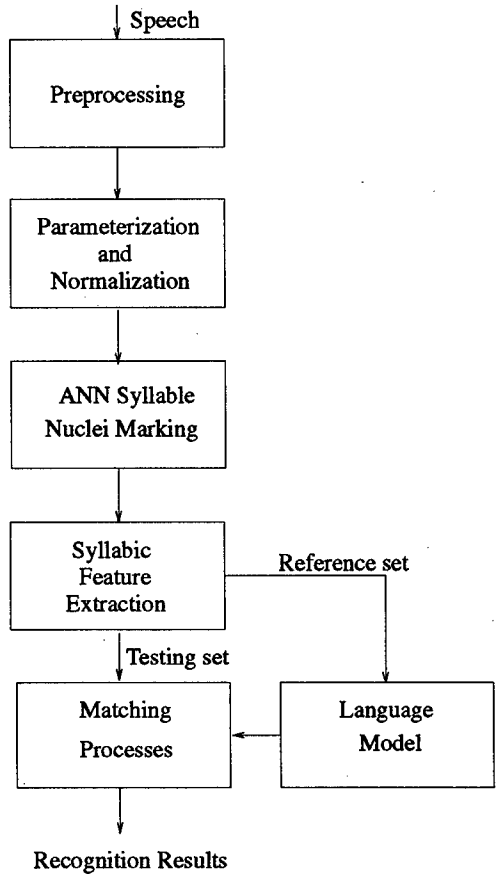


Figure 4.1. Block Diagram of ITTI Baseline DID System.

5. Database

The database used in this contract was constructed from interviews and speech material recorded under AFRL-RRS's direction in Miami, Florida.† The recorded speech data are in two forms and parts:

(1) Twenty digital tapes including data from 143 speakers, mostly Cuban and Liman.

(2) Seventeen analog tapes containing speech from many speakers, mostly from locations other than Cuba or Lima.

The digital tapes include "marking files", which indicate speaker gender, birth country and the locations (on the digital tapes) of the beginning and endpoints of six different utterances recorded during the interview. The speaker information was collected during an interview which was recorded as part of the total speaker data package.

The data on these audio and digital tapes were recorded at a sampling rate of 48 kilo-sample/sec. Down-sampling software was used to reduce the data to a more standard rate of 8 kilo-sample/sec., 16 bit linear PCM data. This reduction of the data bandwidth to a little less than 4 kHz. makes it more representative of data obtainable in tactical environments. (The effect of further bandwidth reduction was examined in the testing phase of the contract.)

The speech collected from subjects in Miami includes response to interview questions, some read phrases and "spontaneous speech". Only the last form of speech, denoted "SP" in the recording files, was used in this contract.

ITTI used data from all 143 speakers from the digital tapes, and data from the analog tapes for an additional (different) 70 speakers. The segment from the analog tapes consists of a small number of Cuban and Liman speakers and is mainly speakers of other dialects. Each speaker has one file. The combination of the two parts of the database contains speech from a total of 213 speakers.

5.1. Concentration on Cuban and Liman

As the AFRL-RRS database was being collected in Miami, the distribution of Spanish dialects collected became increasingly clear. The most frequent origins of the speakers were Cuba and Peru. Since ITTI's baseline algorithm requires about twenty speech samples from both male and female speakers of each dialect to be distinguished to use as reference material, it became clear that there would only be enough material to form adequate reference data sets and test data sets for dialects of Cuba and Lima, Peru. This decision allowed the dialectology literature survey to concentrate on the Cuban-Liman distinction.

Both reference works indicated that Cuba exhibits very little dialect diversity, hence was a reasonable dialect category to use for development and testing, but that Peru is much too diverse linguistically to be a satisfactory category. Professor Lipski

† For a detailed description of this database and its collection, see Beth L. Losiewicz, *Human Expert Identification of Latin American Dialects*, AFRL-RRS Final Report, 1996

put it thus:
"... Peruvian Spanish is really a cover term for dialects which, from a phonetic point of view, are as different from each other as, say, Jersey City, Mobile, Omaha, Sydney, and Port of Spain."

Cuban dialects are uniformly lowland in character, in conformance with the altitude of the country, whereas Peruvians speak lowland versions of Spanish along the coast north of Lima, and highland versions in some of the mountainous region. In some Eastern parts of the country, very little Spanish is spoken and when it is, it is highly distorted by admixture with the indigenous languages, still spoken by the local Indians.

However, it was found that almost all the Peruvian speakers contributing to the new database were, in fact from Lima. This was taken to indicate that the "Peruvian" sample was actually Liman, and therefore dialectally uniform enough to use in algorithm development and evaluation. A difficulty with this solution to the diversity problem is that Lima itself is experiencing an influx of population from all parts of Peru and therefore many forms of Spanish are heard there, and no effort was made to limit the Liman sample to natives of Lima. In fact, doing so might not help much, as even within the native population of Lima, there are important dialect distinctions across the socioeconomic classes. This latter fact may indeed tend to make the Liman sample collected in Miami reasonably homogeneous with respect to dialect, as it predominantly consists of educated speakers from the more affluent classes.

The crucial factor in determining the classes to be studied was that Spanish-speaking experts working for AFRL-RRS found the Cuban and Liman data classes sufficiently distinct to justify using them to study automatic and human DID, so they were adopted for this study. Later, a third class of non-Cubans and non-Limans was added.

5.2. Database Segmentation; Definition of Classes

Automatic LID and DID at ITTI makes use of a sample of the target dialect or language as "training" data, and a second data sample, from different speakers, as the "test" data. A training data set and a test data set are needed for each class of dialect (or language) to be distinguished in the testing. It was therefore necessary to segment the database to meet the recognition program requirements. The segmentation used was influenced by two other organizations working with the same data under AFRL-RRS' direction.

In 1995, Lincoln Labs provided a list of four groups of Cuban and Liman speakers in the database, which they used to test Cuban-Liman separation by their algorithm. ITTI used this assignment to test two-class performance of the speaker-based baseline system.[d]

In 1996, AFRL-RRS directed ITTI to group all other dialects speakers as a third group of non-Cuban, non-Liman speakers. This group of speakers including dialects from Puerto Rico, Costa Rica, Mexico, Columbia, Argentina, Chile, and many other Pan

[d]The best performance of ITTI's baseline system, for whole-file recognition, approaches 90% in two-dialect separation including atypical speakers of these two dialects. This performance is slightly better than the results obtained from Lincoln Lab at that time.

American countries. It is therefore referred to as the "Other" class. It contains a mixture of highland and lowland dialects and makes the separation of Cuban and Liman speakers within the three data sets much more difficult.† It was therefore necessary to create data for three categories for most of the testing and development to be done under the contract. In conformance with AFRL-RRS' direction, all further performance given in this report is for the three-class, Cuban-Liman-Other DID problem.

