Identification of Invariant Acoustic Cues in Stop Consonants Using the Wigner Distribution

by

Harinath Garudadri

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

in

THE FACULTY OF GRADUATE STUDIES
Department of Electrical Engineering

We accept this thesis as conforming to the required standard

THE UNIVERSITY OF BRITISH COLUMBIA
December, 1987

© Harinath Garudadri, 1987
In presenting this thesis in partial fulfilment of the requirements for an advanced degree at the University of British Columbia, I agree that the Library shall make it freely available for reference and study. I further agree that permission for extensive copying of this thesis for scholarly purposes may be granted by the head of my department or by his or her representatives. It is understood that copying or publication of this thesis for financial gain shall not be allowed without my written permission.

Department of Electrical Engineering.

The University of British Columbia
Vancouver, Canada

Date December 1987.
Abstract

It is a common belief that there are invariant acoustic patterns in speech signals, which can be related to their phonetic description. These patterns are expected to remain invariant, independent of the language, speaker, phonetic context, etc. Although many investigations based on short-time spectral analysis have established the feasibility of extracting invariant cues in certain contexts, they could not provide a set of invariant cues in any given phonetic context. In this thesis, the Wigner distribution (WD) was used to analyze speech signals for the first time, to investigate acoustic invariance.

The WD, like the spectrogram, provides a time-frequency description of the signal. Unlike the spectrogram, it provides correct marginals in the time and frequency domains, but it is not a positive distribution. It is demonstrated here that the partially smoothed WD, in which both the properties of positivity and correct marginals are sacrificed to some extent, provides a better time-frequency resolution than short-time spectral analyses methods. An implementation and an interpretation of the partially smoothed WD are presented. The choice of smoothing parameters and the nature of cross-term suppression in a partially smoothed WD are discussed in detail. It is shown that the cross-terms in a partially smoothed WD do not mask the underlying nature of a signal in the time-frequency plane.

A partially smoothed WD was used to investigate acoustic invariance in voiceless, unaspirated stop consonants spoken by native speakers of English, Telugu and French. Contrary to reports in the literature, it was shown that the features “diffuse-rising” and “compact” spectral shapes were not unique to alveolar and velar places of articulation, respectively, but depended on the vowel context. The resulting ambiguities when specifying the place of articulation were resolved using Formant Onset Duration (time taken for the steady state formants to occur in the vocal tract after the consonantal release) and $F_2$ of the following vowel. The place of articulation was specified correctly for 86% of the tokens. Unlike in other investigations, the errors in specifying the place of articulation were uniformly distributed over all vowel contexts.
# Contents

**ABSTRACT**

**LIST OF TABLES**

**LIST OF FIGURES**

**ACKNOWLEDGEMENTS**

## 1 Introduction

1.1 Motivation ................................................. 1
1.2 Statement of the Problem ................................. 2
1.3 Statement of the Results ............................... 3
1.4 Organization of the Thesis .............................. 5

## 2 Acoustic Invariance in Stop Consonants: Earlier Investigations

2.1 Overview .................................................. 6
2.2 Background ............................................... 6
2.3 Earlier Studies for Acoustic Invariant cues ............ 8
  2.3.1 Context dependencies in Earlier Studies ........... 11
2.4 Drawbacks of the Short-Time LP Smoothed Spectra .... 12
2.5 Summary ................................................. 14

## 3 Smoothed Wigner Distribution


## Chapter 4

4.3.4 Comparison of Invariant Patterns from WD Plots and Short-Term Spectra ............................................. 62
4.3.5 Voiced and Aspirated Stop Consonants .......................... 67

4.4 Development of the Program FEATURE ................................. 68
4.4.1 Measurement Section .................................................. 71
4.4.2 Logic Section ......................................................... 72

4.5 The Experiment ........................................................... 76
4.5.1 Training of Judges .................................................... 76
4.5.2 Modification to the Program FEATURE ............................ 79
4.5.3 Data for the Cross-Language Study ................................. 79
4.5.4 Running the Program FEATURE ..................................... 81

4.6 Summary ........................................................................ 82

## Chapter 5

5 Results and Discussions .................................................... 83

5.1 Overview .................................................................... 83

5.2 Overall Performance ...................................................... 84
5.2.1 Performance by Language ............................................. 84
5.2.2 Performance by Place of Articulation ............................... 85
5.2.3 Performance by Vowel Context ....................................... 86
5.2.4 Overall Scores ........................................................ 86

5.3 Performance at Each Node ............................................... 93

5.4 Relevance of the Acoustic Features Measured in WD Plots ........ 95
5.4.1 Spectral Shape ......................................................... 96
5.4.2 Formant Onset Duration and Burst Duration Statistics ... 98

5.5 Summary ..................................................................... 100

## Chapter 6

6 Summary and Suggestions for Future Research ......................... 101
List of Tables

2.1 Results of Blumstein and Stevens, (1979) for voiced CV syllables. 9
2.2 Results of Kewley-Port, (1983) 10
3.1 Examples of some kernel functions and the resulting representations 25
4.1 Organization of the stops in the languages studied. 51
4.2 Comparison of the analysis methods. 62
4.3 FOD and FOD–BD values from the pilot data. 75
4.4 The choice of original thresholds. 76
4.5 The choice of new thresholds. 81
5.1 Overall scores by language. 84
5.2 Overall performance by place of articulation. 85
5.3 Performance by vowel context. 86
5.4 Performance by place of articulation and language. 87
5.5 Performance at each node in the decision tree. 94
5.6 Distribution of the feature “spectral shape” by place of articulation. 96
List of Figures

3.1 Spectrogram of an exponential. ........................................ 17
3.2 Wigner distribution of an exponential. .............................. 17
3.3 Spectrogram of a chirp. ............................................... 17
3.4 Wigner distribution of a chirp. ...................................... 17
3.5 $G(t,\omega)$ for a pseudo-WD. ...................................... 26
3.6 $G(t,\omega)$ for a narrowband spectrogram. ........................ 26
3.7 $G(t,\omega)$ for a wideband spectrogram. ............................ 26
3.8 $G(t,\omega)$ for a smoothed WD with $T\Omega = 0.25$. .............. 26
3.9 Subroutine for computing the WD .................................... 29
3.10 Cause of spectral wrap in a pseudo-WD. ........................... 34
3.11 Spectral wrap in a chirp signal. .................................. 34
3.12 Effective cross-term suppression as the spectral distance between self-terms decreases. ...................................... 36
3.13 WD of English phone [de] with no smoothing. $T\Omega = 0.01$ .... 38
3.14 WD of English phone [de] with partial smoothing. $T\Omega = 0.25$ . 39
3.15 WD of English phone [de] with complete smoothing. $T\Omega = 1.0$ . 40
3.16 Pseudo WD of two chirps showing negligible cross-terms during silence. .................................................. 41
3.17 Smoothed WD of two chirps. $T\Omega = 1.0$. ........................ 42
3.18 Smoothed WD of an alveolar depicting negligible cross-terms during FOD. ................................................ 44
3.19 Smoothed WD of a velar depicting negligible cross-terms during FOD. ................................................. 45

4.1 WD of the beginning of the syllable [ka] (Telugu speaker) ........ 54
4.2 WD of the beginning of the syllable [ga] (English speaker) .... 55
4.3 WD of the beginning of the syllable [tta] (Telugu speaker) .... 56
4.4 WD of the beginning of the syllable [pa] (French speaker) ..... 57
4.5 Examples of typical spectral shapes observed in the WD plots. ... 59
4.6 WD of the beginning of the syllable [ge] (English speaker) .... 65
4.7 WD of the beginning of the syllable [du] (English speaker) .... 66
4.8 WD of the beginning of the syllable [di] (Telugu speaker) ..... 69
4.9 WD of syllable [kʰa] from English. ............................................. 70
4.10 Flowchart for measuring the visual patterns. ......................... 73
4.11 Flowchart for assigning the place of articulation ..................... 74
4.12 Template used for measuring the visual patterns ...................... 78

5.1 WD of the beginning of the syllable [be] (English speaker) .... 89
5.2 WD of the beginning of the syllable [da] (English speaker DG) ... 90
5.3 WD of the beginning of the syllable [te] (Telugu speaker) ...... 92
5.4 Distribution of diffuse non-falling and compact spectral shapes in alveolars and velars. .................................................. 97
5.5 FOD–BD vs FOD scatter plot for English tokens ....................... 99
5.6 FOD–BD vs FOD scatter plot for Telugu tokens ...................... 99
5.7 FOD–BD vs FOD scatter plot for French tokens ..................... 100

B.1 Help Screen of WIG4C ...................................................... 111
B.2 Default Wigner Parameters .............................................. 111
B.3 Default Plot Parameters ............................................... 112
Acknowledgements

I am very grateful to my advisors Dr. Michael Beddoes and Dr. John Gilbert, for their continual help, supervision and support. I would like to thank Dr. Andre-Pierre Benguerel for taking active interest in this dissertation and providing me with the necessary background information in phonetics. I am very thankful to Dr. Peter Lawrence for his help in acquiring the speech data and especially for the encouragement he has offered throughout this work.

The topic of this dissertation originated at the Indian Institute of Technology, Kanpur, in 1981. I would like to thank Dr. Satish Mullick for introducing the Wigner distribution to me and encouraging me to pursue the topic of acoustic invariance.

I thank all my friends, who contributed directly and indirectly to this thesis. Special thanks are due to Prem Lakshmanan, Subroto Bhattacharya, Sudhakar Cherukupalli, Vasudha Vipat, Darrell Wong and John Nicol. I would also like to thank Carol McGeough, Teresa Ukrainetz, Mike Lee and Dr. Jack Bodolec for participating in the study.

I would like to thank Maureen Beddoes, Joanna and Fritz Lehmann, and Dr. Susheela Reddy for making my stay in Vancouver very pleasant. It would be futile to try to express my feelings towards John Gilbert and Carolyn Johnson. I am indeed grateful for everything they did for me and I will always remember Vancouver with fond memories.
Dedicated to Amma and Bhaskar.
Chapter 1

Introduction

1.1 Motivation

From the dawn of civilization, speech has played an important role in society. Evolutionists agree that the development of a speech centre in the brain led to the evolution of homo sapiens and it is speech which sets us apart from primates. Speech is the most effective and natural means of communication among humans, as Edward Sapir said, "the best show that man puts on."

Over the past 50 years, many advances have been made in understanding speech — both its production and its perception. It has become possible to rehabilitate people with speech and hearing disorders by providing them with aids to cope with their impairments. With the advent of computers, many and varied applications such as reading machines for the blind, speech understanding systems, voice operated remote machines to replace humans in hazardous environments or to operate in inaccessible areas, computer generated speech in information retrieval, security systems, etc. have become feasible. However, there are still many aspects of speech production and perception, that are not well understood. A thorough understanding of the nature of speech is essential in order to achieve a desirable level of success in speech related applications, and to assure the quality and reliability of speech technology, comparable with that of, say, home electronics or automobiles.

This study is directed towards understanding the nature of some speech signals, i.e. stop consonants. Stop consonants are the sounds produced by a complete closure of the vocal tract, e.g., [k] and [t] in words like "cat." Unlike vowels,
which are steady state signals, stop consonants are non-stationary and among the
hardest to tackle in signal processing. A new signal processing tool, called the
"Wigner distribution" (WD), is used here in a cross-language study of unaspirated
stop consonants from English, Telugu and French. The WD has not been used
previously, to analyze speech signals extensively. As will be shown, the WD is
particularly suited to study the nature of stop consonants.

1.2 Statement of the Problem

The concept of "acoustic invariance" can be explained as follows:

*Is there some unique acoustic aspect of a given speech segment that
remains invariant, independent of language, speaker, phonetic context,
etc., that can be related to the phonetic description of that speech
segment?*

Acoustic invariance hypothesizes that such unique acoustic properties are present
in speech signals. It is a common belief that these properties (called "invariant
acoustic cues"), are used by humans in perceiving speech. Hence, it should be
possible to extract them from the signal, using signal processing methods. When
these cues refer to the location in the vocal tract, where the airflow is momentarily
stopped during speech production, they are called "invariant acoustic cues for place
of articulation" in stop consonants.

Identification of invariant acoustic cues for place of articulation has been a topic
of active interest for many years. The signal processing tools used previously to
extract acoustic invariance were essentially short-time spectral estimation meth­
ods. Various time-varying spectral properties have been investigated as possible
correlates for place of articulation in stop consonants. Although the feasibility of
extracting some invariant properties from speech has been unequivocally estab­
lished, these investigations have not succeeded in providing a set of necessary and
sufficient acoustic properties of the signal to specify the place of articulation. As
a consequence, the concept of acoustic invariance is still viewed with considerable
reservation [53].
WD is a relatively new signal processing tool. It overcomes some of the drawbacks of conventional Fourier techniques. Over the past few years, it has received much attention in applications involving non-stationary and transient signals. Although the technique has proved to be very useful for analyzing monocomponent signals, when used to analyze multicomponent signals, it is circumscribed by the presence of negative energy regions and cross-terms. As a consequence, WD is not widely accepted as an analysis tool for multicomponent signals such as speech [18]. The use of partially smoothed WD to analyze speech signals is a novel aspect of this thesis. The partially smoothed WD was used to investigate acoustic invariance in stop consonants.

1.3 Statement of the Results

The contributions of this study were in: (i) developing the WD as an analysis tool for processing multicomponent signals such as speech, and (ii) presenting evidence which supports the concept of acoustic invariance in speech signals.

Wigner Distribution

The WD is a bilinear time-frequency distribution. It has the property of correct marginals, and hence provides an accurate description of the time-varying spectral properties of a signal. Unlike short-time spectral analysis techniques, the WD is not a positive function. The presence of cross-terms, due to its non-linear nature, poses problems in many applications. In this study, two most desirable properties of a signal processing tool — (i) positivity and (ii) correct marginals, are both sacrificed by considering a partially smoothed WD. Extremely satisfactory results were obtained when the partially smoothed WD was used to analyze stop consonants and other transient signals.

- The WD can be smoothed to overcome the drawbacks of negative regions and cross-terms. It was shown that, the WD smoothed for positivity (to suppress the cross-terms completely), is identical to a spectrogram computed using Gaussian windows. The partially smoothed WD was shown to provide
greater time-frequency resolution than short-time Fourier techniques. [28,30].

- An interpretation of the partially smoothed WD is presented and the choice of smoothing parameters for multicomponent signals, with particular reference to speech were presented [29]. An implementation of the partially smoothed WD was discussed in detail.

- The nature of cross-terms was discussed with illustrative examples. It was shown that non-zero contributions in the WD, where the speech signal itself is zero, do not mask useful information in the WD.

Acoustic Invariance in Stop Consonants

The partially smoothed WD was used in a cross-language study to investigate invariant acoustic properties for place of articulation in 200 unaspirated consonant-vowel (CV) syllables, spoken by native speakers English, Telugu and French. [31]

- The place of articulation was identified correctly for 78% of the bilabials, 81% of the coronals and 98% of the velars, based on visual features extracted from the WD plots. The distribution of errors in bilabials and coronals was not uniform across all speakers. Identification of velars was best for all speakers in all vowel contents.

- Temporal characteristics of the burst and second formant of the vowel were used to reduce the context dependency in specifying the place of articulation. Unlike other investigations, no context dependencies with respect to the following vowel could be found in the results.

- Usually, the acoustic feature “diffuse non-falling spectral shape” is associated with an alveolar burst and the acoustic feature “compact spectral shape” is associated with a velar burst [68,2]. It was shown that depending on the following vowel, a diffuse non-falling spectral shape can occur in velar bursts and a compact spectral shape can occur in alveolar bursts.

- Dentals and retroflexed alveolars of Telugu were shown to have similar spectral shapes during the burst in the WD plots. This result supports theoretical
predictions of Blumstein and Stevens [2], that dentals and alveolars should have similar acoustic properties, since they both belong to the class of coronals [2,33,12,51].

1.4 Organization of the Thesis

Chapter 2 describes some of the earlier studies, which used LP method for investigating acoustic invariance. The drawbacks of the LP method for analyzing the stop consonants are discussed.

Chapter 3 concerns the Wigner distribution. Computation of partially smoothed WD and its interpretation when applied to multicomponent signals are presented. Examples are presented to illustrate the nature of cross-terms and it is argued that they do not prohibit extraction of useful information from WD.

Chapter 4 describes the experiment conducted to investigate acoustic invariance in stop consonants. Section 4.2 introduces some of the linguistic terms used throughout this thesis. The invariant patterns observed in the pilot study are explained and compared with those reported in the literature. A program called FEATURE was developed, which interactively obtains the relevant visual features of a WD plot from the judges, and assigns a place of articulation to the WD plot.

Chapter 5 discusses the results of the cross language study in detail.

Chapter 6 presents the conclusions and some suggestions for future work.

Appendix A describes the hardware and software used in acquiring the speech data for the pilot study and the final experiment.

Appendix B is intended as a manual for using the program WIG4C.F to compute and plot WD. The program is available on the UBC "General" computer system in a form that can be readily applied to other signals.

Appendix C contains the instructions given to the judges to identify invariant visual patterns in the WD plots.
Chapter 2

Acoustic Invariance in Stop Consonants: Earlier Investigations

2.1 Overview

This chapter describes some of the earlier investigations for specifying the place of articulation in stop consonants, based on their acoustic properties.

Section 2.2 presents a brief background of the concept of acoustic invariance.