In hopes of comparing automatic DID performance with the performance of a group of Spanish dialectologists on the same task, ITTI did the three-class segmentation to match, as closely as possible, conditions of the human testing experiments performed by Dr. Beth Losiewicz of Colorado College. She used 13 Cuban speakers and 15 Liman speakers from the database in her experiments, so ITTI used the same data for testing, and assigned the remaining Cuban (77 speakers) and Liman (37 speakers) data to the training sets. She also used 61 of the 71 available non-Cuban, non-Liman speakers' data in her test.d Since data from ten speakers is inadequate for the training set, it was necessary to use some of the 61 speakers she used in the training set. The method used was to split the 61 speakers' data she used into two nearly equal parts and perform two experiments, using each half of the data she used for testing and the remainder of data (including the ten speakers she didn't use) for the training sets. This is shown in Table 5.1, where the two partitions of the data into training and test sets are designated "A Group" and "B Group". In measuring DID accuracy under various conditions, separate experiments can be conducted using the two partitions of the data. Performance figures given in this report are always the average of these two experiments.

†As demonstrated by the fact that the dialect recognition performance of ITTI's baseline system shows an accuracy of about 92% for Cuban-Liman, which drops to nearly 60% for Cuban-Liman-Other separation.

d For the results of Dr. Losiewicz's test of human dialect identification on this database, see Beth L. Losiewicz, *Human Expert Identification of Latin American Dialects*, AFRL-RRS Final Report, 1996

AFRL-RRS Spanish Dialect Corpus:

Labeled Data (2-10 minutes/speaker)
Cuba: 90 speakers
Lima: 52 speakers
Others: 71 speakers

B. Losiewicz's Listening Test Set: (1 segment/speaker)

Cuba: 13 speakers
Lima: 15 speakers
Others: 61 speakers

ITT Experimental Designation:

A Group:	Train	Test
Cuba	77	13
Lima	37	15
Others	40	31

B Group:	Train	Test
Cuba	77	13
Lima	37	15
Others	41	30

Table 5.1: Designation of Experimental Data

6. Baseline System Testing

A major task in this contract was testing of ITTI's Baseline DID system to establish its performance level and sensitivity to operating parameters, such as amount of speech and Signal-to-Noise ratio, which are important features affecting the success of language-related recognizers in tactical settings. (The baseline DID system was ITTI's LID system as of July, 1996, when baseline system testing began.) The most important of those tests and results are reported here.

A further purpose of this testing was to determine the relationship between the algorithm's performance on dialect recognition *vs* its performance on language identification, for which it had been developed.

6.1. Database and Recognition Task

The tests described here were performed using data from the Spanish Dialect database supplied by AFRL-RRS. The selected data (see Section 5 for details) were subdivided into three categories; Cuban, Liman and "Other". These categories are labels or descriptors assigned to the speech data under the direction of AFRL-RRS. The Cuban and Liman segments of the data are reasonably pure dialect samples. Speech in the segment called "Other" includes many different dialects, presumably different from those in the Cuban and Liman classes.

6.2. Summary of Most Important Results

6.2.1. Dialect and Language Identification Compared

Two primary determinants of DID and LID performance are the length and the number of the speech files used as reference data to represent each class to be distinguished. Experiments with this variable showed that dialect identification accuracy increases with reference file lengths up to two minutes of speech, and that there is a performance degradation when shorter files are used. Language recognition results are available for the baseline system at reference data lengths of about fifty seconds, but with a larger number of reference speakers. For comparison purposes we select dialect identification based on use of entire files (several minutes of speech) but for the smaller number of speakers available in the Spanish dialect database. (See Section 5 for details.) Unless otherwise specified, entire speech files from the AFRL-RRS Spanish dialect database were used as reference data for all tests reported here.

Using this basis of comparison, results show that separating the three Spanish dialect classes is much more difficult for the baseline system than separating any three languages encountered in development of the baseline system. On language identification, the baseline system accuracy is usually above the mid-90% range for any three languages, whereas accuracy at DID only reached the low 60% range. Furthermore, the better identification accuracy for languages is achieved on telephone-quality speech, which is generally more difficult to identify than the laboratory-quality of speech in the Spanish Dialect database.

Most of the difficulty in separating the three Spanish dialect categories, however, can be attributed to the difficulty of distinguishing the heterogeneous class Other from the more homogeneous classes Cuban and Liman, as separation of the latter two is much more successful (nearing 90%) as is shown in Figure 6.1. As the random-choice performance on a two-class separation is 50% and on a three-class problem is 33%, it is seen that the Cuban-Liman separation result is significantly farther above random performance than the Cuban-Liman-Other result is. This observation holds over the entire range of test segment durations examined.

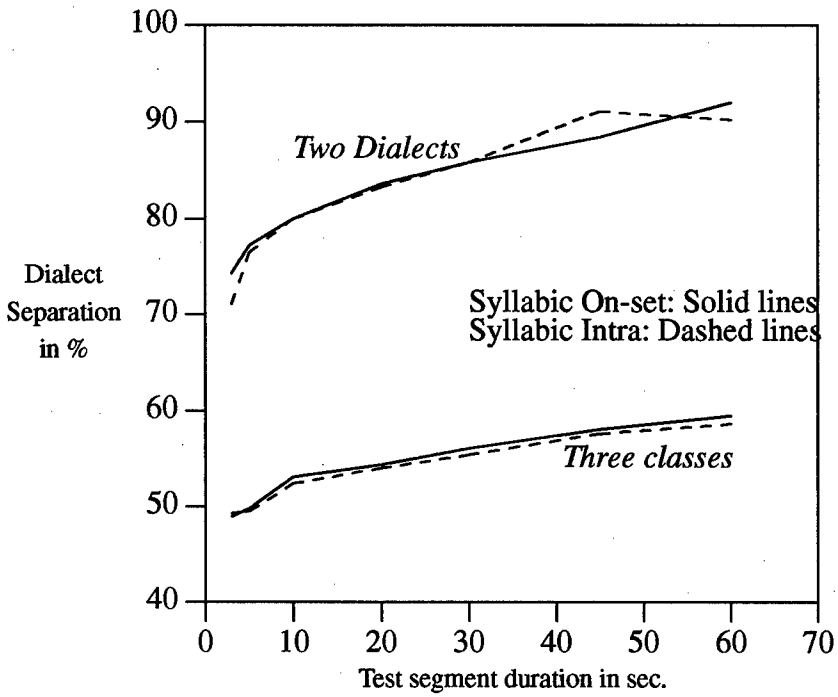


Figure 6.1 Dialect Identification accuracy of the baseline system on two dialects (Cuban *vs* Liman) and three classes (Cuban, Liman and Other).

Unfortunately, the dialectally inhomogeneous character of the Other category makes the comparison between three-class DID and three-class LID performance somewhat imprecise, but the approximately 90% performance of the baseline system on the two-class, Cuban-Liman problem corroborates the conclusion that DID, at least on AFRL-RRS' Spanish dialect database, is more difficult than LID studied to date.

6.2.2. Effect of Test Sample Duration

Another operating parameter which is highly variable in tactical applications and which impacts recognition accuracy is the duration of speech in the sample to be identified. To test the effect of test sample duration it was necessary to control the duration of the test samples in some way. Since there is no marking available for the AFRL-RRS

Spanish dialect database distinguishing speech intervals from non-speech intervals, which would be needed to compute speech duration for a sub-file, ITTI used (with AFRL-RRS' approval) syllable count as an indication of speech duration. (Syllable count is a byproduct of the baseline system feature extraction process.) Cuban and Liman speaker samples were divided into training data and test data as shown in Table 6.1. The 61 "other" dialect speakers were divided into two groups (Group A: 31 speakers, and Group B: 30 speakers) for training and testing. There were 10 additional speakers not used in the listening test called Group C. We ran two tests, differing only in the partitioning of "other" class speakers into training and test sets. Test I used A+C for training and B for test, and Test II used B+C for training and A for test. The numbers of speakers per dialect and training/test partition are shown in the table below. In both experiments, the total number of speakers is 213.

Test	Dialect	Training Speakers	Test Speakers
I	Cuban	77	13
	Liman	37	15
	Other	41(A+C)	30(B)
II	Cuban	77	13
	Liman	37	15
	Other	40(B+C)	31(A)

Table 6.1: Partitioning of AFRL-RRS database into training and test sets.

Experiments involving two-dialect discrimination of Cuban versus Liman used the same training and testing assignments, except that the "other" class was omitted.

Test segment length was measured in syllables, as opposed to elapsed time. We used an assumed average speaking rate of 4.28 syllables per second, derived from previous experience. The maximum number of test segments per speaker was limited to 10. The following table approximately relates number of detected syllables to elapsed segment duration.

Number of Syllables	Approx. Segment Length (Seconds)
13	3
21	5
43	10
86	20
128	30
193	45
257	60

Table 6.2: Relation between syllable count and elapsed segment duration.

Results are shown of the Cuban/Liman test in Figure 6.2, and of the Cuban/Liman/Other test in Figure 6.3. Each data point represents the the average correct classification accuracy measured in Tests I and II. In comparison with the listening test results, two main points are apparent:

1. Machine Dialect ID performance improves significantly with increasing test length, whereas human performance does not.
2. Machine performance appears to be better than human performance, subject to qualifications below.

Data allowing a direct and unequivocal comparison of human versus machine unfortunately does not exist. However, the following facts support the second assertion above. Human accuracy on forced-choice separation of Caribbean versus Highland dialects was measured at 62%. Machine accuracy on forced-choice separation of Cuban versus Liman is approximately 85% (for 30 second test segments). Although the machine accuracy is higher, the relative difficulty of the two tasks is unknown. Machine accuracy on forced-choice separation of Caribbean versus Highland dialects has been measured at 65% using data from an LDC database. Again, the relative difficulty of the tasks is unknown, in this case because the LDC data involves telephone conversations as opposed to interviews using a high-quality microphone. One would expect, however, that equalizing the tasks would widen the performance difference in favor of machine recognition.

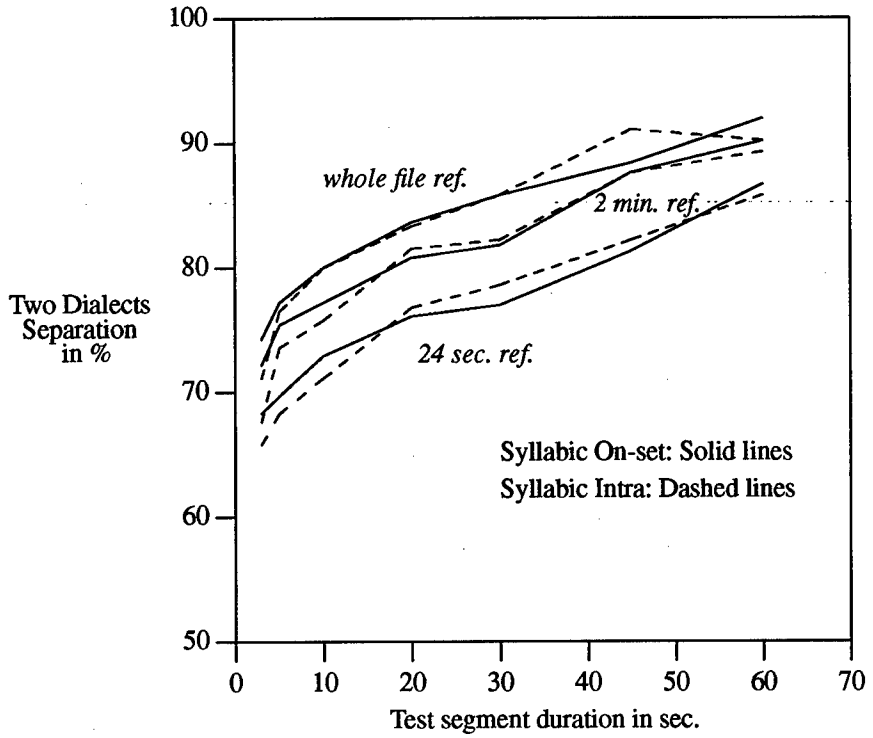


Figure 6.2: Classification of Cuban vs. Liman on AFRL-RRS Database

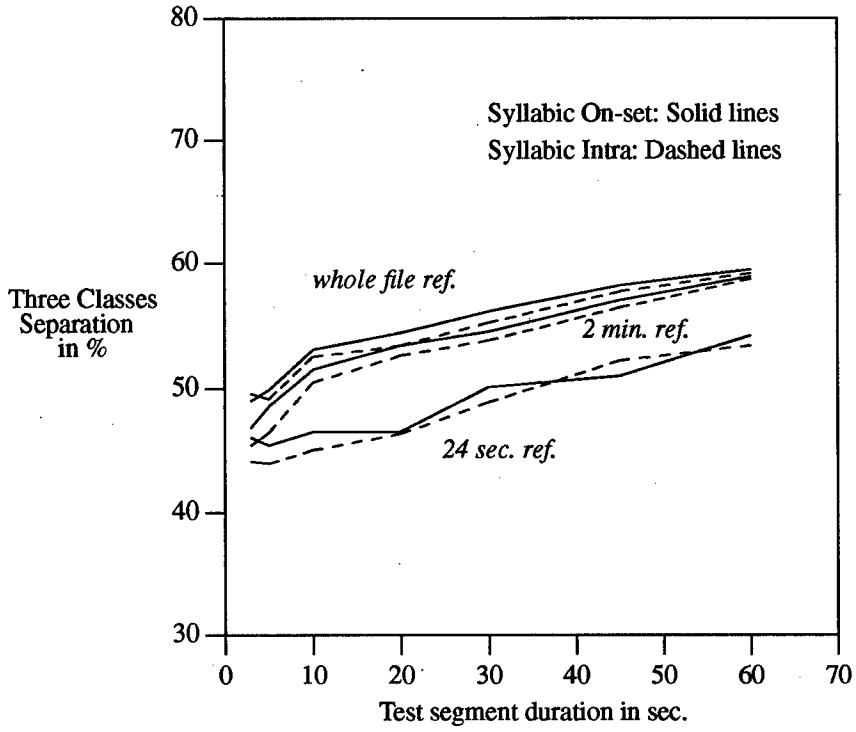


Figure 6.3: Classification of Cuban/Liman/Other on AFRL-RRS Database

We measured an additional data point, using all the available training and test data (4-10 minutes per file). The following table shows the results using onset and intra features for the three-class (Cuban/Liman/Other) experiment.