Section 2.3 summarises three recent investigations pertinent to the present work, based on LP method. The context dependencies in these investigations are summarised in Section 2.3.1.

Section 2.4 presents the drawbacks of using LP technique for studying stop consonants.

2.2 Background

The subject of invariant acoustic patterns as cues to the perception of phonemes has been of active interest for many years. Several investigators have examined various features as possible invariant cues for place of articulation [20,22,68,2,3,49,51]. These investigators have essentially used two basic approaches in an attempt to establish whether a given set of acoustic properties of the signal can function as invariant acoustic cues. The first method was to analyze a corpus of reasonable size and show that the said acoustic properties were present in the corpus to a sta-
tistically reliable extent. The second approach was to synthesize speech segments such that the given acoustic properties were present in them. The perception of the synthetic segments was then investigated to assess the importance of these properties in specifying the place of articulation.

Examples of acoustic properties investigated as possible correlates for place of articulation include the spectral shape of the acoustic signal during consonantal release, the formant loci (or target frequencies), and the direction of spectral change within a short time interval after the closure release. To some extent, most of these features proved to be perceptually relevant in specifying place of articulation. On the whole, they failed to provide a set of necessary and sufficient acoustic properties to specify the place of articulation of a stop consonant, independent of the phonetic context. This context dependency is often quoted to argue that it is not possible to extract invariant acoustic patterns from a speech segment, that would correlate with its phonetic description [52,53].

A possible reason for the failure of these investigations to provide invariant acoustic cues in all phonetic contexts is that, the signal processing tools used were not suitable to analyze, or synthesize the acoustic properties under consideration. This is true to some extent, when short-time spectral analysis methods are used to analyze stop consonants; and when such information is used to synthesize stop consonants. The WD appears to be a more suitable technique for time-varying spectral analysis of transient signals, such as stop consonants. The WD was used in this study to investigate acoustic invariance in stop consonants. It was desired to see, (i) if it was possible to obtain a better representation of the time-varying spectral properties of stop consonants using the WD, and (ii) if it was possible to extract invariant acoustic cues for place of articulation, independent of the vowel context, using the WD.

Recently, acoustic invariance in stop consonants was examined by Blumstein and Stevens [2], Kewley-Port [49] and Lahiri, Gewirth and Blumstein [51] using short-time Linear Prediction (LP) smoothed spectra. Their findings are described briefly, and some drawbacks of LP method when analyzing stop consonants are discussed in rest of this chapter, since they are pertinent to the present investigation.
2.3 Earlier Studies for Acoustic Invariant cues

Blumstein and Stevens

Based on theoretical considerations and perceptual experiments, Stevens and Blumstein [68] predicted the spectral shape of the burst, for different places of articulation during the production of stop consonants, as follows:

- diffuse-falling or diffuse-flat spectral shape for bilabials,
- diffuse-rising spectral shape for alveolars, and
- compact spectral shape for velars.

In order to provide a quantitative measure of the degree to which these spectral shapes correspond to a given place of articulation, Blumstein and Stevens [2] considered 1800 CV and VC syllables, spoken by six (four male and two female) native speakers of English. They developed three templates, each designed to reflect diffuse-falling or flat, diffuse-rising and compact spectral shapes. A 14th order LP filter was used to compute the smoothed spectra of the first 26 ms of the burst of each syllable in the corpus. These spectra were tested against each of the three templates for both correct acceptance and correct rejection. About 85% of CV syllables were correctly accepted or rejected by each template depending on the place of articulation. The results for 450 initial voiced stop consonants ([b, d, g] in the context of vowels [i, e, a, o, u]) reported in table I in [2, p. 1008] are reproduced in table 2.1.

Kewley-Port

Kewley-Port [49] suggested that the invariant acoustic cues might be present in the time-varying spectra, rather than the static spectrum sampled during the burst and the onset of the vowel, as proposed by Blumstein and Stevens. She computed time-varying LP smoothed spectra using a 20 ms or a 25 ms window, starting from the burst and extending into the following vowel. The window was shifted every 5 ms. Time varying spectra of first 40 ms of each token was plotted on a 3-D display.
Kewley-Port proposed three features for the displays; i) Tilt of the spectrum at burst onset, ii) Late onset of low-frequency energy, and iii) Mid-frequency peaks extending over time, based on which the place of articulation was assigned as shown in table I in [49]. A total of 162 CV syllables, spoken by 3 native speakers of English (two male and one female) were analyzed, using a 14th order LP filter for the male speakers and a 12th order LP filter for the female speaker. Four fewer coefficients were used to analyze the voiceless frames; since the spectral sections calculated with fewer coefficients had smoother peaks.

The displays were analyzed by three judges independently and a place of articulation was assigned to each display. A wrong place of articulation was assigned to 46 of the 162 displays by one or more judges. In order to resolve errors which may have occurred due to careless judgement, the judges were given "additional instructions" and asked to analyze the 46 plots again, in a collaborative session. The results in percentage for each place of articulation by vowel context, given in table IV in [49, p. 372] are reproduced here as table 2.2. About 88% of the stop consonants were assigned correct place of articulation in this study.

Although Kewley-Port stressed on the importance of considering time-varying spectral properties to specify the invariant acoustic cues, the time domain did not play any important role in her study. The late onset feature — a direct measure of voice onset time, was used by the judges in only 0.6% of the judgement to assign correct place of articulation [49, p. 328]. The features: (i) tilt of the spectrum and

<table>
<thead>
<tr>
<th></th>
<th>Diffuse-falling template</th>
<th>Diffuse-rising template</th>
<th>Compact template</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correct acceptance</td>
<td>[b] 82.5</td>
<td>[d] 84.0</td>
<td>[g] 86.7</td>
</tr>
<tr>
<td>Correct rejection</td>
<td>[d] 80.7</td>
<td>[b] 86.0</td>
<td>[g] 88.5</td>
</tr>
<tr>
<td>Correct acceptance</td>
<td>[g] 90.0</td>
<td>[b] 88.5</td>
<td>[d] 82.7</td>
</tr>
<tr>
<td>Correct rejection</td>
<td>[d] 82.7</td>
<td>[b] 91.3</td>
<td>[d] 82.7</td>
</tr>
</tbody>
</table>

Table 2.1: Results of Blumstein and Stevens, (1979) for voiced CV syllables.
(ii) mid-frequency peaks, proposed by Kewley-Port are closely related to Blumstein and Stevens’ static spectral templates, and do not take time domain variations into account.

Lahiri et al.

Lahiri et al. [51] studied the time-varying spectral properties of bilabials, alveolars and dentals from Malayalam, French and English for invariant acoustic cues for place of articulation. In a pilot study, they considered the Blumstein and Stevens’ spectral templates to determine if the spectral shape alone was adequate to differentiate bilabials from non-bilabials. They also examined the claim, based on theoretical predictions, that dentals and alveolars share similar spectral properties during the burst (Blumstein and Stevens, [2]; Halle and Stevens, [33]). Lahiri et al. concluded that, the spectral shape during the burst — A) cannot differentiate bilabial stops from dental stops, and B) cannot group dental stops and alveolar stops into one class; contrary to predictions based on phonological and acoustic theories.

Lahiri et al. then computed short-time LP smoothed spectra using a 10 ms window, with a 5 ms shift for each spectral line. Rather than estimating the spectral shape or spectral tilt as Blumstein and Stevens, and Kewley-Port did, they focussed on the relative changes in the distribution of energy from the burst release.

<table>
<thead>
<tr>
<th>Vowel</th>
<th>(N)</th>
<th>b</th>
<th>d</th>
<th>g</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>i</td>
<td>(81)</td>
<td>67</td>
<td>93</td>
<td>56</td>
<td>72</td>
</tr>
<tr>
<td>e</td>
<td>(81)</td>
<td>70</td>
<td>81</td>
<td>89</td>
<td>80</td>
</tr>
<tr>
<td>a</td>
<td>(81)</td>
<td>100</td>
<td>93</td>
<td>100</td>
<td>96</td>
</tr>
<tr>
<td>o</td>
<td>(81)</td>
<td>93</td>
<td>96</td>
<td>100</td>
<td>96</td>
</tr>
<tr>
<td>u</td>
<td>(81)</td>
<td>78</td>
<td>96</td>
<td>100</td>
<td>91</td>
</tr>
<tr>
<td>I</td>
<td>(27)</td>
<td>89</td>
<td>100</td>
<td>100</td>
<td>96</td>
</tr>
<tr>
<td>ε</td>
<td>(27)</td>
<td>100</td>
<td>100</td>
<td>67</td>
<td>89</td>
</tr>
<tr>
<td>æ</td>
<td>(27)</td>
<td>100</td>
<td>100</td>
<td>67</td>
<td>89</td>
</tr>
<tr>
<td>Total</td>
<td>(486)</td>
<td>84</td>
<td>93</td>
<td>87</td>
<td>88</td>
</tr>
</tbody>
</table>

Table 2.2: Results of Kewley-Port, (1983)
to the onset of voicing. Based on further pilot studies, Lahiri et al. showed that the bilabial stop consonants were characterized by either (i) less difference in the energy at low frequencies than at high frequencies, or (ii) same difference in the energy at low and high frequencies, from the burst to the onset of voicing. Dental and alveolar stop consonants were characterized by a greater difference in the energy at low frequencies than at high frequencies. For “low” and “high” frequencies, they chose 1500 Hz and 3500 Hz respectively. The energy measurements were made on a straight line joining the F₂ and F₄ peaks.

The above metric was used in a cross-language study investigating 493 CV syllables spoken by 3 French speakers (298 tokens), 2 Malayali speakers (95 tokens) and 1 English speaker (100 tokens). Lahiri et al. were able to specify the place of articulation correctly for 89% of the bilabials, 94% of the dentals and 92.5% of the alveolars. It should be noted that except for the 198 CV syllables spoken by two French speakers, the rest of the corpus (a total of 295 CV syllables) was analyzed in the pilot experiments to observe and define the invariant acoustic properties reported in their study.

2.3.1 Context dependencies in Earlier Studies

In Blumstein and Stevens [2], most of the errors in specifying the place of articulation for initial stop consonants were in the context of high front vowel [i] and high back vowel [u]. Bilabials in the context of [i] were classified as alveolars and velars. Alveolars in the context of [u] were classified as velars. A large number of velars in the context of [i] and [e] were classified as alveolars. Most of the errors reported in Kewley-Port [49] were also in the context of high front and back vowels. Bilabials in the contexts of [i, e, u] and velars in the contexts of [i, e, æ] were incorrectly identified (see table 2.2). No such context effects were reported by Lahiri et al. [51].

Context dependencies of similar nature were encountered by others in speech perception studies, based on stimuli constructed using tape-splicing techniques (eg., Cole and Scott [21]) and synthetic speech stimuli (eg., Cooper et al. [20], Blumstein and Stevens [3]). Thus, it appears that the acoustic information for place
of articulation is different, in different vowel contexts. This is considered to weaken
the notion of acoustic invariance in speech signals. Many researchers have argued
that it is not possible to extract invariant acoustic cues from the speech signals
(Lindblom [53]; Liberman, Cooper, Shankweiler and Studdert-Kennedy [52]).

2.4 Drawbacks of the Short-Time LP Smoothed Spectra

The burst region of the stop consonants is highly transient. As the closure at the
place of articulation is released, the energy in the speech signal increases. The
characteristics of the resulting plosion and of the ensuing frication noise undergo
a continuous change, till the opening is wide enough to allow the steady state
excitation of the vocal tract resonances corresponding to the vowel sound. The
duration of the transient part can vary from 5 to 100 msec., depending on place
and manner of articulation, language and speaker.

In Section 2.3, it was seen that the typical window lengths used in analyzing
stop consonants vary from 10 to 26 msec. In the LP method, the signal \( s(n) \)
is assumed to be short-time stationary over the analysis window of length, \( N \).
For each window, the short-time spectrum is computed using an all-pole model
(Makhoul, [55]; Markel and Gray, [57]). The smoothed spectrum for a given window
(can be written as

\[
|H(e^{jn2\pi/N})|^2 = \frac{G^2}{|A(e^{jn2\pi/N})|^2} \quad 0 \leq n \leq N - 1,
\]

where

\[
A(e^{jn2\pi}) = 1 + \sum_{k=1}^{p} a_k e^{-jkn2\pi/N}.
\]

In Eqn 2.1, \( G \) is a constant. The LP parameters \( a_k \) are computed such that the total
squared error in approximating the input signal \( s(n) \) with a predicted signal \( \tilde{s}(n) \)
is minimum [57]. The signal spectrum \(|S(e^{j\omega})|^2\) can be represented by \(|H(e^{j\omega})|^2\)
with an arbitrarily small error, by increasing the filter order \( p \).

\[
|S(e^{jn2\pi/N})|^2 = \lim_{p \to \infty} |H(e^{jn2\pi/N})|^2.
\]
The relationship between the signal, error and the LP filter, in the frequency domain can be written as

\[ |S(e^{j\omega})|^2 = \frac{|E(e^{j\omega})|^2}{|A(e^{j\omega})|^2}. \]  

(2.4)

In the present context, it is of interest to see the nature of the above approximation for nonstationary signals and finite values of \( p \). The relationship between the signal spectrum \( |S(e^{j\omega})|^2 \) and the LP spectrum \( |H(e^{j\omega})|^2 \), for finite values of \( p \) can be seen by comparing Eqn 2.4 and Eqn 2.1. \( |E(e^{j\omega})|^2 \) is being modelled by a flat spectrum \( G^2 \). As the all-pole filter order \( p \) is increased, \( |E(e^{j\omega})|^2 \) becomes white noise and \( |S(e^{j\omega})|^2 \) becomes identical to \( |H(e^{j\omega})|^2 \). For finite values of \( p \), \( |E(e^{j\omega})|^2 \) represents the portion of the signal that does not fit the all-pole model. This information is ignored in computing the LP smoothed spectra. The value of \( p \) is chosen based on the number of resonances expected in the signal. For computing the time-varying spectra in speech, usually 10 parameters are used to represent the formant structure, and 4 additional parameters are used to approximate the glottal filter. In the burst region of stop consonants, there is no formant structure, as there are no steady state resonances in the vocal tract. Further, the glottis is associated with the laryngeal action, eg. voicing and aspiration and not the place of articulation. Thus, although a 14\(^{th} \) order LPC filter gives satisfactory results for vowels, it is not ideal for analyzing the burst of stop consonants. Kewley-Port [49] used four fewer coefficients to analyze the frames containing frication, to avoid “rippled peaks” in the spectra. This would further increase the approximations introduced by the LP method. Kewley-Port suggested that the smoother peaks obtained by fewer coefficients were “closer to the underlying fricative spectrum.” This is clearly not valid since, steady state signals like vowels have smooth peaks and to study the nature of stop consonants, their differences from vowels should be preserved in the analysis method.

When overlapping windows are used to compute successive spectra, adjacent spectra are smeared in time. In conclusion, when short-time spectral estimation techniques, which assume an all-pole model for the signal are used to analyze stop consonants, two sources of error can be identified:

1. A smear between the non-stationary burst region and the steady state vowel,
due to overlapping windows.

2. Ignoring the portion of the signal that does not conform to an all-pole model.

Hence, the LP parameters cannot represent the nonstationarities occurring within a window. This property, in fact, is used to detect epileptic spikes in the EEG signals. For a proper choice of the filter order, the LP parameters of the EEG background activity do not reflect the properties of the nonstationary spikes. By monitoring the prediction error, epileptic spikes are detected.

In conclusion, LP smoothed spectra give excellent results when the signals can be modelled by an all pole system. When the nature of the signal deviates from this ideal, as in stop constants, much of the useful information in the error signal is sometimes ignored.

2.5 Summary

Many investigations in the past failed to yield a complete set of necessary and sufficient set of acoustic properties to specify the place of articulation in stop consonants. This is partially due to inadequacies in the signal processing tools used. Three recent investigations for invariant acoustic cues in stop consonants using the LP method were presented. Drawbacks inherent in the LP method for analyzing stop consonants were discussed.
Chapter 3

Smoothed Wigner Distribution

3.1 Overview

In this chapter, the Wigner distribution (WD) is presented as an alternative to short-time spectral analysis of signals that are essentially transient in nature.

Section 2 presents the concept of WD analysis. The literature on its properties and applications is introduced.

Section 3 documents the way in which analytic signals were obtained from real signals in this study.

Section 4 considers the compromise between positivity and enhanced resolution in WD analysis. It is shown that the WD smoothed for positivity is equivalent to a spectrogram. An interpretation of partially smoothed WD is presented.

Section 5 presents the computation of the smoothed WD. A FORTRAN subroutine is included, which requires as inputs: the speech signal, the time domain window, and the frequency domain window; and provides the smoothed WD as output.

Section 6 explains the nature of cross-terms in the WD of multicomponent signals.

Section 7 presents some of the factors to be considered in choosing the smoothing parameters. The effect of partial smoothing is demonstrated and it is shown
that partially smoothed WD is a useful alternative to short-time spectral analysis.

3.2 Introduction

3.2.1 Background

The WD is a powerful signal processing tool. It was first introduced by Wigner for quantum mechanics in 1932 [70], and by Ville for communication theory in 1948 [69]. Over the past few years, the WD has generated much discussion about its potential for the analysis of transient signals. The WD is similar to short-time spectra, in the sense that it describes the time-varying spectral properties of a signal. For a steady state exponential signal, both the WD and the short-time FFT are equivalent as shown in Figures 3.1 and 3.2. For time-varying signals, however, WD provides a much better resolution than a short-time FFT. This is illustrated in Figures 3.3 and 3.4, which show the spectrogram and the WD of a chirp signal, respectively. Both displays have been computed using the same resolution in the frequency domain. Note the wide main lobe in Figure 3.3, due to temporal smear. Figure 3.4 shows that the WD preserves time domain characteristics of the signal better than a spectrogram.