Class	Onset Features	Intra Features
Cuban	22/26 = 84.6%	23/26 = 88.5%
Liman	19/30 = 63.3%	18/30 = 60.0%
Other	37/61 = 60.7%	36/61 = 59.0%
Average	78/117 = 66.7%	77/117 = 65.8%

Table 6.3: Classification of Cuban/Liman/Other on AFRL-RRS Database using all available training data.

Note that different numbers of trials are observed for Cuban, Liman, and Other-class data. The reported average accuracy is equal to the total number of correct identifications divided by the total number of trials. Average accuracy, excluding the Other class, is 73.2%. The expected accuracy due to chance is 33.3%.

6.2.3. Effect of SNR

Perhaps the most troublesome operational parameter in tactical applications of speech-related algorithms is the presence of noise with the speech signal. Usually the noise is additive and broad in bandwith. To examine the sensitivity of the baseline DID system to additive noise, a program was developed that estimates the speaking level within a speech file, and adds white Gaussian at the required level to achieve a specified signal-to-noise ratio. The result is a new, "noisy" 16-bit sampled waveform file.

Experiments involving added noise used data with three signal-to-noise ratios (SNRs): 15, 10 and 6 db. Results are summarized in the table below for both "onset" and "intra" syllabic features. Dialect ID three-class performance is not seriously affected by added noise when training and test speech have the same SNR. The results for clean training speech versus noisy test speech show 3 to 10% loss of accuracy compared with noisy training versus noisy test speech.

Test Length	Onset	Intra
3 sec.	49.31	49.22
5 sec.	51.09	49.55
10 sec.	51.95	52.38
20 sec.	56.05	53.98
30 sec.	57.00	55.38

Table 6.4: Clean training, clean test.

Test Length	Training SNR = 15 dB		Clean Training	
	Onset	Intra	Onset	Intra
3 sec.	48.80	46.41	42.81	43.39
5 sec.	47.94	47.18	42.21	43.91
10 sec.	48.58	49.02	45.67	46.08
20 sec.	52.30	51.55	46.94	48.33
30 sec.	53.30	54.58	49.47	50.39

Table 6.5: Test SNR = 15 dB.

Test Length	Training SNR = 10 dB		Clean Training	
	Onset	Intra	Onset	Intra
3 sec.	47.60	47.59	41.36	40.17
5 sec.	49.22	48.71	43.15	40.09
10 sec.	49.47	50.99	43.53	42.05
20 sec.	51.69	53.81	46.78	43.84
30 sec.	52.21	55.37	49.17	47.15

Table 6.6: Test SNR = 10 dB.

Test Length	Training SNR = 6 dB		Clean Training	
	Onset	Intra	Onset	Intra
3 sec.	47.94	47.94	40.58	40.17
5 sec.	49.39	50.33	42.73	41.05
10 sec.	52.30	51.50	43.45	41.39
20 sec.	54.74	53.87	45.50	42.71
30 sec.	54.27	54.26	48.41	44.56

Table 6.7: Test SNR = 6 dB.

These data are summarized graphically in Figures 6.4 and 6.5 below.

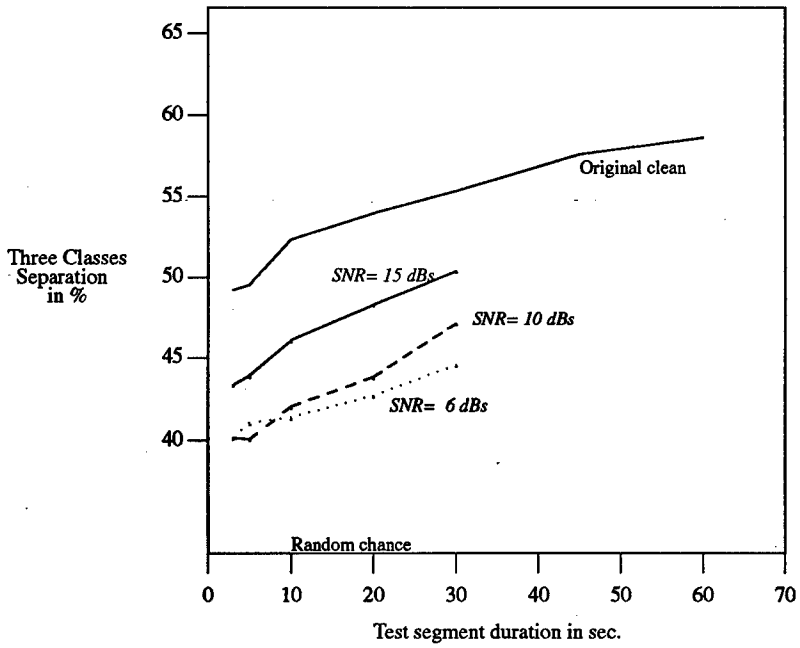


Figure 6.4: Dialect Identification accuracy of the baseline system with clean reference data and noise added to the test data, at various SNR levels.

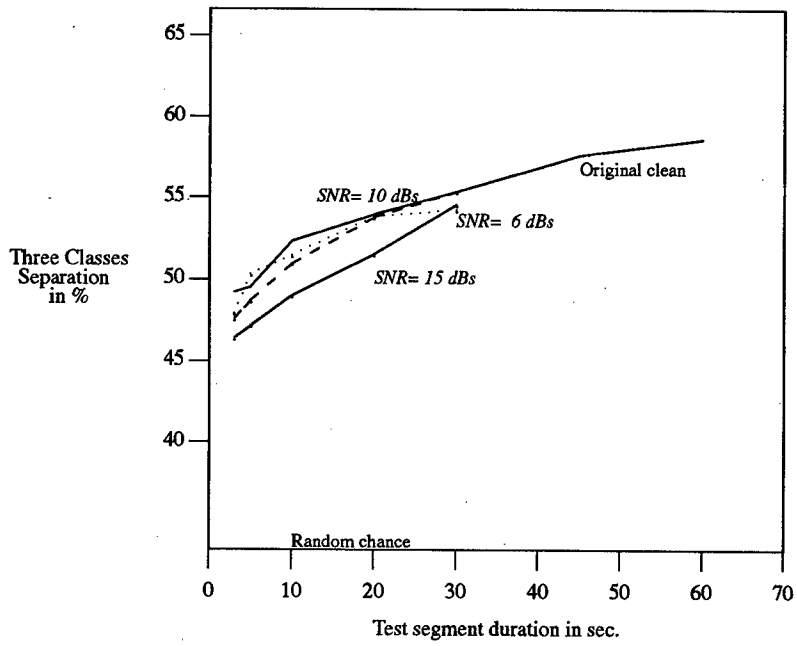


Figure 6.5: Dialect Identification accuracy of the baseline system with noise added to the test and reference data, at various SNR levels.

6.2.4. Effects of Channel Bandwidth

The baseline system nominally uses information in the band from about 250 Hz. to 3.3 kHz. The bandwidth used was artificially truncated at both the upper and lower limits to test sensitivity to data bandwidth, by reducing the number of filterbank channels used in the baseline system front end. The result shows a greater sensitivity to bandwidth loss at the low frequency limit than at the high frequency limit, because loss of low frequency information affects the syllabic marking process adversely. (This occurs when the lower band limit is raised enough to deny first formant information.)

6.2.4.1. Elimination of Highest-Frequency Filter Channels

Experiments were started involving dialect ID using various numbers of filterbank channels. The original filterbank has 14 filters covering the range 300-3560 Hz. The filter bandwidths are equal when measured on the Mel frequency scale. Dialect ID three-class performance is summarized in the tables below for both "onset" and "intra" syllabic features. The first table, labeled "14 Filters", shows performance of the baseline system using the original filterbank. The following tables, "13 Filters" and "12 Filters", are for systems in which the highest one or two frequency channels, respectively, are deleted. This effectively reduces the frequency range of the system to 300-3094 Hz (13 filters) or 300-2687 Hz (12 filters).

Test Length	Onset	Intra
3 sec.	49.31	49.22
5 sec.	51.09	49.55
10 sec.	51.95	52.38
20 sec.	56.05	53.98
30 sec.	57.00	55.38
45 sec.	57.67	57.62
60 sec.	58.77	58.64

Table 6.8: 14 Filters (Baseline).

Test Length	Onset	Intra
3 sec.	51.04	50.47
5 sec.	48.88	51.01
10 sec.	52.37	53.83
20 sec.	56.30	56.48
30 sec.	56.71	56.37
45 sec.	58.29	58.69
60 sec.	60.70	60.69

Table 6.9: 13 Filters.

Test Length	Onset	Intra
3 sec.	50.17	49.92
5 sec.	49.55	50.48
10 sec.	52.13	53.57
20 sec.	56.94	55.53
30 sec.	56.75	55.98
45 sec.	58.58	58.90
60 sec.	60.41	60.39

Table 6.10: 12 Filters.

Comparison of this 14-filter, 13-filter, and 12-filter data shows that the performance of the ITT LID system is quite insensitive to the location of the high-frequency band edge. The performance differences, if significant, favor the narrower bandwidth analysis. This suggests that salient LID information is concentrated in the lower frequencies. We hypothesize that the principal component analysis is better able to model the relevant low-frequency information when the analysis bandwidth is restricted.

We performed additional testing to determine how many high-frequency filters can be eliminated without adversely affecting accuracy. The table below shows accuracy using "intra" features only with various numbers of filters.

Test Length	Filters 1-11	Filters 1-10	Filters 1-8	Filters 1-6
3 sec.	48.71%	50.77%	50.01%	48.46%
5 sec.	49.63%	49.55%	48.27%	45.79%
10 sec.	50.92%	50.75%	48.78%	44.25%
20 sec.	52.99%	54.45%	51.66%	44.97%
30 sec.	54.31%	54.55%	51.68%	44.65%

Table 6.11: DID accuracy using 11, 10, 8, and 6 filters.

Performance data for various numbers of filters (deleting filters from the high-frequency end) is summarized in Figure 6.6.

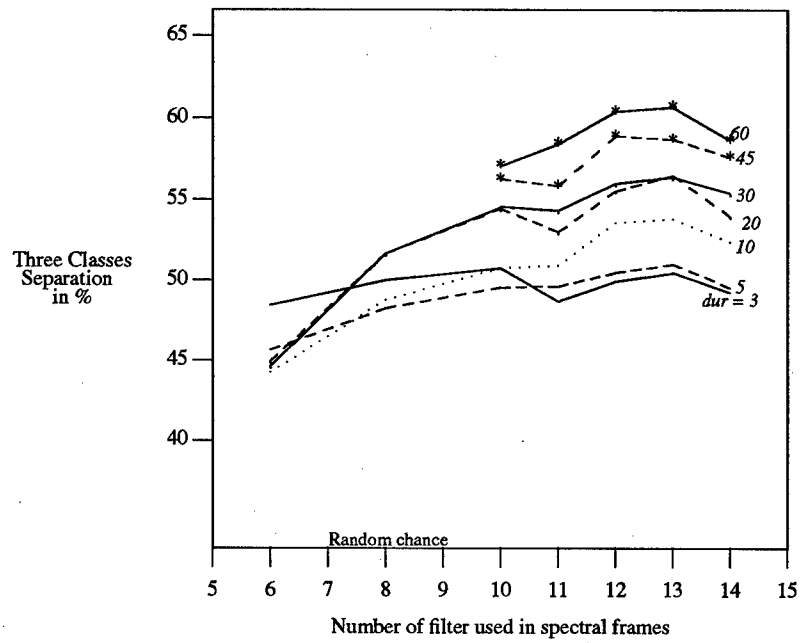


Figure 6.6: Dialect Identification accuracy using various numbers of filters.

The "knee" of performance versus number of filters depends on the length of the test data. For three-second test utterances, the number of filters can be reduced from 14 to 6 with no apparent loss of accuracy. For thirty-second tests, the number of filters can be reduced from 14 to about 11 with no apparent loss of accuracy. Interestingly, the sixth and eleventh filters correspond to about the upper limits of the first and second formants, respectively. We offer the following hypothesis to account for these experimental results. Given three-second speech samples, it is possible to detect language-dependent patterns involving only one degree of freedom of articulation: high-low tongue movements, primarily influencing F1. Three seconds is not long enough to

observe reliable patterns involving combinations of high-low and front-back movements (F1 and F2) because there are too many such combinations possible. Therefore, there is no benefit to increasing bandwidth beyond the first six filters. However, thirty seconds is sufficient to observe a more reliable statistical sampling of patterns involving both F1 and F2 movements. Therefore, accuracy improves with increasing bandwidth up to about 11 filters.

If this hypothesis is correct, it indicates that the measurements used for dialect identification should be dependent on the expected length of test utterances. For three-second utterances, it may be useful to restrict bandwidth to about 1000 Hz, and possibly to do more detailed modeling within that band.

6.2.4.2. Elimination of Lowest-Frequency Filter Channels

Experiments were completed in which the lowest filter or two filters were eliminated. In these cases, the filters used were 2-14 and 3-14, respectively. Recognition results using "intra" features are shown below, together with those for the baseline (filters 1-14) system.

Test Length	Filters 1-14	Filters 2-14	Filters 3-14
3 sec.	49.23%	48.71%	48.71%
5 sec.	49.55%	48.02%	50.49%
10 sec.	52.38%	53.22%	50.31%
20 sec.	53.99%	54.97%	53.59%
30 sec.	55.38%	54.29%	55.42%

Table 6.12: DID accuracy with deleted low-frequency filters.

Eliminating low frequency filters does not appear to cause appreciable degradation, independent of test utterance length within the 3-30 second range tested. These results are shown graphically in Figure 6.7.