The WD of a signal \( f(t) \) is defined as

\[
W_f(t, \omega) = \int_{-\infty}^{+\infty} f(t + \tau/2)f^*(t - \tau/2)e^{-j\omega \tau} d\tau. \tag{3.1}
\]

For two signals \( f_1(t) \) and \( f_2(t) \), the cross WD is defined as

\[
W_{f_1,f_2}(t, \omega) = \int_{-\infty}^{+\infty} f_1(t + \tau/2)f_2^*(t - \tau/2)e^{-j\omega \tau} d\tau. \tag{3.2}
\]

Some of the useful properties of the WD which make it attractive for transient signal analysis are:

- The WD is a real function of time and frequency.

- The WD gives correct marginal distributions in time and frequency, i. e., integrating the WD at a given time over all frequencies yields the instantaneous
Figure 3.1: Spectrogram of an exponential.

Figure 3.2: Wigner distribution of an exponential.

Figure 3.3: Spectrogram of a chirp.

Figure 3.4: Wigner distribution of a chirp.
power, and integrating at a given frequency over all time yields the power spectral density.

\[
\frac{1}{2\pi} \int_{-\infty}^{+\infty} W_f(t, \omega) d\omega = |f(t)|^2 \\
\int_{-\infty}^{+\infty} W_f(t, \omega) dt = |F(\omega)|^2
\] (3.3) (3.4)

- The WD preserves the shifts in time and frequency domains. For signals shifted in time and/or frequency, the WD is also shifted in time and/or frequency by the same amount.

\[
f_1(t) = f(t + t_1)e^{j\omega_1 t} \\
W_{f_1}(t, \omega) = W_f(t + t_1, \omega + \omega_1)
\] (3.5) (3.6)

- WD of time limited signals is also limited to the same time duration. Similarly, WD of band limited signals is also limited to the same bandwidth.

Unlike conventional spectral estimation methods, WD is periodic in the frequency domain with a period of \(\pi\) rather than \(2\pi\). This makes it necessary to sample the signal at twice the Nyquist rate, or to consider the analytic version of the input signal, in order to avoid aliasing in discrete WD. While the WD is excellent for analyzing monocomponent signals (as shown in Figures 3.2 and 3.4), it has some drawbacks for analyzing multicomponent signals.

- For a given Fourier transform pair \(f(t) \leftrightarrow F(\omega)\), the WD is not always positive for all values of \(t\) and \(\omega\).

- The WD can be non-zero when the signal \(f(t)\) itself is zero. Similarly, the WD can have non-zero contributions for frequencies where the spectrum \(F(\omega)\) is zero. These non-zero contributions have been called "cross-terms" or "interference terms." They occur in the WD because it is bilinear in nature. Consider the signal \(f(t)\) defined as \(f_1(t) + f_2(t)\). The WD of \(f(t)\) can be shown to be

\[
W_f(t, \omega) = W_{f_1}(t, \omega) + W_{f_2}(t, \omega) + 2\Re[W_{f_1,f_2}(t, \omega)].
\] (3.7)
The first two terms in the above equation will be called self-terms since they correspond to the signal itself. The third term is the cross term and its support in the time-frequency plane is generally midway between the self-terms.

The presence of negative regions and cross-terms is frustrating in many applications of WD. It is, however, shown that these drawbacks need not deter one from taking advantage of its useful properties. Higher order distributions with correct marginals and positivity have been proposed in literature [16,17]. Discussion in this thesis will, however, be limited to WD and its applications.

3.2.2 Literature Review

Over the last few years, many advances have been made in applying WD techniques to the processing of non-stationary signals. Some of these, pertinent to the present topic, are summarized in this section.

Properties of WD

The properties of the WD in the context of signal processing were discussed extensively by Claasen and Mecklenbräuker [13] and Flandrin and Escudié [24]. The dual nature in time and frequency of WD and Ambiguity function (AF) were studied by Hlawatsch [36,37]. An excellent tutorial review of the WD and its relationship to other time-frequency distributions was given by Boudreaux-Bartels [7]. Cohen gave a generalized equation for generating all possible bilinear distributions, based on a kernel function [15]. These distributions, known as Cohen's class of generalized bilinear distributions, can be written as

\[ C_f(t, \omega; G) = W_f(t, \omega) * G(t, \omega) \]  

i.e., any desired distribution \( C_f(t, \omega; G) \) can be obtained by convolving WD of the signal \( W_f(t, \omega) \) and an arbitrary kernel function \( G(t, \omega) \). The convolution is in both \( t \) and \( \omega \). \( G \) is related to \( G \) by a two dimensional Fourier transform. If the kernel function \( G(t, \omega) \) is chosen as \( 2\pi \delta(t)\delta(\omega) \), then \( C_f(t, \omega) \) becomes the WD. Thus,
the kernel function also allows the comparison of WD with other distributions in Cohen’s class, with respect to spread in time and frequency. 1

A serious drawback to using WD in signal analysis was in the presence of negative regions and crossterms. Their presence is particularly annoying in speech research, since the spectrogram, a popular time-frequency analysis tool in speech for nearly four decades, does not have this disadvantage. The signals for which the WD is always positive were presented by Hudson in 1974 [40]. These signals are chirp-like in nature, but the WD of most of the signals encountered in real life (speech, biomedical signals, etc.) do not belong to this class. de Bruijn in 1967 [9] and Janssen in 1981 [45] have shown that smoothing the WD in time and frequency with appropriate Gaussian windows will always result in a positive distribution2. For the bilinear distributions of Cohen’s class, the conditions of positivity and correct marginals are incompatible [71,14]. However, in this class of distributions, the WD has been shown to be superior to other distributions in terms of minimum spread in time and frequency [46]. Further, Janssen and Claasen [47] showed that there is no distribution in Cohen’s class, which becomes positive for all signals after smoothing with \( T\Omega < 1 \) and that the WD is the only distribution in this class which becomes positive for all signals after smoothing with \( T\Omega = 1 \). Janssen and Claasen also pointed out that when the signal deviates from the subclass of signals given by Hudson in [40], considerable smoothing was required to get a non-negative distribution. In later sections of this chapter, it is shown that smoothing with \( T\Omega = 1 \) makes WD and spectrograms equivalent, and a smoothing value \( T\Omega \) much less than 1 is adequate in many applications to make meaningful interpretations regarding the nature of the signal.

**Analytic input for WD**

For real signals, WD is symmetric with respect to the \( \omega = 0 \) line in the frequency domain. Consequently, contributions at positive and negative frequencies interact to give rise to cross-terms at \( \omega = 0 \) [13,4]. In applications where the information

---

1See Table 3.1 on page 25 for examples of some kernel functions.

2The smoothing operation is as shown in Equations 3.10, 3.11. In Section 3.4.1, this smoothing was parameterized on the time-bandwidth product \( T\Omega \) of the Gaussian windows used in smoothing so that \( T\Omega = 1 \) results in a positive distribution.
at low frequencies is important, the analytic version of real signal should be used to compute WD [4,5]. Examples of applications in which analytic signals were used to compute WD can be found in [4,5,26,29,42,44,58,60,65]. It is particularly important to consider the analytic signals in computing WD of speech, since the high amplitude cross-terms at $\omega = 0$ not only mask the low frequency information, but also obliterate the low amplitude higher formant structure of speech. In two recent applications of the WD in speech analysis [10,11] and [1], however, the real signal was used in computing the WD. The authors did not comment on the cross-terms at $\omega = 0$ due to the real nature of the signal. Applications of WD, in which analytic versions of speech signals were used can be found in [28,29,65,72].

Cross-terms in WD

For signals with fairly simple structure in the time-frequency plane, it is not difficult to interpret the WD, in spite of cross-terms and negative regions. Examples of this are chirp-like signals with fairly localized support in time and frequency, encountered by Boashash and others in seismic prospecting and oceanography [4,5,6,42]. For speech and many biological signals, however, cross-terms and negative regions pose many problems in making meaningful interpretations regarding the nature of the signal. Even though it has been argued that any kind of smoothing will destroy the desirable properties of WD [18], mild smoothing was found inevitable in order to get interpretable results for many types of signals [25,29,31,59,60,61,72]. Flandrin discussed the nature of cross-terms in multicomponent signals, in general, and suggested some explicit ways of smoothing [25]. Hlawatsch classified cross-terms as inner and outer cross-terms and discussed methods to distinguish cross-terms from self terms [35]. Smoothing WD for positivity eliminates all the cross-terms in the distribution [28]. This, however, results in a distribution similar to a spectrogram. Partial smoothing was used by Flandrin and Martin [58,60], for biological signals and Garudadri, Beddoes, Gilbert and Benguerel [28,29,30,31], Wokurek et al. [72] for speech. Riley [65] and Atlas et al. [1] smoothed the WD for positivity in analyzing speech signals. Such smoothing makes the WD and spectrogram equivalent. Indeed, in their examples of WD plots and spectrograms, the formants can be seen equally well in both WD plots and spectrograms. The WD plots did not display
It was pointed out by Cohen [18] that the WD is not necessarily zero in the regions in time or frequency where the signal itself is zero. It was remarked that these non-zero regions could mask some relevant information about the signal (e.g. silences in speech). It was shown that these regions contain the cross-terms in the time domain, which are smoothed when computing the discrete WD using finite windows.

Applications of WD

The WD has been successfully applied to many signal analysis problems. The WD of the impulse response of loud speakers was found to be useful in characterising them [44]. In studying seismic signals and microstructure of underwater temperature gradients, the WD was found to be a useful tool by Boashash and others [4,6,42]. Martin and Flandrin used WD to study the nature of biological signals by defining generalized spectral estimators [59,61]. The problem of synthesizing the time domain signal from an arbitrary description in the time-frequency plane using the WD was studied by Boudreaux-Bartels [7,8], Yu and Cheng [73], and Krattenthaler et al. [48]. The applications of WD in analyzing somatosensory evoked potentials for early detection of multiple sclerosis, EEG signals, harmonic analysis in power systems etc., are being investigated [32].

3.3 Analytic Signals

In the previous section, the importance of using analytic signals to compute the WD was discussed (see page 20). The analytic version of a real signal \( s(n) \) is defined as

\[
f(n) = s(n) + j\hat{s}(n).
\]

where \( \hat{s}(n) \) is the Hilbert transform of \( s(n) \). A Hilbert transformer of length 112 points and passband from 100 Hz to 4.9 kHz was designed using the IEEE signal processing program EQFIR [41,64]. The program FASTFILT [41] was modified to read the signal \( s(n) \) in blocks of 256 samples and output the complex analytic
signal $f(n)$ directly. For a signal of duration $N$, the output of the FIR filter was of length $N+112$. The first 112 points were rejected when forming the complex signal $f(n)$.

### 3.4 Positivity and Resolution Considerations

From equation 3.1, it can be seen that the signal $f(t)$ is needed from $-\infty$ to $+\infty$ to compute the WD. In practice, the signal will be windowed in the time domain using a window $h(t)$. The windowed WD, known as the pseudo WD (PWD) was given by Claasen and Meklenbräuker [13] as follows:

$$
\tilde{W}_f(t, \omega) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} W_f(t, \xi)W_h(0, \omega - \xi)d\xi.
$$

(3.9)

where $W_h(t, \omega)$ is the WD of window $h(t)$. The above equation is a convolution between $W_f(t, \omega)$ and $W_h(0, \omega)$ in the frequency variable $\omega$. As a consequence, the window introduces bias or smear parallel to the frequency axis only, and the property of correct marginals with respect to time is not lost. Hence, there is no smear parallel to the time axis, even when successive WD spectra are computed with overlapping windows.

In most applications, the presence of cross-terms and negative regions tends to mask the underlying nature of the signal. In practice, it is necessary to apply some sort of smoothing in the time and frequency domains in order to suppress the negative regions and the cross-terms [25,29]. The smoothing process is a 2-D convolution of the WD with a function $G(t, \omega)$ given by

$$
\tilde{W}_f(t, \omega) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} W_f(\tau, \xi)G(t - \tau, \omega - \xi)d\tau d\xi.
$$

(3.10)

deBruijn showed that using $G(t, \omega)$ defined as

$$
G(t, \omega) = \frac{1}{T\Omega} e^{-t^2/T^2} e^{-\omega^2/\Omega^2}
$$

(3.11)

with $T\Omega \geq 1$ yields a positive WD [9]. $G(t, \omega)$ acts as a low pass filter in time and frequency, and filters out the oscillating cross-terms. Here, $T$ and $\Omega$ define the amount of smear in the time and frequency domains respectively. As the range of smoothing is increased by increasing the values of $T$ and $\Omega$, the amplitude of
the cross-terms decreases. For $T\Omega \geq 1$, there are no cross-terms in the WD and positivity can be assured. The positivity of WD smoothed with $T\Omega \geq 1$ was shown by de Bruijn [9] and Janssen [45] using Hermite polynomials. A simpler proof, based on the fact that a smoothed WD computed with $T\Omega = 1$ and a spectrogram computed using a Gaussian window are identical, is now presented.

### 3.4.1 Positivity of a Smoothed WD

Consider the spectrogram of the signal $f(t)$, given by

$$S_f(t, \omega) = \left| \int_{-\infty}^{+\infty} f(\tau) h(t - \tau) e^{-j\omega \tau} d\tau \right|^2. \quad (3.12)$$

Here, $h(t)$ is the time domain window used in computing the short-time Fourier transform. The spectrogram is related to the WD by a 2-D convolution as shown in Equation 3.13 [13].

$$S_f(t, \omega) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} W_f(\tau, \xi) W_h(t - \tau, \omega - \xi) d\tau d\xi \quad (3.13)$$

where $W_h(t, \omega)$ is the WD of the window $h(t)$ given by

$$W_h(t, \omega) = \int_{-\infty}^{+\infty} h(t + \tau/2) h^*(t - \tau/2) e^{-j\omega \tau} d\tau. \quad (3.14)$$

For a Gaussian window $h(t)$, defined as

$$h(t) = \frac{1}{2\sqrt{\pi T}} e^{-t^2/2T^2} \quad (3.15)$$

Eqn 3.14 can be shown to be

$$W_h(t, \omega) = e^{-t^2/T^2} e^{-\omega^2T^2}. \quad (3.16)$$

If we choose $T\Omega = 1$ in Equation 3.11, the smoothing function in Equation 3.10 and the spectrogram in Equation 3.13 will be identical to each other. Thus, the WD smoothed according to Equation 3.10 with $T\Omega = 1$ is identical to a spectrogram computed using a Gaussian window. Since the spectrogram is a positive function, we conclude that smoothed WD is also positive for $T\Omega = 1$. Spectrograms computed using other windows (Hamming, Hanning etc.) are similar WD plots smoothed with $T\Omega > 1$, since Gaussian windows have the smallest time-bandwidth product.
### 3.4.2 Interpretation of the Smoothed WD

An interpretation of the PWD for monocomponent signals was given by Flan-drin [27]. Since the WD of monocomponent signals do not have any cross-terms, there is no necessity for smoothing. For multicomponent signals, partial smoothing can be viewed as a compromise between the desired resolution and cross-term suppression in the distribution. Such a compromise sacrifices both the properties correct marginals and positivity, to some extent. An interpretation of the partially smoothed WD is considered here. An insight into the smoothed WD of multicomponent signals like speech can be gained by considering Cohen’s class of generalized bilinear time-frequency distributions [15]. All members in Cohen’s class can be described in terms of the WD and a kernel function as shown in Equation 3.10. By choosing different functions for $G(t, \omega)$, different time-frequency representations of Cohen’s class can be obtained. A few of the choices for $G(t, \omega)$, relevant to this discussion, are shown in Table 3.1. The function $G(t, \omega)$ is plotted for a pseudo-WD, for a narrowband spectrogram, a wideband spectrogram, and a smoothed WD in Figures 3.5 – 3.8.

For the pseudo-WD, it can be seen from Figure 3.5 that the kernel $G(t, \omega)$ does not introduce any temporal smear. For a narrowband spectrogram, on the other hand, the kernel produces substantial smear in the time domain, as shown in Figure 3.6. For a spectrogram, choice of $T$ (or $\Omega$), implies the choice of $\Omega$ (or $T$) based on Heisenberg’s inequality. Thus, reducing $T$ for better time resolution in turn increases the frequency domain smear, as shown in Figure 3.7. The kernel for a

<table>
<thead>
<tr>
<th>$G(t, \omega)$</th>
<th>Representation</th>
</tr>
</thead>
<tbody>
<tr>
<td>$2\pi \delta(t) \delta(\omega)$</td>
<td>Wigner distribution</td>
</tr>
<tr>
<td>$W_0(0, \omega)$</td>
<td>Pseudo Wigner distribution</td>
</tr>
<tr>
<td>$W_n(t, \omega)$</td>
<td>Spectrogram</td>
</tr>
<tr>
<td>$\frac{1}{\tau_n} e^{-t^2/\tau^2 - \omega^2/\Omega^2}$</td>
<td>Smoothed Wigner distribution</td>
</tr>
</tbody>
</table>

Table 3.1: Examples of some kernel functions and the resulting representations
Figure 3.5: $G(t, \omega)$ for a pseudo-WD.

Figure 3.6: $G(t, \omega)$ for a narrowband spectrogram.

Figure 3.7: $G(t, \omega)$ for a wideband spectrogram.

Figure 3.8: $G(t, \omega)$ for a smoothed WD with $T\Omega = 0.25$. 
smoothed WD with \( T \Omega = 0.25 \) is shown in Figure 3.8. The option of specifying both \( T \) and \( \Omega \) explicitly and independent of each other makes the partially smoothed WD an attractive tool. Since WD has correct marginals, it may be viewed as the perfect time-frequency description of the signal. The kernel function may be considered as an aperture through which the signal is being observed. As the kernel deviates from \( \delta(t)\delta(\omega) \), the WD is smeared in time and frequency. While the area of this aperture is constant, and defined by uncertainty relations in Fourier analysis, it can be explicitly defined in the partially smoothed WD.

Besides introducing a smear in the time and frequency domains, smoothing causes a loss of the phase information. This can be seen from Example I on page 30, where the phase terms \( \phi_i \) and \( \phi_k \) are present only in the cross-term in Equation 3.25. Thus, smoothing for positivity eliminates all the phase information in the WD; however, partially smoothed WD has proved to be a very useful tool. A priori information about the nature of the signal is often helpful in choosing the smoothing parameters.

### 3.5 Computing the Smoothed WD

In practice, large computational savings can be achieved by computing the smoothed WD directly, rather than by computing the WD and then performing the 2-D convolution for smoothing. The formulation for computing the partially smoothed WD is given below:

By choosing a separable smoothing function such as in Equation 3.11, Equation 3.10 can be rewritten as:

\[
\tilde{W}_f(t, \omega) = [W_f(t, \omega) \ast G(\omega)] \ast g(t) \tag{3.17}
\]

where \( G(\omega) = \frac{1}{\pi} e^{-\omega^2/\alpha^2} \) and \( g(t) = \frac{1}{T} e^{-t^2/T^2} \). The bracketed term is the PWD in Equation 3.9 computed using a time window \( h(t) \) such that

\[
W_h(0, \omega) = G(\omega). \tag{3.18}
\]

Substituting Equation 3.14 into Equation 3.18 and taking the inverse Fourier transform on both sides, we have

\[
h^2(\tau/2) = \sqrt{\pi} e^{-\pi^2(\tau/2)^2}. \tag{3.19}
\]
We can now rewrite Equation 3.17 as

$$\tilde{W}_f(t, \omega) = \tilde{W}_f(t, \omega) * g(t)$$  \hspace{1cm} (3.20)$$

where

$$\tilde{W}_f(t, \omega) = \int_{-\infty}^{+\infty} f(t + \tau/2)f^*(t - \tau/2)h^2(\tau/2)e^{-j\omega\tau} d\tau.$$  \hspace{1cm} (3.21)$$

Substituting for $g(t)$ in Equation 3.20 and interchanging the order of integration, we have

$$\tilde{W}_f(t, \omega) = \int_{-\infty}^{+\infty} \left[ \int_{-\infty}^{+\infty} f(t + \tau/2 - \tau')f^*(t - \tau/2 - \tau')g(\tau') d\tau' \right] h^2(\tau/2)e^{-j\omega\tau} d\tau.$$ \hspace{1cm} (3.22)$$

The bracketed term is the convolution of $f(t + \tau/2)f^*(t - \tau/2)$ with $g(t)$. This term is windowed using $h^2(\tau/2)$, and its Fourier transform computed to yield the smoothed WD at time $t$. In discrete form, Equation 3.22 is given by:

$$\tilde{W}_f(n, m) = 2 \sum_{k=-K}^{K-1} \sum_{i=-(I-1)/2}^{(I-1)/2} f(n + k - i)f^*(n - k - i)g(i) h^2(k)e^{-jk\omega n/2M}$$  \hspace{1cm} (3.23)$$

where $M = 2K$. This equation is the same as the one given by Flandrin [25,26], except that the smoothing windows have been chosen as Gaussian and defined in terms of $T$ and $\Omega$. This enables one to specify the amount of smoothing in a smoothed WD, relative to that of a spectrogram.

Figure 3.9 shows an implementation of Equation 3.23 for analytic signals as a FORTRAN callable subroutine WIG. The input arrays F, GT and GF are the input signal, time window and frequency window respectively. Array F should be dimensioned at least IT + IF in the calling routine where IT and IF are the lengths of the arrays GT and GF respectively. A cyclic shift of $(IF - 1/2)$ points is performed on GF before passing it to WIG. Each time WIG is called, a WD of length IF points corresponding to the time point $(IF + IG - 1)/2$ is returned in RL. WIG calls the subroutine FFT described in [64]. This may be replaced by a call to a local FFT program for faster execution.
F(IF+IT) : Input /Output data Array.
GT(IT) : Time smoothing window. IT is odd.
GF(IF) : Freq. smoothing window. IF is power of 2

DO 10 loop performs time smoothing. DO 20 and DO 30 loops form the product for FFT computation. IPOWER(IF) returns IZ such that IF=2**IZ. IZ is used in CALL FFT.

SUBROUTINE WIG(F, GT, GF, IF, IT, RL)
COMPLEX F(1), S(256)
REAL GT(1), GF(1), RL(1)

IFBY2 = IF / 2
IF2M1 = IFBY2 - 1
IF2P1 = IFBY2 + 1
IZ = IPOWER(IF)

DO 20 K = 1, IF2P1
KPLUS = IF2P1 + K - 1
KMINUS = IF2P1 - K + 1
S(K) = (0.,0.)
DO 10 I = 1, IT
10 S(K) = S(K) + F(KPLUS + I - 1) * CONJG(F(KMINUS + I - 1)) * GT(I)
20 S(K) = 2. * S(K) * GF(K)
DO 30 I = 1, IF2M1
30 S(IF2P1 + I) = CONJG(S(IFBY2 - I + 1))

CALL FFT(S, IZ, IF)
DO 40 I = 1, IF
40 RL(I) = REAL(S(I))

RETURN
END

Figure 3.9: Subroutine for computing the WD
3.6 Nature of Cross-terms

An understanding of the nature of cross-terms is essential if they are to be handled successfully and if useful information is to be extracted from the WD. Generally speaking, cross-terms oscillate in the time-frequency plane and occur midway between the self-terms. Unfortunately, analytic evaluation of these terms is lengthy and cumbersome [19]. An insight into the nature of the cross-terms, however, can be gained by considering simple cases. Three tutorial examples are considered here to provide such insight. The approach adopted is to express the signal in terms of simple signals for which the cross-terms can be readily written.

Example I: \( f(t) \) in terms of harmonics

Consider a quasi-stationary analytic signal \( f(t) \), with the duration of stationarity at least \( T' \) seconds. \( f(t) \) can be expanded in terms of its Fourier coefficients as

\[
 f(t) = \sum_{i=0}^{n} A_i e^{j\omega_i t + j\phi_i} \quad t_1 \leq t \leq t_1 + T'
\]

(3.24)

where \( n \) can be made sufficiently large to make the right hand side arbitrarily close to \( f(t) \). The WD of this signal can be shown to be

\[
 W_f(t, \omega) = 2\pi \sum_{i=0}^{n} A_i^2 \delta(\omega - \omega_i) \\
 + 4\pi \sum_{i>k} A_i A_k \cos((\omega_i - \omega_k)t + (\phi_i - \phi_k)) \delta(\omega - (\omega_i + \omega_k)/2)
\]

(3.25)

The first sum in the above equation corresponds to self terms the direct contributions which are always positive. The second sum corresponds to the cross-terms due to the components at frequencies \( \omega_i \) and \( \omega_k \). These cross-terms occur midway between \( \omega_i \) and \( \omega_k \), and are modulated parallel to the time direction, at a frequency \( \omega_i - \omega_k \). These terms will be called “frequency domain cross-terms.”

The frequency domain cross-terms are suppressed by smoothing parallel to the time direction. The term \( \frac{1}{T} e^{-t^2/T^2} \) in Equation 3.11 performs this smoothing. As the value of \( T \) is increased, the amount of suppression increases.
Example II: \( f(t) \) in terms of impulses

Consider a time limited signal

\[
f(t) = \sum_{i=0}^{n} A_i \delta(t - t_i).
\]

Self-terms in the WD of \( f(t) \) are of the form \( 2\pi \delta(t - t_i) \) for any \( \omega \). The cross WD due to the terms at \( t_i \) and \( t_j \) can be written as

\[
W_{-i,j}(t, \omega) = A_i A^*_j \int_{-\infty}^{+\infty} [\delta(t + \tau/2 - t_i) \delta(t - \tau/2 - t_j)]e^{-j\omega \tau} d\tau.
\]  \hspace{1cm} (3.26)

The bracketed term is a function of both \( t \) and \( \tau \). Solving for \( t \) and \( \tau \) at \( t + \tau/2 - t_i = 0 \) and \( t - \tau/2 - t_j = 0 \), the above can be rewritten as

\[
W_{-i,j}(t, \omega) = A_i A^*_j \int_{-\infty}^{+\infty} [\delta(t - t_i + t_j/2) \delta(\tau - (t_i - t_j))]e^{-j\omega \tau} d\tau
\]  \hspace{1cm} (3.27)

\[
= A_i A^*_j \delta(t - t_i + t_j/2)e^{-j\omega(t_i - t_j)}
\]  \hspace{1cm} (3.28)

\[
2\Re[W_{-i,j}(t, \omega)] = 2|A_i A^*_j| \delta(t - t_i + t_j/2) \cos(\omega(t_i - t_j)).
\]  \hspace{1cm} (3.29)

Hence, the cross-term due to components at \( t_i \) and \( t_j \) occurs at \( (t_i + t_j)/2 \) and is modulated parallel to the frequency direction. The frequency of modulation is a function of \( t_i - t_j \). Cross-terms of this nature will be called “time domain cross-terms.”

This example is useful for understanding the nature of non-zero terms in the WD, when the signal is zero. It will be used to counter the argument given by Cohen [18] that these cross-terms mask useful information in the signal. Equation 3.29 shows the structure of these cross-terms. The time domain cross-terms are suppressed by the term \( \frac{1}{n}e^{-\omega^2/n^2} \) in Equation 3.11. In Section 3.4.1, it was seen that this smoothing was inherently present in computing WD. The frequency response of the lowpass filter in this smoothing is \( G(t, \omega) \) at \( t = 0 \) in Figures 3.5 and 3.8.

Example III: General signals

The classification of cross-terms as time domain and frequency domain cross-terms has been adopted here simply as a convenience. For a given signal, the cross-
terms can be modulated parallel to both time and frequency directions, simultaneously. This can be seen from the example considered by Boudreaux-Bartels [7] and Hlawatsch [35].

Let \( h(t) \) be a signal concentrated at \((0,0)\). The cross WD of the signals \( f_1(t) \) and \( f_2(t) \), defined as

\[
\begin{align*}
    f_1(t) &= A_1 h(t - t_1) e^{j\omega_1 t} \\
    f_2(t) &= A_2 h(t - t_2) e^{j\omega_2 t}
\end{align*}
\]

is given by

\[
W_{f_1,f_2}(t,\omega) = |A_1 A_2^*| e^{j[(\omega - \omega_c) t_d - (t - t_c) \omega_d]} W_h(t - t_c, \omega - \omega_c)
\]  

(3.30)

where

\[
\begin{align*}
    \omega_c &= \frac{\omega_1 + \omega_2}{2}; \\
    \omega_d &= \omega_1 - \omega_2 \\
    t_c &= \frac{t_1 + t_2}{2}; \\
    t_d &= t_1 - t_2
\end{align*}
\]

Thus, the cross-terms due to \( f_1 \) and \( f_2 \) occur at their geometric center; the frequency of modulation parallel to the time direction is a function of \( \omega_d \) and that parallel to the frequency direction is \( t_d \).

### 3.7 Choice of the Smoothing Parameters

The smoothing parameters \( T \) and \( \Omega \) determine the amount of cross-term suppression parallel to the time and frequency directions, respectively. In a digital implementation, it is necessary to truncate the Gaussian windows in Equation 3.11. Since the Fourier transform of a time-limited signal cannot be band-limited, the spectra of the windows exist for all frequencies. The smoothing operation shown in Equation 3.10 is essentially a spatial low pass filtering operation, in both time and frequency directions. Hence, it follows that it is not possible to generate a positive distribution using truncated windows, even with \( T \Omega \geq 1 \). WD smoothed for positivity is not of much interest, since the resolution is lost. By applying partial smoothing over small regions, it is possible to limit the spread of components...
in time and frequency to regions defined by $T$ and $\Omega$. The factors considered in choosing the values for $\Omega$ and $T$ for computing the smoothed WD of speech signals are discussed below. They should also be applicable to other multicomponent signals.

### 3.7.1 Choice of $\Omega$

The length $IF$ of the window $GF$, in Figure 3.9 is normally chosen as a power of 2 for efficient computation. For a 60 dB sidelobe suppression, $IF$ should span at least 6 times the standard deviation of the window [34]. This is achieved by choosing $\Omega' = IF/3$, where $\Omega'$ is $\Omega$ in number of samples. For this choice of $\Omega$, the effective smear is given by $\Omega = \frac{\Omega'}{2^n}$, where $f_s$ is the sampling frequency.

**Spectral Wrap**

The periodicity of the WD in the frequency domain is $\pi$ rather than $2\pi$. Because of this, a peculiar phenomenon called "spectral wrap" occurs in the WD [30]. The effect of spectral wrap is to make very high frequencies appear at very low frequencies and vice versa. The way spectral wrap occurs in the WD is illustrated in Figure 3.10. The range of smear in the frequency domain is shown schematically by the window (perpendicular to paper). Because of the finite width of the main lobe of this window, components that are close to 0 or $\pi$ spill over to the other end of the spectrum. Spectral wrap can be minimized by making the width of the main lobe as small as possible. This is achieved by increasing the value of $IF$ at the cost of an increase in computation. For a given value of $IF$, the width of the main lobe can be further reduced by making $\Omega' > IF/3$. This, of course, causes an increase in the relative amplitude of the side lobes. Figure 3.11 illustrates the spectral wrap in a chirp signal.

In this study, speech signals were sampled at 10 kHz. For this sampling frequency, $IF = 128$ and $\Omega' = IF/3$ were found to keep spectral wrap to a satisfactory level. The effective smear $\Omega$ in the frequency was 117 Hz. In the next section, it is demonstrated that for this choice of $\Omega$, the time domain cross-terms did not mask the presence of silence durations in the signal.
Figure 3.10: Cause of spectral wrap in a pseudo-WD.

Figure 3.11: Spectral wrap in a chirp signal.
3.7.2 Choice of $T$

From Equation 3.25, it can be seen that the cross-term of two signals that are $(\omega_i - \omega_k)$ apart is modulated parallel to time with a frequency of $(\omega_i - \omega_k)$. To eliminate this cross-term, the time domain smoothing $T$ should extend to more than one period of $(\omega_i - \omega_k)$. As the frequency $(\omega_i - \omega_k)$ decreases, the value of $T$ should be increased for reasonable cross-term suppression. In the limiting case when $(\omega_i - \omega_k) = \Omega$, the value of $T$ is $1/\Omega$ such that $T\Omega = 1$. In this situation, $\Omega$ can be further reduced by choosing a longer window for GF, so that satisfactory cross-term suppression ensured with $T\Omega < 1$. Figure 3.12 demonstrates frequency domain cross-term suppression in two chirps of opposite slopes. The smoothed WD was computed with $T\Omega = 0.06$. As the spectral separation $(\omega_i - \omega_j)$ between self-terms decreases, the amplitude of cross-terms increases, indicating the need for larger values of $T$ for satisfactory cross-term suppression.

In the present work, the value of $T$ was based on the choice of $\Omega$ and the desired value of $T\Omega$. After experimenting briefly with various values, $T\Omega = 0.25$ was chosen for the entire corpus. For the choice of $\Omega$ as selected in the previous section, this corresponded to $T = 2.1$ msec. or 21 sample points. This choice was judged satisfactory since the duration of the burst in stop consonants generally varies from 5 to 40 msec. Ideally, a value of $T$ that is smaller than $T'$ in Equation 3.24 is desired. The length of the Gaussian window $G_T$ in Figure 3.9 was chosen as $1T = 2T$ rather than $3T$, in order to save computational costs. For the WD plots with $T\Omega = 0.25$, the time domain resolution is roughly equal to that of a narrowband spectrogram and the frequency domain resolution is roughly equivalent to that of wideband spectrogram.

3.7.3 Cross-term Suppression in Smoothed WD

It is frequently argued in the literature that the presence of negative regions and cross-terms limit the usefulness of WD for multicomponent signals. In the course of this work, the WD was applied extensively to synthetic and real multicomponent signals. As a result, it was concluded that the problem was not as grave as it is usually portrayed to be (see [18]). A few examples are presented here to illustrate
Figure 3.12: Effective cross-term suppression as the spectral distance between self-terms decreases.
the effectiveness of partial smoothing.

In many situations, surprisingly little smoothing is necessary to bring out the structure of a signal. The smoothing in Figure 3.12 occurs over 5 sample points or 0.5 msec. This smoothing can be seen to be adequate when the self-terms are more than 2 kHz apart. For most multicomponent signals, the frequency domain cross-terms can completely mask the signal structure. This is illustrated in Figure 3.13, depicting the pseudo-WD of the syllable [de] in English. Two versions of the WD of this syllable, smoothed with $T \Omega = 0.25$ and $T \Omega = 1$ are shown in Figures 3.14 and 3.15. Figure 3.15 is equivalent to a spectrogram. The formant structure can be seen clearly in both 3.14 and 3.15. Figure 3.14, however, provides much better temporal resolution than its spectrographic equivalent.