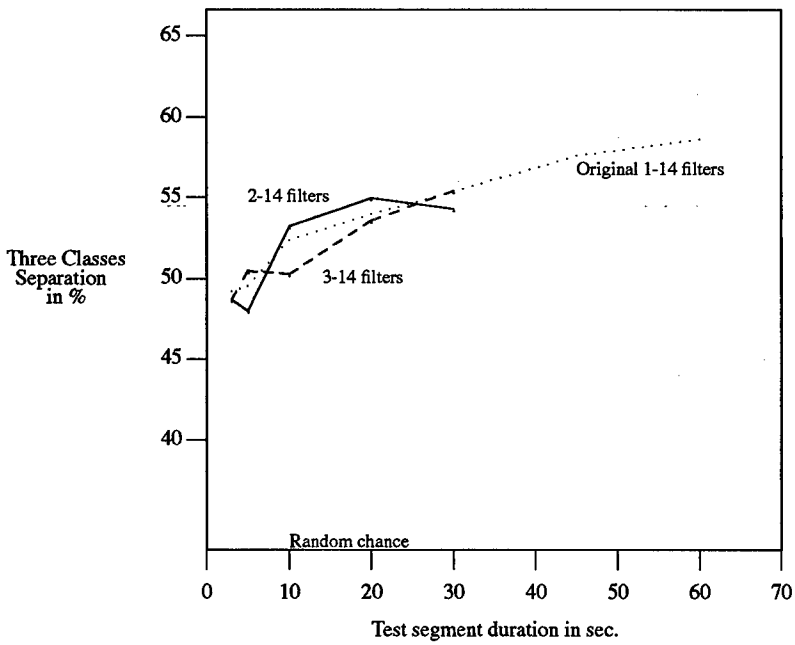


Figure 6.7: DID accuracy with deleted low-frequency filters.

6.2.4.3. Modified Filter Banks

Previous testing showed that recognition accuracy on short test segments does not change significantly when the effective bandwidth of the acoustic analysis is reduced from 4 kHz to about 1 kHz by using only the first 6 out of 14 filterbank channels. While frequencies below 1 kHz (in the vicinity of F1) are critical to dialect ID, high frequencies appear to be less important. A possible explanation is that the LID algorithm cannot effectively utilize the higher dimensionality of observations produced by the 14-channel filterbank, given the available quantity of training and test data. This motivated us to experiment with filterbanks incorporating the six low-frequency filterbank channels, plus either one or two broad filterbank channels representing frequencies above 1 kHz. These modified filterbanks covered the same range of frequencies as the original 14-channel filterbank, but with much less detail above 1 kHz.

Condition	3 Sec	5 Sec	10 Sec	20 Sec
6 filters	48.5%	45.7%	44.3%	45.0%
6 filt + 1	48.8%	46.2%	47.6%	47.6%
6 filt + 2	47.9%	47.7%	48.6%	50.2%
8 filters	50.0%	48.3%	48.8%	51.7%

Table 6.13: Dialect Identification accuracy using modified filterbanks.

These results give no indication of any significant differences for 3 second test segments. As before, there is slight improvement with increasing dimensionality for longer test segments. However, the modified filterbank does not perform better than the original when the dimensionality is fixed.

6.2.5. Spectral Tilt

The slope of the long-term spectrum of audio data is sensitive to the characteristics of the audio channel through which the data are obtained. In operational environments data may be received from a large variety of different audio channels, and this leads to uncontrolled variations in the spectral slope of the speech data. Sensitivity to this operational parameter was tested by passing the database speech through a pre-emphasis or de-emphasis filter to tilt the spectrum up or down 6 dB/octave. It was found that, when both reference and test signal are modified in the same way, the performance does not show any significant change compared to performance with the original clean speech.

Experiments were performed in which the speech data was pre-processed by applying a filter with a spectral tilt of either -6db/octave (deemphasis) or +6db/octave (preemphasis). Results are shown below.

Test Length	Baseline	-6dB/octave	+6dB/octave
3 sec.	49.23%	50.00%	49.14%
5 sec.	49.55%	49.63%	50.17%
10 sec.	52.38%	51.87%	52.46%
20 sec.	53.99%	53.81%	53.48%
30 sec.	55.38%	54.90%	54.68%

Table 6.14: Dialect Identification accuracy with spectral tilts.

There is no significant change of performance if the clean speech is processed with either +6db or -6 db/octave filtering. These results are shown graphically in Figure 6.8.

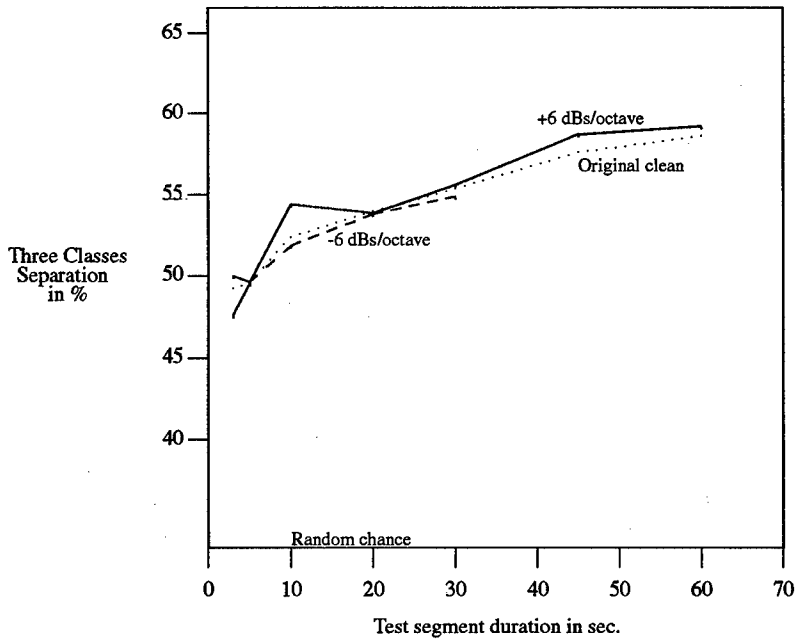


Figure 6.8: Dialect Identification accuracy with spectral tilts.

6.2.6. Performance Versus Number Reference Speakers per Dialect Class

In the baseline test, there were 70 Cuban reference speakers, 37 Liman, and 40 Other. We abbreviate this training condition as (70,37,40). We performed an experiment involving reduced numbers of speakers per dialect group. The following table compares accuracy of the baseline with that of the (25,19,20) training condition.

Test Length	(70,37,40)	(25,19,20)
On_set		
3 sec.	48.96%	47.51%
5 sec.	49.80%	47.01%
10 sec.	53.06%	50.00%
20 sec.	54.33%	50.89%
30 sec.	56.05%	51.57%
45 sec.	58.09%	53.46%
60 sec.	59.44%	55.10%
Intra		
3 sec.	49.23%	47.78%
5 sec.	49.55%	45.20%
10 sec.	52.38%	48.72%
20 sec.	53.98%	49.27%
30 sec.	55.38%	52.30%
45 sec.	57.62%	51.81%
60 sec.	58.64%	53.34%

Table 6.15: DID accuracy versus numbers of reference speakers.

The numbers of speakers per dialect group were chosen with the conflicting goals of cutting the numbers in half, while also making the number of speakers per dialect group nearly equal. The selection of speakers used or not used was random. The table shows that accuracy is reduced significantly. It is concievable that degradation could be reduced through systematic, as opposed to random, selection of reference speakers.

The number of speakers represented in the reference data was known to be an important factor for the performance of the baseline system in language recognition experiments. A similar result was found for the baseline system in its application to DID. As the AFRL-RRS Spanish dialect database only provided a small number of reference speakers for each dialect, tests of this sensitivity were very limited, and we can only reach the qualitative conclusion that, for each dialect class, each gender group should have more than 20 speakers to cover the variation due to speaker differences. If the number of reference data speakers for each dialect and gender is less than that value, the performance will suffer. There is some evidence that speaker differences are more critical for dialect recognition than in language identification.

The above data are plotted graphically in Figure 6.9

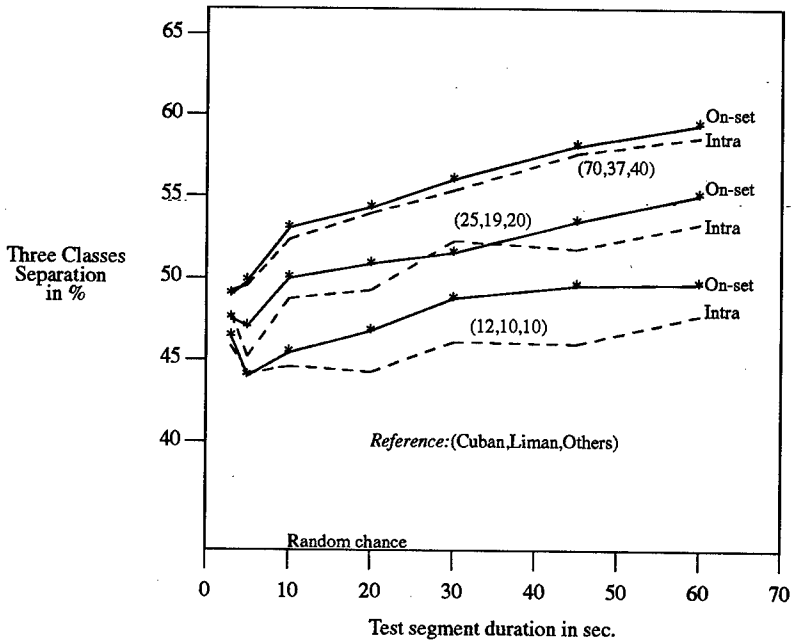


Figure 6.9: Dialect Identification as a function of the number of reference speakers.

7. DID System Development

After testing the baseline DID system, as described in the preceding Section, research was undertaken to improve its performance on the dialect identification task. This research addressed four main areas; error analysis of the dialect class confusions; experiments with a speaker-independent VQ system; experiments using syllabic prosodic features; and extraction of listener-identified dialect-specific segments. These researches are described, together with the results obtained, in this Section.

7.1. Error Analysis

In the testing performed as described in Section 6, only average recognition rates over the three classes to be separated are reported. The distribution of errors over the three possible misclassifications was examined to gain insight into the weaknesses of the baseline system. An error analysis of the confusion matrix among three classes showed that unbalanced numbers of test samples in each of three classes, and the diversity of many dialects in the "other" group have significant effects on the performance results.

In order to derive a performance score less affected by the accidental distribution of test data size, a new scoring scheme was developed. It uses the average of the percentage accuracy on each dialect instead of the percentage of the total number of test samples correctly recognized. When performance is computed in this way, the unbalanced number of test samples for each class has less effect on the experimental results. Although this improved scoring method gives a more representative performance figure, and is clearly preferable to the gross correct percentage, it does not materially change the performance and sensitivity results reported in the previous Section. Qualitative and comparative performance trends with all operating parameters remain essentially unaltered.

The dialectally ambiguous "Other" group was the source of most errors. That is, more errors were observed in distinguishing Other samples from both Cuban and Liman samples than were observed for separation of Cuban and Liman samples. Fortunately the Other group included enough speakers from Colombia to construct what is perhaps a more dialectally homogeneous three-class test. When only Colombian speakers were used for the Other category, performance showed a statistically significant improvement of two to five percent. The improvement is thought to be attributable to the reduced dialectal diversity of the Other group under this restriction, but it may also be partially attributable to a more balanced test. Unfortunately, dialectologists did not compare the dialectal diversity of the initial Other group of speakers or the Colombian-only Other group.

7.2. The Highland-Lowland Distinction

A test which would have been very interesting, could it have been carried out, would have been evaluation of recognition performance on the two-class problem of separating speakers of highland and lowland versions of Spanish. Cotton and Sharp, and many other dialectologists, find this to be the dominant distinction in Latin

American Spanish (and in fact in world-wide Spanish. See Section 3.) Although the birth cities of speakers is known for the AFRL-RRS Spanish dialect database, AFRL-RRS was not able to provide a classification of the speakers as highland and lowland dialect speakers. A small amount of data were classified in this way, in connection with some expert listening tests performed by Dr. Beth Losiewicz, but it was not of sufficient size to permit testing with the automatic system. We recommend that this be done in the future.

7.3. Speaker-independent System Tests; Comparison with LID

As mentioned in the Introduction, ITTI has developed two main types of LID systems; so-called "speaker-dependent" and "speaker-independent" systems. The speaker-dependent versions retain individual identity of reference speakers and incorporates a search for speakers similar to the unknown speaker as an initial step in DID. In contrast, the speaker-independent version pre-processes all of the reference speaker data to extract speaker-independent models of the dialect classes (in the form of cluster centers of Vector Quantized data.)

In application to language identification, it has been found that when the amount of reference data is sufficient (usually more than 1024 cluster centers), the speaker-independent language recognition performance is only slightly less than the performance of the speaker-dependent system. This may be significant in operational systems because the computational requirement for speaker-independent is at least two orders of magnitude less than for speaker-dependent recognition.

When the speaker-independent system was applied to the three-class DID problem, it was found to perform significantly worse than the speaker-dependent system (yielding less than 40% accuracy), for all test segment durations. We conclude that dialect recognition depends upon or requires the more precise speaker characteristic matching afforded by the speaker-based system, as it searches for the best group of speakers for scoring the testing samples. However, this result should be considered tentative, as it may be affected by the small size and the makeup of the AFRL-RRS Spanish dialect database.

7.4. Experiments Using Syllabic Prosodic Features

The Literature Survey, as reported in Section 3, found that dialectologists cite prosodic features, including vowel coloration and the "rhythmical" properties of speech, as dialect determinants. However, it is difficult or impossible to define objective criteria for application of these properties to distinguish dialects. Automatic DID can, however, access and measure acoustic correlates of these prosodic features for comparison with reference data, and thereby make use of these properties.

The acoustic correlates of prosodic features are found in the pitch and amplitude contours of syllables or vowels. ITTI had developed a prosodic feature extraction capability which operates in the vicinity of vocalic nuclei for use in its LID system, and have used it to merge spectral features as part of combining multiple systems to improve language identification recognition performance. In language identification the

performance of the syllabic prosodic features by themselves have shown very poor results, so for DID they were combined with spectral vocalic features to form a single feature space. This method of processing was compared with the performance using only syllabic spectral features. It was found that recognition performance shows some degradation at the short duration (3-30 sec.) range. It was concluded that at present "syllabic prosodic" feature extraction does not show any benefit for dialect recognition, contrary to what might have been expected by reading the comments of Spanish dialect experts. Perhaps the features explored in this research fail to capture the linguistic properties referred to by these experts. It would be necessary to work closely with an expert, using many, well-chosen examples to ensure that the percepts they reference are in fact detected by the acoustic features actually extracted. This is a line of research which dialectologists might welcome, since a side benefit, if the effort were successful, would be an automatic system for detecting and measuring - hence reducing to objective terms - a speech property they have long noted but been unable to quantify.