The time domain cross-terms are generally easier to handle than the frequency domain cross-terms because of the inherent smoothing present in the computation. In a recent paper, Cohen [18] pointed out that the WD has non-zero contributions when the signal is zero. He implied that these contributions might mask important information about the signal, for example, silences in speech. In light of this, a closer examination of the time domain cross-terms, viz-a-viz the seriousness of these objections, is in order.

The time domain cross-terms are suppressed depending on the choice of $\Omega$ (or $\text{IF}$ in Figure 3.9). From examples II and III on page 31, it can be seen that the frequency of time domain cross-terms increases as the (temporal) distance between the self-terms increases. As the distance between the self-terms increases, the cross-term due to them is attenuated more. The effect of these terms is most pronounced when the self-terms are separated by a very short silence. In these situations, the WD is a better choice, since the temporal detail is rather poor in a spectrogram. This is illustrated in Figures 3.16 and 3.17, which depict the pseudo-WD and the spectrogram of a synthetic signal with a small silence duration. It is clear that the presence of a silence can be easily identified in the WD.

It should also be noted that the time domain cross-terms appear only when the silence duration is smaller than the window length $\text{IF}$. This has some important implications in the WD analysis of speech signals since silences occur in speech for
Figure 3.13: WD of English phone [de] with no smoothing. $T\Omega = 0.01$
Figure 3.14: WD of English phone [de] with partial smoothing. $T\Omega = 0.25$
Figure 3.15: WD of English phone [de] with complete smoothing. $T\Omega = 1.0$
Figure 3.16: Pseudo WD of two chirps showing negligible cross-terms during silence.
Figure 3.17: Smoothed WD of two chirps. $T\Omega = 1.0$. 
many different reasons, for example:

- Silences maybe due to breath pauses which are used to mark grammatical boundaries in speech are usually greater than 50 msec. in duration.

- Silences may also occur because of particular phonetic sequences in a given speech segment. An example of this is the closure of voiceless stop consonants, such as the sound to [p] in words like "speech" and "april." These silences are usually of the order of 100 ms or more.

These silences do not produce any contributions in the WD, since the windows used in speech are generally much shorter.

- Silences that are shorter than the window length occur in speech between the burst and the onset of the vowel in stop consonant-vowel sequence. They vary from 5 to 50 ms.

A visual feature related to this duration, called "formant onset duration" (FOD), is defined on page 58. Figures 3.18 and 3.19 show FOD in an alveolar and a velar stop consonant. It can be seen that the time domain cross-terms do not mask the presence of silence during FOD. Results presented in Chapter 5 establish that the FOD carries useful information regarding the place of articulation in stop consonants, thus making it clear that the time domain cross-terms do not inhibit extraction of useful information from the WD of speech.

### 3.7.4 Other Considerations Specific to Speech

Similar to spectrographic analysis it was necessary to pre-emphasize the speech signal prior to computing the WD. In order to observe the higher formants in the WD, some form of compression (e.g., logarithmic) of the data had to be performed. The compression equation used was:

$$ W(t, \omega) = 10 \log_{10} [W'(t, \omega) \ast \lambda + 1] $$  \hspace{1cm} (3.31)

where $W'(t, \omega)$ is the half rectified WD, normalized with respect to the maximum value of the WD. The term $\lambda$ was the scaling factor. In order to keep the logarithmic datum as zero, unity was chosen as an offset.
Figure 3.18: Smoothed WD of an alveolar depicting negligible cross-terms during FOD.
Figure 3.19: Smoothed WD of a veiler depicting nebulosity cross-terms during PPD.

Page 332: #640/210/65-576.10 #I265/612 #I7/21/52 #204 #74185
3.8 Summary

In this chapter, the concept of WD was introduced and a brief review of the literature dealing with properties and applications of WD was presented. The cross-terms in WD mask the underlying nature of speech signals. It was shown that smoothing the cross-terms completely makes WD and spectrogram equivalent. Partially smoothed WD was found to be a useful alternative in many applications. The extent of partial smoothing was parameterized on $T\Omega$ by using truncated Gaussian windows. An interpretation of the partially smoothed WD and its computational aspects, was presented. The nature of cross-terms was discussed in detail, as were factors involved in choosing the smoothing parameters. It was shown that it is possible to obtain a better time-frequency resolution (or localization) by using partially smoothed WD, than by using short-time spectral analysis methods.
Chapter 4

Study of Acoustic Invariance: A Cross-Language Study

In Chapter 2, it was suggested that, in earlier investigations, the analysis methods were partly responsible for the context dependency in specifying invariant cues for place of articulation. In Chapter 3, it was demonstrated that the partially smoothed WD overcomes the drawbacks of short-term spectral analysis techniques. The next task was to determine whether the WD would be more successful in specifying invariant acoustic cues for place of articulation in stop consonants, with less context dependency. A cross-language study was conducted to establish this. The experimental procedure is presented in this chapter.

4.1 Overview

Section 4.2 introduces some of the linguistic terms used throughout this dissertation.

Section 4.3 presents the invariant patterns observed in the WD plots of the pilot study. These patterns were compared with the invariant acoustic cues proposed in the literature [2,49,51].

Section 4.4 concerns the program FEATURE, which interactively obtains data regarding the invariant patterns of a WD plot, and assigns a place of articulation to it.
Section 4.5 deals with the final experiment, in which three judges visually identified the invariant patterns in the WD plots. The program FEATURE assigned a place of articulation to each WD plot based on the response of the judges.

4.2 Preliminaries

4.2.1 Stop Consonants

Stop consonants can be classified articulatorily in terms of (i) the vocal tract parameter, place of articulation and the source parameters, voicing and aspiration. Formal definitions of the place of articulation, voicing and aspiration can be found in Peterson and Shoup [63]. These terms are described below:

**Place of Articulation** is the place along the vocal tract, where the airflow is stopped completely for a brief period and suddenly released, in articulating the stop consonant. The stop consonants used in this study have four different places of articulations:

- **Bilabial** — airflow is stopped at the lips, as in the phones [p, b].
- **Dental** — airflow is stopped behind the teeth as in the phones [t, d].
- **Alveolar** — airflow is stopped at the alveolar ridge, as in the phones [t, d].
- **Velar** — airflow is stopped at the velum, as in the phones [k, g].

**Voicing** occurs when the vocal folds are set into a quasi-periodic motion during articulation, giving rise to a low frequency resonance in the vocal tract. This low frequency resonance in a time-varying spectral representation of the speech signal is sometimes called the “voice bar.” For stop consonants, the duration between the instant of release and onset of voicing is called “Voice Onset Time” (VOT) [54]. If the onset of voicing occurs prior to or at the

---

1The diacritic “_” will be used instead of “’” to represent dental stops throughout this thesis.
release, VOT has a negative or 0 value, and the stop consonant is said to be voiced. If the onset of voicing occurs at least 15 ms after the release or after the release, VOT has a positive value and the stop consonant is said to be voiceless.

**Aspiration** occurs when a turbulence noise source at the glottis excites the vocal tract, following the release. This excitation usually lasts for at least 30 ms after the release [63, p. 59].

"Invariant acoustic cues for place of articulation in stop consonants" can be defined as those unique properties of a speech segment corresponding to a given place of articulation, which can be extracted from the signal using signal processing methods. Given these properties, it should be possible to determine the place of articulation in a speech segment using signal processing methods. It is a common belief that some form of these properties is used by the auditory system in transducing speech signals.

### 4.2.2 Organization of Stop Consonants in English, French and Telugu

Stop consonants used in this study were spoken by native speakers of three different languages — English, French and Telugu. (Telugu is a Dravidian language, spoken by about 50 million people in the state of Andhra Pradesh, India.) These languages were chosen because of their different linguistic origins. A brief description of the organization of stop consonants in these languages follows:

For a given place of articulation, there are four different combinations of source parameters: voicing and aspiration, that can excite the vocal tract. Not all combinations of these source parameters are utilised by the native speakers of a given language. For example, voiced aspirated stop consonants occur in Telugu, but not in English and French. In addition, native speakers of different languages group the consonants with similar articulatory features into different phonemic categories. The stop consonants of English, French and Telugu are described below in terms of articulatory parameters.
English stop consonants can be grouped into two different phonemic categories — /p, t, k/ and /b, d, g/. These groups are contrasted primarily by aspiration and secondarily by voicing. /p, t, k/ are aspirated and voiceless while /b, d, g/ are unaspirated. Usually, /b, d, g/ have no pre-voicing, though some English speakers tend to pre-voice these stops. Phonemically, the above two groups in English have been termed voiceless and voiced stops. There are three places of articulation in English: bilabial /p, b/, alveolar /t, d/ and velar /k, g/.

French stop consonants can be grouped into two different phonemic categories — /p, t, k/ and /b, d, g/. But unlike in English, the primary contrast is voicing and the secondary contrast is aspiration. /p, t, k/ have a positive VOT while /b, d, g/ have a negative VOT, i.e., they are prevoiced. Aspiration in French depends on the syllable context. /p, t, k/ are unaspirated in the word initial position and aspirated in the word final position while /b, d, g/ are never aspirated. Phonemically, these two groups in French are also termed as voiceless and voiced stops. There are three places of articulation in French: bilabial /p, b/, dental /t, d/ and velar /k, g/.

Telugu uses all the four modes of source excitation [50]. Thus, the stops of Telugu are grouped in four different phonemic categories — voiceless unaspirated /p, t, t, k/, voiceless aspirated /pʰ, tʰ, tʰ, kʰ/, voiced unaspirated /b, d, d, g/, and voiced aspirated /bʰ, dʰ, dʰ, gʰ/. Telugu has four places of articulation; bilabial /p, pʰ, b, bʰ/, dental /t, tʰ, d, dʰ/, post-alveolar retroflex /tʰ, tʰ, d, dʰ/ and velar /k, kʰ, g, gʰ/. Retroflexed alveolar stops are produced with the tip of the tongue curled backwards, while making the constriction at the alveolar ridge. For brevity, these sounds will be referred to as Telugu alveolars, throughout this dissertation.

Dentals, alveolars, and retroflexed post-alveolars, etc. are referred to as coronals [12]. Table 4.1 summarizes the organization of word initial stop consonants in these three languages. The signs ‘+’ and ‘−’ before voicing and aspiration denote their presence or absence in the realization of that phoneme. It should be noted that the

---

2The English informants who participated in this study had no such pre-voicing.
voiced stops of English are articulatorily similar to the voiceless stops of French and the voiceless unaspirated stops of Telugu.

4.3 Invariant Patterns in the Wigner Plots

4.3.1 Pilot Study

In the pilot study, three native speakers of English, French and Telugu (JG, AP and HG respectively), served as informants. The corpus for this study contained each of the consonants shown in Table 4.1, except Telugu dentals, followed by the vowels [i a u]. This gave 18 syllables each from English and French and 36 syllables from Telugu. The informants were provided with the list of syllables on a type written page. They were requested to pause before each CV syllable in order to capture each syllable in isolation. Two sets of the CV syllables were recorded and the second list was chosen for digitization.

The analog recordings were made in a sound proof booth on a Revox tape recorder. An AKG microphone was used for recording the pilot data. A monotone signal at 120 Hz was played through a pair of headphones and the subjects were
requested to try and maintain constant pitch as given by the signal. The data was
digitized using a 12 bit A/D convertor. The WD of the discrete data was computed
and plotted. Further hardware and software details of this setup are presented in
appendix A.

The WD plots of the pilot data were examined for unique patterns, for each
place of articulation. The invariant patterns observed were similar to those reported
in the literature [2,49,51] in many respects, with some essential differences. In this
section, these invariant patterns are described and compared with those reported
in literature.

4.3.2 Feature Definitions

In order to see how well the invariant patterns of the WD plots characterized a
given place of articulation, a set of visual features of the WD plots were defined;
these features are described below.

Vowel Pattern

The vowel pattern was characterized by 2 to 5 peaks distributed in frequency
and extending in time. These peaks correspond to the formants of the steady
state vowel. The formant structure was connected by a wide spectral line, at the
beginning of every pitch period. In Figure 3.14, the WD of a steady state vowel
can be seen from time 24 ms onwards.

Spectral Shape

The spectral shape of the burst region was broadly classified into three classes.

a) Compact Pattern

The compact pattern was characterised by the presence of a single prominent
spectral peak above 700 Hz. The compact pattern was sometimes character-
ized by several successive peaks in the time domain as shown in Figure 4.1,
which depicts the WD of a velar from Telugu in the context of [a]. This was
named a "compact spectrum with multiple peaks." The compact spectrum
was usually centered around the frequency of the $F_2$ of the following vowel. When there was more than one peak in frequency at a given time, and the number of contours in the tallest peak were more than twice the number of contours in other peaks, the spectral shape was also classified as compact. This is illustrated in Figure 4.2, which shows a velar from English, in the context of [a].

b) Diffuse Non-Falling Pattern

A diffuse non-falling pattern was characterised by the presence of two or more significant peaks during the burst. These peaks usually covered a range of more than 1.5 kHz and their amplitude decreased, prior to the start of the vowel. A diffuse non-falling spectral shape is illustrated in Figure 4.3, which shows the WD of a Telugu alveolar in the context of [a].

The diffuse non-falling spectral shape was a typical characteristic of bursts of most alveolars in this study. The bursts of velars in the context of high front vowels also displayed a diffuse non-falling spectral shape. There were, however, many differences in the diffuse non-falling spectral shapes of alveolar and velar bursts, that were not reflected in the gross spectral shape.

c) Diffuse Falling Pattern

The diffuse falling pattern was characterized by an energy concentration in the low frequency region, which rapidly increased over 5 to 10 ms, to its maximum value. This was followed by the vowel pattern. The diffuse falling pattern was typical to bilabials in the pilot data. Sometimes, the burst was completely absent in the bilabial burst, and the WD plot started with a typical vowel pattern. This spectral shape was named “diffuse falling” rather than “falling,” partly to be consistent with the terminology used in the literature and partly to include the diffuse vowel pattern when the burst was not noticeable. When the burst was present, it did not excite any higher formants. Thus, the diffuseness of the bilabial burst previously reported in the literature was essentially due to the diffuseness of the following vowel. A typical diffuse falling pattern is illustrated in Figure 4.4, depicting a French bilabial in the context of [a].
Figure 4.1: WD of the beginning of the syllable [ka] (Telugu speaker).
Figure 4.2: WD of the beginning of the syllable [ga] (English speaker).
Figure 4.3: WD of the beginning of the syllable [tə] (Telugu speaker).
Figure 4.4: WD of the beginning of the syllable [pa] (French speaker).
The voice bar was usually continuous with the vowel in the WD plots and there was no decrease in its energy prior to the start of the vowel. The energy in the burst, however, usually decreased prior to the start of the vowel. The energy due to voicing, in the range 0 to 700 Hz, was ignored when determining the diffuse non-falling and compact spectral shapes. Some examples of typical spectral shapes observed in the WD plots are shown as schematic energy density diagrams in Figure 4.5. The vowel region is represented by a mesh. The horizontal lines represent the formant structure and the vertical lines represent the speech spectra at each pitch pulse. The burst is represented by dark areas.

**Formant Onset Duration**

Formant Onset Duration (FOD) was defined as: "the duration from the start of the burst to the start of the vowel." The start of the vowel was taken as the time instant where a formant structure could be observed in the WD plots. FOD may be related to the time duration needed for resonances to set in the vocal tract, after the consonantal release. It is always positive. FOD is different from VOT (see page 48) as defined by Lisker and Abramson [54], which reflects the duration between the start of the burst and the start of voicing. In the pilot data, FOD was observed to become progressively longer as the place of articulation moved from the lips to the velum. Typical values for the FOD are shown in Figures 4.1 - 4.2.

**Burst Duration**

Burst duration (BD) was defined as: "the duration from the start of the burst to the end of the burst." For bursts with diffuse non-falling and compact spectral shapes, there was a decrease in the amplitude of the burst, prior to the start of the vowel. For these spectral shapes, the end of the burst was taken as that point in time at which the number of contours was half the maximum. For the pseudo-log compression described in section 3.7.4, each contour corresponded to approximately 3 dB difference. This strategy of measuring BD was adopted because it was a convenient visual pattern to measure. The BD in compact and diffuse non-falling bursts is shown in Figures 4.3 and 4.2. For bursts with diffuse falling spectral
No noticeable burst.
Only vowel pattern can be seen.

Peaks at low frequency have more energy.

No noticeable burst
Second formant shows up more clearly.

a) Examples of diffuse-falling spectral shape during the burst.

The peak at low frequency is not due to pre-voicing.

The low frequency peak is due to pre-voicing.

b) Examples of diffuse non-falling spectral shape during the burst.

No multiple peaks in time. No pre-voicing.

No multiple peaks in time. Pre-voicing is present.

Compact with multiple peaks.

Compact with multiple peaks.

VOWEL

BURST

c) Examples of compact spectral shape during the burst.

Figure 4.5: Examples of typical spectral shapes observed in the WD plots.
shape, the burst was usually connected with the vowel pattern. There was no noticeable decrease in the burst amplitude. The end of the burst was then taken to coincide with the start of the vowel. See Figure 4.4.

This definition of BD ignores subsequent peaks in a compact spectrum with multiple peaks in time as shown in Figure 4.1. It is suggested that these peaks may also have a role in the perception of velars. Blumstein and Stevens [2] have mentioned that for a stimulus with an abrupt onset to give rise to an auditory representation containing a narrow spectral peak, the spectral peak must persist for a few tens of milliseconds following the onset. In a task which requires identification of velars from bilabials and coronals, however, a spectral shape with multiple peaks was felt to be a sufficient visual feature, since, it was observed that this pattern occurred only in velars in the pilot study. The results presented in Chapter 5 also indicate that a compact spectral shape with multiple peaks was typical to only velars in English, Telugu and French.

Second Formant

Reliable second formant ($F_2$) measurements are not possible in the vicinity of the burst, because of formant transitions. Originally, it was intended to study only the burst region of the stop consonants for invariant acoustic cues, and the use of $F_2$ was not anticipated at the time of computing the WD plots. Consequently, the WD plots included only 2 to 3 pitch periods from the start of the vowel. An approximate value of $F_2$ was nevertheless obtained for each token, as explained on page 118.

4.3.3 Place of Articulation

In terms of the visual patterns described in Section 4.3.2, the invariant visual patterns observed in the WD plots of pilot data are described below.
Bilabials

Bilabials were characterized by diffuse falling spectra, during the burst. The BD was typically less than 10 ms and the burst was usually connected with the vowel pattern. The FOD was also typically less than 10 ms. Figure 4.4 shows an example of a typical bilabial in which the low frequency burst was noticeable. When the steady state vowel started within 2 to 3 ms from release, the burst pattern of bilabials was not noticeable and the WD plot appeared to start directly with the vowel pattern.

Alveolars

Alveolars were mainly characterized by a diffuse non-falling spectral shape. The BD was typically 5 to 15 ms after which there was a noticeable attenuation in the energy of the burst. The FOD was typically between 5 to 20 ms. Figure 4.3 shows an alveolar, spoken by Telugu speaker, in the context of [a]. Sometimes, alveolars in the context of [o] or [u] displayed a compact spectral shape.

Velars

Velars were characterized by diffuse non-falling spectral shape when followed by front vowels, and compact spectral shape when followed by back vowels. BD for velars was around 5 to 15 ms, and FOD was around 30 to 50 ms. Figures 4.1 and 4.2 show (respectively), the WD of syllable [ka] in Telugu and syllable [ga] in English. Figure 4.2 depicts a compact spectrum and Figure 4.1 depicts a compact spectrum with multiple peaks. Compact spectra with multiple peaks usually occurred in the context of [o] and [u] in French and Telugu velars. They were less frequent in the English velars.

It appears that the bursts of velars are most affected by the following vowel. This can be appreciated by considering that the shape of the front cavity has a relatively larger degree of freedom than the shape of the back cavity. Since, for a velar place of articulation the length of the front cavity is the maximum, so then are the coarticulation effects. In the WD plots, the spectral shape of velar bursts
was seen to vary progressively from diffuse non-falling to compact and compact spectrum with multiple peaks (in the time domain), depending on the vowel context and the language.

### 4.3.4 Comparison of Invariant Patterns from WD Plots and Short-Term Spectra

In this section, the invariant patterns obtained using the WD are compared with those obtained using short-time spectral estimation techniques (summarised in Section 2.3). The form of acoustic invariance observed in the speech signal depends to a large extent on the signal processing tool used. In the investigations considered in Section 2.3, LP smoothed spectra were computed using different window sizes. In the present work, the smoothed WD was used to study the time-varying spectral properties of the signal.\(^3\) Table 4.2 compares the equivalent temporal resolutions in each of these methods.

<table>
<thead>
<tr>
<th>Method</th>
<th>Window length</th>
<th>Window shift</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blumstein and Stevens [2]</td>
<td>LP spectra</td>
<td>26 ms</td>
</tr>
<tr>
<td>Kewley-Port [49]</td>
<td>LP spectra</td>
<td>20 ms, 25 ms</td>
</tr>
<tr>
<td>Lahiri et al. [51]</td>
<td>LP spectra</td>
<td>10 ms</td>
</tr>
<tr>
<td>Garudadri et al. [31]</td>
<td>Smoothed WD</td>
<td>4.2 ms</td>
</tr>
</tbody>
</table>

Table 4.2: Comparison of the analysis methods.

The invariant acoustic cues for place of articulation reported in [2,49,51] can be considered to be time-averaged versions of the invariant patterns described in Section 4.3.3. By collapsing these patterns in the time domain, the invariant cues in [2,49,51] are obtained. The enhanced temporal detail in the WD plots serves to explain some of the context dependencies reported in literature (see Section 2.3.1). These context dependencies appear to be mainly due to (i) time domain smear in the analysis methods, and (ii) coarticulation effects. In [2,49], it

---

\(^3\)Use of the term “smoothing” as in: “smoothed WD,” and “LP smoothed spectra,” is unfortunate, since there is nothing in common between the terms, when used in these contexts. In the LP method, “smoothing” corresponds to ignoring the portion of the signal that does not fit the all pole model (see page 12). In smoothing the WD, local averages are computed to suppress the cross-terms (see Section 3.4.2).
was not possible to specify, using LP method, invariant acoustic properties in the phonetic contexts where coarticulation may be expected.

Effects of Temporal Smear

The effects of temporal smear were maximum when short-time spectral analysis methods were used to study bilabials, since bilabials have a weak burst which usually lasts for about 5 to 10 ms. In the WD plots, the energy in this burst was seen to be concentrated in the low-frequency region. Thus, averaging the spectrum of a bilabial burst with the diffuse spectrum of a vowel results in a diffuse spectral line with falling slope. This characteristic of bilabials is adequate to distinguish bilabials from alveolars and velars in many vowel contexts, since, the high amplitude burst associated with alveolar and velar release [74] would make the averaged spectral shape non-falling. For bilabials in the context of high front vowels, however, the high frequency $F_2$ can dominate the low amplitude burst when smeared across each other. This was perhaps the reason for more errors when identifying bilabials in the context of [i, e] in Blumstein and Stevens [2] and Kewley-Port [49]. The relatively short window (10 ms) used by Lahiri et al. limits the smear across the burst and the vowel to some extent. The important innovation in their work, however, was to consider energy gradients over time, at low and high frequencies. By doing so, they were able to specify the place of articulation, independent of the vowel context.

It thus, appears that bilabials can be characterised by a low frequency energy in the burst. Typical values of BD and FOD are in the range of 0 to 10 ms. The diffuse nature of bilabials reported in the literature can be attributed to the temporal smear across the burst and the vowel.

Coarticulation Effects

In Section 4.3.3, it was mentioned that the spectral shapes of the velar bursts were seen to vary most, depending on the vowel context. It was also suggested that the shape of the front cavity varies more than the shape of the back cavity when articulating velars in different vowel contexts. A compact spectral shape is known to occur when the cavities posterior and anterior to the place of articulation
resonate approximately at the same frequency (Stevens, [66]). For velars in the context of back vowels, the resonance frequency of the back cavity coincides with the F2 of the following vowel, giving rise to a compact spectrum. For high front vowels, however, the frequency of F2 is quite high [62]. Thus, for velars in the context of high front vowels, it would be expected that the resonance frequency of the front cavity would be high in anticipation of the following vowel. The spectral shape of velar bursts in the context of high front vowels then, becomes diffuse rather than compact, since the front and back cavities do not resonate at the same frequency. Figure 4.6 shows the WD of the beginning of the syllable [ge], spoken by an English speaker. The spectral shape during the burst can be seen to be diffuse non-falling, instead of compact.

In the WD plots, the spectral shapes of alveolar bursts were observed to be less diffuse in the context of back vowels than in the context of front vowels. The above explanation regarding diffuse non-falling spectral shape of velars in the context of front vowels can be extended to explain this. The lower resonance frequency of the front cavity (in anticipation of a back vowel), enhances the resonance frequency of the back cavity, giving rise to a compact spectral shape, rather than a diffuse non-falling spectral shape. Blumstein and Stevens have reported that a large number of alveolars in the context of [u] were incorrectly accepted as velars by the compact template [2, p. 1011]. Figure 4.7 shows the WD of the beginning of the syllable [du], spoken by an English speaker. The spectral shape can be seen to be compact, rather than diffuse non-falling. This was the only alveolar token of this speaker that was classified as a velar in this study.

It thus appears that the spectral shapes "diffuse non-falling" and "compact" are not unique to alveolars and velars, respectively. Both alveolars and velars share these two features, depending on the vowel context. Given this ambiguity in the spectral shapes, it is possible to explain many error patterns encountered by researchers when specifying the place of articulation information in the acoustic signal. In Blumstein and Stevens [2], a large number of velars in the context of front vowels were incorrectly accepted by the diffuse-rising template as alveolars and incorrectly rejected by the compact template as velars. Alveolars in the context of [u] were incorrectly accepted by the compact template as velars. In Kewley-Port
Figure 4.6: WD of the beginning of the syllable [ge] (English speaker).
Figure 4.7: WD of the beginning of the syllable [du] (English speaker).
[49], a large number of velars in the context of front vowels [i, ɛ, æ] were identified as alveolars. Blumstein and Stevens have explained such context dependencies as coarticulation effects of the tongue body in anticipation of the following vowel [2].

Kewley-Port [49] used the feature “late onset of low-frequency energy,” a measure of VOT, to distinguish velars from bilabials and alveolars. The judges in her study used this feature for only 0.6% of the corpus, indicating that the place of articulation was assigned based on the spectral shape alone in her study. She reported that the feature definition system she used did not capture the importance of this feature. The late onset feature is probably more unsuitable for (pre)voiced velars. Voicing can occur at any time with respect to the consonantal release, since, the place of articulation is associated with the vocal tract and voicing corresponds to the source (glottis), driving the system. Thus, a measure of onset of the vowel, rather than the onset of voicing would serve better to disambiguate velars from non-velars, since, a steady state formant pattern cannot occur until the release is complete.

Formant onset duration defined in Section 4.3.2 is related to time duration required for the steady state resonances to occur in the vocal tract, after the consonantal release. Unlike VOT, the value of FOD is always positive. In the pilot study, FOD was found to increase as the place of articulation moved from the lips to the velum. While the spectral shape during the burst is determined mainly by the vocal tract configuration, FOD appears to be related to inertia of the articulators. Hence, the FOD was considered to be a good choice as a secondary feature for specifying invariant acoustic patterns for place of articulation. Section 4.4.2 presents the manner in which FOD and F2 were used to disambiguate (i) alveolars in the context of back vowels from velars and (ii) velars in the context of front vowels from alveolars.

4.3.5 Voiced and Aspirated Stop Consonants

The invariant patterns described in the previous section were more readily observable in the WD plots of voiceless, unaspirated stop consonants, than in the WD plots of voiced and aspirated stop consonants. When voice onset occurred prior to
release, the voice bar was superimposed on the invariant patterns described above. An example of this is depicted in Figure 4.8, which shows the WD of syllable [di], spoken by a Telugu speaker. The voice bar superimposed on the burst with diffuse non-falling spectral shape can be observed. In the WD plots of aspirated stops, it was relatively more difficult to identify the invariant patterns described in Section 4.3.3. The burst region in an aspirated CV syllable consists of a sudden increase in energy associated with the release, fricative noise at the place of articulation and aspiration. The energy due to aspiration can be fairly wideband. Figure 4.9 shows the WD of the beginning of syllable [ka], spoken by an English speaker. The noisy structure in this plot is perhaps due to aspiration. A compact spectral shape, corresponding to velar place of articulation, can be seen superimposed on a noisy background.

In Table 4.1, it may be seen that only unaspirated voiceless consonants occur in all the three languages. Consequently, the cross-language study envisaged in this study was possible only for the unaspirated voiceless stops of English, Telugu and French. Further, the source for voicing and aspiration is at the glottis, while the source for the place of articulation is at the location of the stop in the vocal tract. Since the invariant cues for the place of articulation were the focus of this study, voiced and aspirated consonants were not considered.

4.4 Development of the Program FEATURE

In the previous section, it was shown that the WD describes the time-varying spectral properties of the burst with much greater precision than short-time spectral analysis methods. In addition to the visual patterns described in section 4.3.3, there were many other interesting aspects of the signal displayed in the partially smoothed WD plots, which were lost when the smoothing was made equivalent to that in a spectrogram, i.e. $T\Omega = 1.0$, as illustrated in Figures 3.14 and 3.15. From these figures, and other WD plots presented in this dissertation, it is clear that the partially smoothed WD is a powerful tool for investigating acoustic invariance in stop consonants and analyzing other transient signals.

In this dissertation, the extent to which the invariant patterns described in
Figure 4.8: WD of the beginning of the syllable [di] (Telugu speaker).
Figure 4.9: WD of syllable [kʰa] from English.
Section 4.3.3 specified the place of articulation was investigated. A program called FEATURE was developed to assign the place of articulation to a set of values for the visual features defined in section 4.3.2. Three judges performed visual pattern matching measured the values of visual features. This procedure was similar to the one used by Kewley-Port [49]. It is reasoned that, if human judges can find invariant patterns that lead to correct identification of the place of articulation, then developing sophisticated pattern matching algorithms for recognition by a computer would be justified.

The program FEATURE included two procedures: (i) a measurement section in which values corresponding to the visual patterns were obtained interactively from the judges, and (ii) a logic section in which the place of articulation was assigned to the WD plot based on an assessment of the visual patterns. These sections are described here.

4.4.1 Measurement Section

In this procedure, judges were asked to indicate: (i) the start of the burst, (ii) the start of the vowel, (iii) the F2 of the vowel and (iv) the spectral shape within the burst region. Simple ad-hoc tests were then performed to determine whether the burst region corresponded to the release of a stop, or some acoustic artefact not relevant to the task of assigning place of articulation in this experiment. Weak bursts of less than 5 ms duration, with spectral shapes that were clearly different from those of bilabials, alveolars and velars discussed in Section 4.3.3, were candidates for these tests. Such bursts were observed in 4 of the 9 bilabials in the pilot data. Because of equipment limitations, it was not feasible to thoroughly investigate the nature of these bursts. It was, however, noted that ignoring these bursts resulted in a typical bilabial pattern. The strategy of ignoring these bursts was adopted whenever their duration was found to be less than 5 ms, and when there was no spectral peak in the burst at the F2 of the following vowel. Until such a time as it will be possible to study these bursts extensively, they are presumed due to some acoustic artefact. When an artefact was detected, judges were requested to ignore this region of the plot and asked to identify another region for the burst. New values for the start of the burst, the start of the vowel, and the spectral shape were
obtained interactively. This process was repeated until reliable burst cues were identified, or until the judges decided to evaluate the next plot. If a reliable burst region was identified, the judges were prompted for the burst duration. The place of articulation was then assigned as explained in the next section. The flowchart of the measuring section of the program is shown in Figure 4.10.

4.4.2 Logic Section

The flowchart for the logic section is depicted in Figure 4.11. Decisions regarding the place of articulation were made at ten different nodes in the flowchart, based on: (i) the spectral shape, (ii) FOD, (iii) BD and (iv) second formant F₂ of the following vowel, as shown in Figure 4.11. This section has four time thresholds — TT2, TT2B, TT3 and TT4, and two frequency thresholds — FT₁ and FT₂ against which the parameter values were compared in assigning the place of articulation.

The place of articulation was assigned as “bilabial” at node 1 for tokens with diffuse falling spectral shape, and velar at node 10 for tokens with compact spectral shape with multiple peaks in time domain. In the pilot data, FOD, BD and F₂ were redundant for the tokens with the foregoing spectral properties.

Alveolars in the context of front vowels and velars in the context of back vowels are said to have a homorganic place of articulation, with the following vowel. For such tokens, the values of the parameters measured from the WD plots do not have conflicting values. These tokens are assigned the place of articulation at nodes 2, 3, 8 and 9. Specifically, tokens with diffuse non-falling spectral shape and FOD—BD less than TT2 or FOD less than TT3 are assigned alveolar place of articulation at nodes 2 and 3, respectively. Similarly, tokens with compact spectral shape and FOD—BD greater than TT2B or FOD greater than TT4 are assigned velar place of articulation at nodes 9 and 8, respectively.

As noted earlier (section 4.3.4), alveolars in the context of back vowels, and velars in the context of front vowels, can have compact and diffuse non-falling spectral shapes, respectively, because of coarticulation in segments with non-homorganic place of articulation. For such tokens with conflicting cues, i.e., diffuse non-falling spectral shape and a long FOD or compact spectral and a relatively short FOD,
Figure 4.10: Flowchart for measuring the visual patterns.
Figure 4.11: Flowchart for assigning the place of articulation.
the ambiguity was resolved using $F_2$ of the following vowel. Thus, for tokens with compact spectral shape and FOD less than TT3, if $F_2$ was less than FT1 indicating a back vowel, then an alveolar place of articulation was assigned at node 4. Similarly, for tokens with diffuse non-falling spectral shape and FOD greater than TT4, if $F_2$ was greater than FT2 indicating a front vowel, then a velar place of articulation was assigned at node 7.

For tokens in which ambiguities of the visual features could not be resolved, the place of articulation was assigned on the basis of the spectral shape alone at nodes 5 and 6. Tokens with diffuse non-falling spectral shapes were assigned an alveolar place of articulation, and tokens with compact spectral shape were assigned a velar place of articulation as in [2,49]

The choice of the time domain thresholds was based on the FOD and FOD—BD values of the 27 voiceless unaspirated stop consonants of the pilot data, measured by the experimenter. These values are shown in Table 4.3. Each entry corresponds to a given phone in the context of [i], [a] and [u], respectively. The frequency thresholds FT1 and FT2 were selected from the formant data given in Peterson and Barney [62]. Most of the informants in [62] spoke General American English.