7.5. Listener-identified Dialect-specific Segments

An attempt was made to use the results found in the Literature Survey. In that study, several acoustic-phonetic features were cited as features which Spanish dialect experts have identified as major differences between Cuban and Liman. A Spanish-speaking listener who was familiar with the results of the Literature Survey claimed that by listening to only a short utterance he could distinguish Cuban and Liman dialect speakers. To test the potential implication of this observation for automatic DID, the listener marked some phonetic sequence segments in the test data as typical samples for each dialect. If those marked segments do contain the distinctive features for these two dialects, they could be used as a reference data set and search for the best match distance measure in a test sample, and use that distance to identify the dialect. A test of this idea was performed.

The Spanish-speaking listener used a graphic aid system to mark boundaries of speech segments in which he heard evidence of each dialect. These segments comprised 30 to 50 seconds of speech extracted from a small number of speakers from each of three dialect groups. A recognition experiment using the combined syllabic spectral and prosodic features to represent the speech he selected was then performed.

The recognition performance of this system, for short duration (3 to 10 second) test segments, shows the same performance as the speaker-based ITTI baseline system. However, for longer duration (longer than 10 seconds) test segments, the performance is not only significantly lower than for the baseline system, but also at test segment durations longer than 45 seconds, the performance was even lower than for test segment durations of 3 to 5 seconds. This is contrary to the expectation that longer test samples always produce better recognition. (At the longer durations, the performance is about the same as the speaker-independent VQ system.)

There are several possible inferences from this mixed experimental result. On the positive side, the reference set in this experiment contains only 30 to 50 seconds of speech for each dialect, so we may speculate that if the number of marked segments is

increased several fold, the recognition performance might increase significantly and at some point exceed that of the baseline system. With the small amount of reference data used in this experiment, processing is more than two orders of magnitude faster than with the normal baseline system reference data, so the reference data could be increased a lot and still process much faster than with baseline system reference data. So there is some indication that listener-identified data might be used both to increase performance and reduce processing time. Also, if the short test segment result indicate that the listener is in fact identifying dialect-diagnostic data, it indicates that dialect descriptions in the literature have potential for automatic DID, if the listening process can be automated.

However, the degradation of DID accuracy at longer durations is puzzling and renders any conclusion hazardous. Normally, longer test samples give greater accuracy. However, longer test samples do bring a wider range of linguistic performance into play; perhaps the small number of marked segments is in some sense "swamped" by the greater diversity. If so, the reversal at longer durations may vanish if many more samples are marked and incorporated as reference material. Although the need to mark a large amount of data is incompatible with field operation of a DID system for tactical applications, this is nevertheless a fruitful avenue of further research, as it might eventually become possible to automate the segment selection being done by the listener.

8. Conclusions

The major conclusions to be drawn from the results of this contract are as follows:

1. The available literature for Latin American Spanish and Arabic dialects was reviewed. It was found that no wholly satisfactory definition exists for the term "dialect", as many experts disagree on the criteria differentiating dialects. Nevertheless, it was possible to compile a list of major distinctive acoustic-phonetic units of Arabic dialects and Cuban and Liman Spanish dialects. It is to be expected that the structure, approach and available data for differentiating dialects will vary widely over languages, as was the case between Arabic and Spanish.

2. Robustness of the baseline DID system was evaluated in noisy environments, under channel variations and under variations in other operating parameters likely to be important in tactical environments. It was found that there is strong system robustness in recognition performance if the training and testing data are under the same environment. Major dependencies of the system are the number of reference speakers used for training for each dialect and the duration of the unknown speech segment.

3. Recognition errors on the three-class DID separation of Spanish Cuban-Liman-Other were analyzed. The dialectally ambiguous class "Other" is responsible for most errors. Replacing that class by one consisting of Colombian speakers caused slight improvement. It is not known how dialectally diverse the Colombian group is. Much better performance (relative to the random chance result) was obtained for the two more dialectally homogeneous classes Cuban and Liman.

4. A computationally efficient speaker-independent VQ-based system for dialect recognition was tested. It was found to perform much poorer on DID than on LID, from which we conclude that dialect recognition requires more precise normalization of speaker differences than does LID.

5. Under motivation provided by the Literature Survey, the usefulness of syllabic prosodic features for DID was investigated. It was found that dialectologists sometimes cite vowel coloration and the "rhythm" of speech as diagnostic of dialect. It was found that incorporating these features does not improve performance unless the duration of the tested sample is longer than 30 seconds.

6. Replacing whole-file reference data representing the dialect classes to be recognized by a small number of diagnostic speech segments selected by a Spanish-speaking listener produced recognition accuracy comparable to the whole-file results, with orders of magnitude less reference data, for short duration test

segments. This suggests several lines of investigation to improve DID performance, and may indicate that some of the phonological features used by dialectologists to distinguish dialects may be usable in an automatic system. However, very poor performance on longer duration test segments is not understood and further evaluation with a larger number of marked samples is necessary to clarify the result.

References

Char64a.

B. Charity, k. Stowasser, and R. Wolfe, in *A Dictionary of Iraqi Arabic*, 1964.

Cott88a.

Eleanor Greet Cotton and John M. Sharp, *Spanish in the Americas,* Georgetown University Press, Washington, D.C., 1988.

Harr62a.

R.S. Harrell, in *A Short Reference Guide to Morrocan Arabic*, p. 10, 1962.

Kayea.

Alan S. Kaye and Judith Rosenhouse, "Arabic Dialects and Maltese," in *The Semitic Languages*, ed. Robert Hetzron, Routledge, London.

Li94a.

K. P. Li, "Automatic language identification using syllabic spectral features," *ICASSP-94*, pp. I-297-300, April, 1994.

Li95a.

K. P. Li, "Experimental Improvement of a Language ID System," *ICASSP-95*, pp. I-3515, May, 1995.

Li96a.

K. P. Li, "Language and Dialect Identification by Syllabic Spectral Features," *J. Acoustical Society of America*, vol. 100, no. 4, pp. I-2760, October, 1996.

Losi94a.

Beth L. Losiewicz, *Preliminary Report on the Feasibility of Machine Spanish Dialect Identification and Description of Latin American Dialect Database (LADD) Development,* Expeimental Psychology, The Colorado College, Colorado Springs, CO, August, 1994. Sponsored by Air Force Office of Scientific Research, Bolling AFB and Rome Labs

Reka94a.

D. M. Rekart and M. A. Zissman, *Dialect Labels for the Spanish Segment of the OGI Multi-language Telephone Speech Corpus,* Lincoln Laboratory, MIT, Lexington, MA, September 7, 1994.

Resn75a.

Melvyn C. Resnick, *Phonological variants and dialect identification in Latin American Spanish,* Mouton, Paris, 1975.

☆U.S. GOVERNMENT PRINTING OFFICE: 1999-610-130-81150

MISSION OF AFRL/INFORMATION DIRECTORATE (IF)

The advancement and application of information systems science and technology for aerospace command and control and its transition to air, space, and ground systems to meet customer needs in the areas of Global Awareness, Dynamic Planning and Execution, and Global Information Exchange is the focus of this AFRL organization. The directorate's areas of investigation include a broad spectrum of information and fusion, communication, collaborative environment and modeling and simulation, defensive information warfare, and intelligent information systems technologies.