The choice of thresholds is shown in Table 4.4. There were a small number of tokens in the pilot study to arrive at statistically reliable values for the thresholds. It was not feasible to consider a larger corpus, since computation costs were pro-

<table>
<thead>
<tr>
<th></th>
<th>FOD (ms)</th>
<th>FOD—BD (ms)</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>[b] 8, 6, 15</td>
<td>0, 0, 0</td>
</tr>
<tr>
<td></td>
<td>[d] 17, 12, 22</td>
<td>7, 5, 15</td>
</tr>
<tr>
<td></td>
<td>[g] 32, 28, 25</td>
<td>23, 19, 19</td>
</tr>
<tr>
<td>Telugu</td>
<td>[p] 11, 12, 6</td>
<td>0, 0, 0</td>
</tr>
<tr>
<td></td>
<td>[t] 28, 25, 24</td>
<td>16, 15, 15</td>
</tr>
<tr>
<td></td>
<td>[k] 42, 63, 65</td>
<td>37, 55, 59</td>
</tr>
<tr>
<td>French</td>
<td>[p] 12, 10, 8</td>
<td>0, 0, 0</td>
</tr>
<tr>
<td></td>
<td>[t] 24, 14, 22</td>
<td>18, 7, 10</td>
</tr>
<tr>
<td></td>
<td>[k] 55, 24, 40</td>
<td>46, 18, 34</td>
</tr>
</tbody>
</table>

Table 4.3: FOD and FOD—BD values from the pilot data.
hibitive; therefore, a strategy of adjusting the thresholds for tokens with conflicting cues was adopted. For the tokens that were not classified at one of the nodes 3, 4, 7 or 8 in the first pass, the thresholds TT3, TT4, FT1 and FT2 were relaxed (as shown in Table 4.4), and classification was attempted in a second pass. If classification was remained impossible in the second pass, place of articulation was assigned based on the spectral shape alone, as explained earlier. In principle, it should be possible to further investigate the acoustic signal for additional invariant patterns, when it is not possible to resolve the ambiguities by using the simple features considered here.

The program FEATURE has five check points to trace the various branches traversed in assigning the place of articulation for each token. At the conclusion of analyzing each token, the values of the visual features and the check points were recorded in a file for future analysis. Considerable effort was expended to make this program easy to use and resistant to user errors.

### 4.5 The Experiment

#### 4.5.1 Training of Judges

Three judges participated in this study. Two of them (CM and TU), graduate students in the School of Audiology and Speech Sciences, were conversant with the terminology of acoustic phonetics and one of them (ML), an undergraduate student

<table>
<thead>
<tr>
<th></th>
<th>Pass 1</th>
<th>Pass 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>TT1</td>
<td>5 ms</td>
<td>—</td>
</tr>
<tr>
<td>TT2</td>
<td>16 ms</td>
<td>—</td>
</tr>
<tr>
<td>TT2b</td>
<td>26 ms</td>
<td>—</td>
</tr>
<tr>
<td>TT3</td>
<td>16 ms</td>
<td>20 ms</td>
</tr>
<tr>
<td>TT4</td>
<td>28 ms</td>
<td>24 ms</td>
</tr>
<tr>
<td>FT1</td>
<td>1.0 kHz</td>
<td>1.2 kHz</td>
</tr>
<tr>
<td>FT2</td>
<td>1.7 kHz</td>
<td>1.5 kHz</td>
</tr>
</tbody>
</table>

Table 4.4: The choice of original thresholds.
in Electrical Engineering, was familiar with the signal processing terminology. It was possible to run the program and measure the parameter values from the WD plots as prompted by the program, without knowledge of either disciplines. All the judges were paid for their participation in the study.

Judges were provided with instructions for measuring the visual patterns in the WD plots and for using the program FEATURE (Appendix C) and familiarized with the procedure in an oral presentation and a demonstration, by the experimenter. A set of templates xeroxed onto transparencies were provided to facilitate measurements (see Figure 4.12).

The 27 plots of the pilot study were randomised and given to each of the judges in a loose leaf binder. When the judges performed the experiment independently, the program was able to assign correct place of articulation to 80% of the tokens. When the experimenter performed the same task, a score of 96% was obtained. A collaborative judgement was then performed, in which the experimenter also participated. Since the collaborative session was also part of the training, a unanimous agreement was requested for each parameter value measured from the plots. A score of 92% was obtained in this session. Syllable [ti] by the French informant and syllable [gu] by the English informant were unanimously assigned velar and alveolar place of articulation, respectively. The syllable [ti] had a very weak burst, with only one conspicuous peak in the burst. As a result, the spectral shape of the burst in this token was judged as compact. The token was assigned velar place of articulation at node 6. For the token [gu], the voicing started 10 ms after the release and the vowel started 25 ms after the release. Based on instructions given to the judges, it was unanimously decided that the vowel started 10 ms after the burst, making the value of FOD 10 ms rather than 25 ms. As a result, this token was classified as an alveolar at node 4 instead of a velar at node 8.

Based on feedback from the judges, some changes were made in the program which are summarised in the next section. Some clarifications regarding the original instructions were provided (see Section C.4 in Appendix C).
Figure 4.12: Template used for measuring the visual patterns.
4.5.2 Modification to the Program FEATURE

A cursory analysis of the data files containing responses from the judges revealed that the values of the parameters measured by the judges were not quite similar. The program FEATURE is designed such that small measurement errors in the parameters do not drastically affect its performance. Of the five parameters measured for each plot, the $F_2$ of the vowel does not correspond to the burst region of the CV syllable. As noted earlier, the use of $F_2$ was not anticipated prior to computing and plotting the WD (see page 60). When analyzing data from pilot study, it was found that measuring $F_2$ from the WD plots (as explained on page 118) was quite time consuming and not very accurate. Measurement of $F_2$ is not a difficult task and many techniques for its accurate measurement have been proposed in the literature [57]. Since it is not proposed that the WD be used to measure $F_2$ it was decided to keep the $F_2$ errors uniform across all the judges. In order to achieve this uniformity, the experimenter measured the $F_2$'s of the corpus and an $F_2$ data base was created. This data base was then expanded to include the maximum time duration in each WD plot. The program was modified to prompt judges with the proper template to be used in the measurements, and the $F_2$ of the vowel, for each token.

The errors made by the judges in the training session were mainly recognizing various spectral shapes. A series of schematic WD plots were made to illustrate various forms of diffuse falling, diffuse non-falling and compact spectral shapes in the WD plots, and provided to judges.

4.5.3 Data for the Cross-Language Study

In the final experiment, recorded tokens were obtained from six informants: JG and DG for English, AP and JB for French and HG and CS for Telugu. Tokens for the final experiment was recorded approximately a year after recording tokens for the pilot study. The informants were provided with a list of CV syllables containing the nasal consonant /m/, all the consonants of their language shown in Table 4.1, and the nasal consonant /n/, followed by the vowels /i, e, a, o, u/. The recording conditions were similar to those used for preparing the pilot set.
Each informant recorded three sets of repetitions. The last two sets were chosen for digitization. WD plots were generated for the unaspirated voiceless stops, as explained in Appendix A.

Dental Tokens from Telugu

Based on theoretical considerations and spectral analysis of Swedish dental stops, Fant showed that dentals have a diffuse spectral shape with predominance of high-frequency energy during the burst [23]. Blumstein and Stevens suggested that alveolars, dentals and other phonetic categories broadly termed "coronals," share similar spectral properties during the burst, namely, diffuse-rising spectral shape and that they would differ in other attributes of the acoustic signal, such as the direction of formant transitions or onset frequencies of particular formants [2]. Phonologically, alveolars and dentals are also described with the same feature, i.e., [coronal] [12,33]. These observations prompted Lahiri et al. [51] to investigate alveolar and dental stops in various languages to determine whether they displayed the diffuse-rising spectral shape. Unfortunately, their prediction that the spectral shape was diffuse-rising for both alveolars and dentals was not substantiated. They did show that the relative changes in the distribution of energy from the burst release to the onset of voicing was the same for these two phonetic categories. Since dentals and alveolars are phonologically contrasted in Telugu, dentals of Telugu were included in the corpus. The program FEATURE was modified such that if it assigned an alveolar place of articulation to a dental stop, the classification was considered correct.

The final data set consisted of tokens spoken by three new speakers (DG, CS, JB), two new vowels vowel contexts ([e, o]), and a new place of articulation (dentals of Telugu), which were not considered in the pilot study. This resulted in 60 plots for English, 80 plots for Telugu (including the plots for 20 dental stops) and 60 plots for French. The WD plots were randomized within each language and kept in loose-leaf binders for analysis by the judges. These plots were not analyzed by the experimenter prior to finalizing the program FEATURE.
4.5.4 Running the Program FEATURE

The final experiment was conducted in three stages. In the first stage, judges spent an average of 7 hours each and analyzed the data independently. In tokens that were incorrectly classified in the independent judgement, a distinct pattern was observed in the five visual features measured by the judges. A closer examination of the parameter values showed that the single set of thresholds used in the program FEATURE was not optimum for all the three languages. For instance, the FOD's of alveolars in Telugu were much longer than those in English. In order to establish the extent to which parameter values measured by the judges were invariant for a given place of articulation, a different set of thresholds for each language was arrived at by re-examining the pilot data — this time, independently for each language. The second stage of the experiment involved running the program with different threshold values for each language, but for the same parameter values obtained in the independent judgement. The choice of thresholds for English were the same as in the independent judgement. The new thresholds used for the tokens of Telugu and French are shown in Table 4.5.

In the third stage, the three judges participated in a collaborative judgement and analyzed 50 WD plots for which the assignment of spectral shape had not been unanimous. Unlike the collaborative judgement reported in [49], the corpus did not include the tokens for which the spectral shape was unanimous and wrong.
place of articulation was assigned to the consonant. Similarly, tokens for which the same place of articulation was assigned by the three judges, in spite of different spectral shapes, were also analyzed in the collaborative judgement. No additional instructions were given to judges for their collaborative judgement. This judgement mainly served the purpose of correcting errors due to oversight and/or careless judgement.

The results of the three stages of this experiment are presented in the next chapter.

4.6 Summary

In this chapter, a cross-language study of stop consonants from English, Telugu and French, to investigate acoustic invariance was presented. Some of the linguistic terms used throughout this dissertation and the organization of stop consonants in the languages considered were explained. The invariant patterns for each place of articulation, observed in WD plots of the pilot data were explained. A comparison of these patterns with the invariant cues reported in the literature indicated that the difficulties reported in [2,49,51], in specifying the place of articulation (independent of the vowel context) were (i) partly due to temporal smear in the analysis techniques, and (ii) partly due to coarticulation effects. The WD was useful in preserving the temporal detail and identifying some coarticulation effects. A formal study was conducted to establish if it was possible to specify the place of articulation, independent of the vowel context, based on the invariant patterns observed in the WD plots. The results of this study are described in Chapter 5.
Chapter 5

Results and Discussions

In this chapter, the results of a cross-language study to investigate acoustic invariance in stop consonants are presented. The purpose of the experiment was to identify visually unique patterns in the WD plots of CV syllables spoken by native speakers of different languages. The corpus consisted of a total of 200 utterances, 60 from English, 80 from Telugu and 60 from French, spoken by two native speakers of each language. Five vowel contexts ([i, e, a, o, u]) were considered for each place of articulation. The English tokens had bilabial, alveolar and velar places of articulation. The Telugu tokens had bilabial, dental, retroflexed post-alveolar and velar places of articulation. The French tokens had bilabial, dental and velar places of articulation. As noted earlier, for brevity, retroflexed post-alveolar stop consonants from Telugu will be referred to as Telugu alveolars. Three judges analyzed the WD plots and provided the values for 5 visual parameters defined in the Section 4.3.2. A place of articulation was assigned to each plot by the program FEATURE, based on the responses of the three judges. The performance of this program in assigning the correct place of articulation is discussed.

5.1 Overview

Section 5.2 presents the overall performance in identifying the invariant patterns. The data is collapsed to allow discussion of the results in each language and for each place of articulation, independently. The complete data is then presented to allow discussion by (i) language, and (ii) place of articulation.
Table 5.1: Overall scores by language.

Section 5.3 presents the performance at each of the ten nodes in FEATURE where a place of articulation was assigned. The number of tokens that were assigned the correct place of articulation, and the number of tokens that were assigned a wrong place of articulation, at each node are presented.

Section 5.4 presents the role played by various visual parameters measured from the WD plots, in arriving at a correct place of articulation.

5.2 Overall Performance

The overall performance of the program FEATURE, in assigning the correct place of articulation to a WD plot has been arranged as a confusion matrix in Table 5.4. In the following two sections, the performance is explained by language and by place of articulation, by collapsing the confusion matrix accordingly. Since the performance of the judges was fairly consistent in measuring parameter values from the WD plots, results for individual judges are not presented.

5.2.1 Performance by Language

Table 5.1 shows the overall performance in English, Telugu and French for the independent session, the experiment with new thresholds and the collaborative session.

The 10% and 0.6% increase in scores for Telugu and French tokens from the independent session to the session with new thresholds show that the choice of original thresholds was particularly poor for Telugu tokens. Although only 75% of Telugu tokens displayed invariance in the independent session, 85% of them,
nevertheless shared invariant acoustic properties in the parameters measured by the judges, as demonstrated by the experiment with new thresholds.

For the collaborative judgement, there was an improvement of 5.5%, 2.2% and 6.6% in the results for English, Telugu and French respectively, resulting in an overall improvement of 4.8%. The small improvement in the results of the collaborative judgement implies that the distribution of errors was fairly random in the independent session, indicating that the training of the judges was good enough to enable them to make reliable measurements from the WD plots. The 2.2% increase in the results for Telugu indicate that there were relatively fewer errors in the independent judgement of Telugu tokens.

5.2.2 Performance by Place of Articulation

Table 5.2 shows the confusion matrices for place of articulation in: (i) the independent session, (ii) the experiment with new thresholds and (iii) the collaborative session. The scores for the alveolar and dental tokens have been combined as coronals in this table. Correct scores are indicated by bold face.

From the flowchart in Figure 4.11, it may be deduced that the choice of new thresholds did not affect the classification of bilabials. There was relatively small improvement in the classification of bilabials in the collaborative session (only 3.3%, corresponding to 2 tokens), indicating that errors in bilabials were not due to careless judgement.

Most of the improvements in the session with new thresholds and in the collaborative session were with the identification of coronal tokens. From Tables 5.1

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Lab.</td>
<td>75.0</td>
<td>11.7</td>
<td>13.3</td>
</tr>
<tr>
<td>Cor.</td>
<td>10.3</td>
<td>60.9</td>
<td>28.8</td>
</tr>
<tr>
<td>Vel.</td>
<td>1.1</td>
<td>2.2</td>
<td>96.7</td>
</tr>
</tbody>
</table>

Table 5.2: Overall performance by place of articulation.
Table 5.3: Performance by vowel context.

<table>
<thead>
<tr>
<th></th>
<th>Bilabial</th>
<th>Coronal</th>
<th>Velar</th>
</tr>
</thead>
<tbody>
<tr>
<td>i</td>
<td>75</td>
<td>75</td>
<td>92</td>
</tr>
<tr>
<td>e</td>
<td>83</td>
<td>81</td>
<td>100</td>
</tr>
<tr>
<td>a</td>
<td>75</td>
<td>75</td>
<td>100</td>
</tr>
<tr>
<td>o</td>
<td>92</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>u</td>
<td>67</td>
<td>75</td>
<td>100</td>
</tr>
</tbody>
</table>

and 5.2, it may be seen that for the session with new thresholds, improvements were in the identification of coronal tokens of Telugu. Most of the errors in the independent session were associated with identification of the alveolars of English and dentals of French. About 8.8% of these were corrected in the collaborative session.

The identification of velars was uniformly good, across the three languages (greater than 96%). Their classification was robust with respect to the threshold values and there were relatively fewer errors in the independent session.

5.2.3 Performance by Vowel Context

Table 5.3 shows the results of the collaborative session for each place of articulation, with respect to vowel context. For the limited corpus considered in this study, no context dependencies could be found in the performance of the program FEATURE. Errors were distributed, relatively uniformly, across front and back vowels.

5.2.4 Overall Scores

Table 5.4 shows the performance by both language and place of articulation, for each speaker. The salient points to be noted from this table are discussed here briefly.
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Eng.</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>b</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>JG</td>
<td>66.7</td>
<td>16.7</td>
<td>16.7</td>
</tr>
<tr>
<td>DG</td>
<td>96.7</td>
<td>3.3</td>
<td>0.0</td>
</tr>
<tr>
<td>d</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>JG</td>
<td>0.0</td>
<td>76.7</td>
<td>23.3</td>
</tr>
<tr>
<td>DG</td>
<td>6.7</td>
<td>60.0</td>
<td>33.3</td>
</tr>
<tr>
<td>g</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>JG</td>
<td>0.0</td>
<td>10.0</td>
<td>90.0</td>
</tr>
<tr>
<td>DG</td>
<td>0.0</td>
<td>3.3</td>
<td>96.7</td>
</tr>
<tr>
<td><strong>Tel.</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>p</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HG</td>
<td>56.7</td>
<td>16.7</td>
<td>26.7</td>
</tr>
<tr>
<td>CS</td>
<td>76.7</td>
<td>16.7</td>
<td>6.7</td>
</tr>
<tr>
<td>t</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HG</td>
<td>0.0</td>
<td>46.7</td>
<td>53.3</td>
</tr>
<tr>
<td>CS</td>
<td>3.3</td>
<td>50.0</td>
<td>46.7</td>
</tr>
<tr>
<td>ɾ</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HG</td>
<td>0.0</td>
<td>80.0</td>
<td>20.0</td>
</tr>
<tr>
<td>CS</td>
<td>13.3</td>
<td>60.0</td>
<td>26.6</td>
</tr>
<tr>
<td>k</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HG</td>
<td>0.0</td>
<td>0.0</td>
<td>100</td>
</tr>
<tr>
<td>CS</td>
<td>0.0</td>
<td>0.0</td>
<td>100</td>
</tr>
<tr>
<td><strong>Fre.</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>p</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AP</td>
<td>90.0</td>
<td>0.0</td>
<td>10.0</td>
</tr>
<tr>
<td>JB</td>
<td>63.3</td>
<td>16.7</td>
<td>20.0</td>
</tr>
<tr>
<td>ɾ</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AP</td>
<td>36.7</td>
<td>53.3</td>
<td>10.0</td>
</tr>
<tr>
<td>JB</td>
<td>10.0</td>
<td>56.7</td>
<td>33.3</td>
</tr>
<tr>
<td>k</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AP</td>
<td>6.7</td>
<td>0.0</td>
<td>93.3</td>
</tr>
<tr>
<td>JB</td>
<td>0.0</td>
<td>0.0</td>
<td>100</td>
</tr>
</tbody>
</table>

Table 5.4: Performance by place of articulation and language.
Bilabials

The results from the collaborative session indicate that identification of bilabials was relatively poor for Telugu informants HG, CS and French informant JB. There was a weak burst, a few tens of milliseconds prior to the diffuse falling pattern in the WD plots of 26 bilabial tokens out of a total of 60 tokens. The strategy of ignoring these regions (as explained in Section 4.4.1) was effective for 13 of these tokens. For the other 13 tokens, the program was unable to ignore these bursts, even though the spectral shape of these bursts was quite different from those of alveolars and velars. It was not possible to study these bursts in detail, at this stage, due to limitations in the A/D, D/A and waveform editing facilities. Further investigations are needed to establish the perceptual relevance of these bursts and to incorporate this phenomenon into the set of invariant patterns. From Table 5.4, it can be seen that this characteristic of bilabials could be specific to language and/or speaker, since these bursts were predominant in the bilabials of both Telugu informants and French informant JB. Figure 5.1 shows the WD of the syllable [be] spoken by a native English speaker. This figure presents an example of a weak burst, prior to the diffuse falling pattern.

Dentals and Alveolars

The identification of alveolars was rather poor for tokens spoken by English speaker DG. All errors for DG (5 out of 30 tokens) occurred only in alveolars and 4 of these tokens occurred were in one set of CV syllables. Two of these tokens had weak bursts, leading to incorrect classifications of spectral shape. For the other tokens, the start of the vowel was identified as being a couple of pitch periods later by all judges, even though the WD plots displayed clear formant structure. Fig 5.2 shows the beginning of the syllable [da], spoken by DG. For this token, the judges identified the start of the vowel at 40 ms, instead of 12 ms. Ambiguities of this nature were not present in the pilot data.

Lahiri et al. [51] studied the dentals and alveolars of French and Malayalam and showed that the spectral shape of dentals during the burst was not diffuse-rising, across all vowel contexts. In Chapter 2, it was suggested that the LP technique
Figure 5.1: WD of the beginning of the syllable [be] (English speaker).
Figure 5.2: WD of the beginning of the syllable [da] (English speaker DG)
was not suitable for studying the nature of stop consonants. In the present study all dental from tokens of Telugu were found to have a diffuse non-falling spectral shape during the burst. This is extremely encouraging, since, the program was designed for alveolars and neither the judges nor the experimenter had analyzed the WD of a dental token. This suggests that the original predictions of Fant [23] and Blumstein and Stevens [2] were correct, and that problems encountered by Lahiri et al. might be attributed to the signal processing tool. Figure 5.3 shows the WD of the beginning of a Telugu dental in the context of [e]. The spectral shape during the burst can be seen to be diffuse non-falling.

The FOD was much longer for the alveolars and dentals of Telugu, than for the alveolars of English and dentals of French, which lead to a number of Telugu coronals being classified as velars. It was apparent that a single set of thresholds was not appropriate for all languages. A new set of thresholds, different for each language, was obtained from Table 4.3. With the new thresholds, there was an improvement of 30% in the results of coronals of Telugu. Thus, although the original thresholds failed to demonstrate invariance in the coronals of Telugu, there were invariant patterns in the parameters obtained in the independent session, in approximately 89% of Telugu coronals. About 8% of errors occurred due to careless judgement; these were corrected in the collaborative session. From table 5.4, it can be seen that the experiment with new thresholds mainly served to improve the scores of Telugu coronals. The improvement for the rest of the tokens was fairly marginal.

Performance for French dentals was poor, with an average score of 55%. In the pilot study, dental tokens from French were found to have weak bursts, leading to incorrect classification of the spectral shape. Similar problems were also encountered with the French dentals in the final experiment. A more extensive pilot study might have helped in obtaining a better performance for the French dental tokens. It was encouraging to note that not all errors were due to improper choice of thresholds or to tokens of poor quality. About 20% of the errors were due to over-sight and careless judgement, which were corrected in the collaborative session, ultimately resulting in an overall performance of 75% for French alveolars.
Figure 5.3: WD of the beginning of the syllable \[\text{[te]}\] (Telugu speaker)
Velars

Performance of velars was excellent in all sessions (greater than 90%). After the collaborative session, there was only one velar ([gi] of English speaker JG) that was classified as an alveolar. The following conclusions can be made concerning invariant acoustic patterns in velars:

- Velars can have both diffuse non-falling and compact spectral shapes, during the burst, depending on the vowel context.

- The FOD following the burst and the F2 of the vowel can be used along with spectral shape, to specify the invariant acoustic patterns for velars, independent of the following vowel.

In Blumstein and Stevens [2] and Kewley-Port [49], considerable context dependency was reported in specifying place of articulation for velars. Using the WD, it was possible to resolve these context dependencies using the FOD and F2 information.

5.3 Performance at Each Node

Table 5.5 shows the number of tokens assigned a place of articulation at each of the ten nodes in flowchart in Figure 4.11. Each entry corresponds to the number of times a correct place of articulation was assigned at a given node. The quantity in parentheses indicates the number of times an incorrect place of articulation was assigned at a given node.

A large number non-bilabial French tokens were classified as bilabials at node 1. In assigning the bilabial place of articulation, the program FEATURE used only the spectral shape and ignored the FOD–BD and FOD information. It is possible to use the parameters FOD-BD and FOD together with the diffuse falling spectral shape in the flowchart in Figure 4.11, to prevent non-bilabials being classified as bilabials. The pilot data, however, indicated that such measures were unnecessary.

Alveolars in the context of back vowels and velars in the context of front vowels tend to have compact and diffuse non-falling spectral shapes respectively. Nodes 4
<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Ind. Sess.</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Eng.</td>
<td>49</td>
<td>17</td>
<td>21</td>
<td>0</td>
<td>4</td>
<td>12</td>
<td>9</td>
<td>12</td>
<td>0</td>
<td>24</td>
</tr>
<tr>
<td>Tel.</td>
<td>40</td>
<td>15</td>
<td>13</td>
<td>1</td>
<td>1</td>
<td>42</td>
<td>0</td>
<td>12</td>
<td>2</td>
<td>35</td>
</tr>
<tr>
<td>Fre.</td>
<td>46</td>
<td>2</td>
<td>4</td>
<td>1</td>
<td>0</td>
<td>27</td>
<td>0</td>
<td>22</td>
<td>7</td>
<td>25</td>
</tr>
<tr>
<td><strong>New Thrs.</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Eng.</td>
<td>49</td>
<td>17</td>
<td>21</td>
<td>0</td>
<td>4</td>
<td>12</td>
<td>9</td>
<td>12</td>
<td>0</td>
<td>24</td>
</tr>
<tr>
<td>Tel.</td>
<td>40</td>
<td>84</td>
<td>9</td>
<td>1</td>
<td>2</td>
<td>13</td>
<td>2</td>
<td>12</td>
<td>5</td>
<td>35</td>
</tr>
<tr>
<td>Fre.</td>
<td>46</td>
<td>2</td>
<td>13</td>
<td>1</td>
<td>1</td>
<td>19</td>
<td>2</td>
<td>21</td>
<td>8</td>
<td>25</td>
</tr>
<tr>
<td><strong>Col. Sess.</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Eng.</td>
<td>57</td>
<td>18</td>
<td>20</td>
<td>0</td>
<td>4</td>
<td>11</td>
<td>9</td>
<td>14</td>
<td>0</td>
<td>24</td>
</tr>
<tr>
<td>Tel.</td>
<td>39</td>
<td>91</td>
<td>8</td>
<td>1</td>
<td>17</td>
<td>2</td>
<td>12</td>
<td>5</td>
<td>6</td>
<td>35</td>
</tr>
<tr>
<td>Fre.</td>
<td>45</td>
<td>2</td>
<td>19</td>
<td>1</td>
<td>20</td>
<td>3</td>
<td>21</td>
<td>12</td>
<td>1</td>
<td>23</td>
</tr>
</tbody>
</table>

Table 5.5: Performance at each node in the decision tree.
and 7 were provided to allow the program FEATURE to assign the correct place of articulation to these tokens by using FOD and $F_2$ information. A very small number of alveolar tokens arrived at node 4, in all three languages. This is partly due to the fact that coarticulation effects are greater for velars than for alveolars and partly due to poor choice for the value of threshold $FT1$.

For tokens that arrived at nodes 5 and 6, conflicting visual cues could not be resolved and a decision about place of articulation was based on the spectral shape alone. For these tokens, the set of visual patterns defined in Section 4.3.2 was not sufficient to assign a reliable place of articulation.

For the English tokens, the nodes 2, 3 and 8, 10 appear to be quite robust for alveolars and velars respectively. For Telugu and French velars, the nodes 7, 8, 9 and 10 all seem to be quite reliable, with a large number of judgement favouring nodes 7 and 10. With the new thresholds, node 2 was found to be the most favoured for Telugu alveolars.

### 5.4 Relevance of the Acoustic Features Measured in WD Plots

In the past, many acoustic features have been proposed as invariant acoustic cues for place of articulation in stop consonants. To some extent, all these features have been shown to be perceptually relevant, i.e. by synthesizing stimuli containing these features and investigating the extent to which these are perceived as a given stop consonant [20,22,68,49,51]. This suggests that there might be a set of properties in the acoustic speech signal, which ultimately conveys perceptual information of a linguistic nature. The goal is to discover all the elements of such a set and to specify the extent to which each member conveys information about place of articulation. Experiments to establish the perceptual relevance of spectral shape and the combination of FOD, BD and $F_2$ were not attempted as part of this dissertation. The role played by these features in determining place of articulation was nevertheless assessed by studying the judges' responses, as described below.
Table 5.6 shows distribution of the feature “spectral shape,” by place of articulation, in each language. For each place of articulation, in each language, a total of 20 tokens were analyzed by the three judges.

### Diffuse Falling

From Table 5.4, it can be seen that 95% of bilabials in English, 65% of bilabials in Telugu and 75% of the bilabials in French were judged to have diffuse falling spectral shape, after errors due to careless judgement were corrected. Thus, out of a total of 20 bilabial tokens from each language 1 token in English, 7 tokens in Telugu and 4 tokens in French, were found not to have a diffuse falling spectral shape. As discussed in section 5.2.4, these tokens displayed a brief burst prior to the diffuse falling spectrum and were assigned an alveolar or velar place of articulation only by default. Thus, the absence of a diffuse falling spectral shape was not a sufficient cause to reject a token as bilabial. The presence of a diffuse falling spectral shape during the burst was nevertheless found to be a reliable cue for bilabials, since only 2.9% of non-bilabials were judged to have a diffuse falling spectral shape.
Figure 5.4: Distribution of diffuse non-falling and compact spectral shapes in alveolars and velars.

**Diffuse Non-Falling Vs Compact Spectral Shapes**

The distribution of diffuse non-falling and compact spectral shapes in coronals and velars is shown in Figure 5.4. It can be seen that the number of velars judged to have a diffuse non-falling spectral shape was substantially more than the number of coronals that were judged to have a compact spectral shape. About 20% of velars in English and Telugu and 35% of velars in French were judged to have a diffuse non-falling spectral shape. Similarly, 6.7%, 1.7% and 5% of the coronals in English, Telugu and French were judged to have a compact spectral shape. All dental tokens of Telugu were judged to have a diffuse non-falling spectral shape. From Figure 5.4, it may also be deduced that co-articulation effects were more prominent for velars than for coronals. Such context dependencies, with respect to the following vowel were not observed during the bursts in bilabials.

In Section 4.3.3, it was suggested that the spectral shape of velar burst was most affected by the vowel context. For velar place of articulation, it has been shown that the cavities posterior and anterior to the constriction have roughly the same resonance frequencies [66]. Since the shape of the front cavity has a greater degree of freedom than the shape of the back cavity, its resonance frequency varies more than the resonance frequency of the back cavity. Velars in the context of back
vowels displayed a compact spectrum during the burst in the WD plots. For velars followed by front vowels, the front cavity has a higher resonance frequency than the back cavity. This difference in resonance frequencies appears to cause a diffuse spectral shape in the WD plots of velars in the context of front vowels. The FOD, might be related to the inertia of the articulators, rather than to the shape of the vocal tract. Hence, spectral shape along with FOD, should specify the invariant cues for the place of articulation in velars. Results of the present study indicate that this indeed the case. Figure 4.6 shows the WD of the beginning of syllable [ge], spoken by a native English speaker. The spectral shape is diffuse non-falling, rather than compact, because of the vowel context. The FOD for this token was 32 ms.

5.4.2 Formant Onset Duration and Burst Duration Statistics

Figures 5.5, 5.6 and 5.7 show respectively, scatter plots of FOD—BD vs FOD, for English, Telugu and French. The value of BD is around 10 ms in most situations. These figures show that there is considerable overlap in the value of FOD for coronals and velars. It may also be deduced that FOD increases as the place of articulation moves from lips to the velum.

FOD is believed to be a more robust measure to distinguish different phonetic categories than the “late onset of low-frequency energy” feature proposed by Kewley-Port. The late onset of low-frequency energy is a direct measure of the VOT [49, pp. 325] and would not work for pre-voiced velars. For a given place of articulation, voicing can occur prior to the release, along with the release or after the release, depending on the phonemic category [54,12]. For this reason, the FOD was defined such that it is related to the time taken for release, rather than the onset of voicing. Unlike VOT, the value of FOD is always positive. Its value is related to the time taken for the opening at the place of articulation to reach sufficient area, to allow steady state resonances in the vocal tract.
Figure 5.5: FOD–BD vs FOD scatter plot for English tokens

Figure 5.6: FOD–BD vs FOD scatter plot for Telugu tokens
Figure 5.7: FOD–BD vs FOD scatter plot for French tokens

5.5 Summary

In this chapter, the results of a cross-language study, to investigate acoustic invariance in stop consonants were presented. It was possible to specify the place of articulation for about 87% of the English tokens, 88% of the Telugu tokens and 83% of the French tokens, based on invariant patterns in the WD plots. The performance was best for velars in all the three languages (average of 98.3%) and alveolars and dentals of Telugu (average of 97.5%). There were no context dependencies in specifying the place of articulation, i.e. the errors in specifying the place of articulation were uniformly distributed over all vowel contexts. Further investigations, with larger corpora, are suggested for coronals of English and French and bilabials of Telugu and French. The results presented in this chapter clearly indicate that the WD is more suited to study the nature of stop consonants, than the conventional short-time spectral estimation methods.
Chapter 6

Summary and Suggestions for Future Research

6.1 Summary

In this investigation, the partially smoothed WD was used in a cross-language study of stop consonants. The purpose of the investigation was to identify invariant acoustic patterns in the WD plots which might specify the place of articulation in stop consonants. Contributions of this study include: (i) developing the partially smoothed WD as a tool for transient signal analysis, and (ii) demonstrating evidence in support of acoustic invariance in stop consonants. A summary of both topics is presented in this chapter, and some suggestions for extending this work are presented.

6.1.1 Wigner Distribution

The WD is a new signal processing tool that is more suited for analyzing transient signals than conventional short-time spectral analysis methods. It has been successfully used in many areas, such as loud-speaker analysis, biological signal processing, seismic signal analysis, etc. [44,59,61,4,42,43]. For speech signal analysis, however, reports on applications of WD in the literature have been limited [10,28,29,31,1,65,72]. In Chester's work [10], real signals were used to compute the WD and the problem of cross-terms in the WD was not addressed. Atlas et al. [1] also used real signals to compute the WD. In the works of Atlas et al. and Riley [1,65], the WD analysis was equivalent to that of a spectrogram, since the
WD was smoothed for positivity.

For bilinear distributions of Cohen’s class, the requirements of positivity and correct marginals are incompatible. The spectrogram and related short-time spectral estimation techniques are positive, but do not have the property of correct marginals. Consequently, they fail to localize time or frequency events in the signal, to their respective time-frequency coordinates in the distribution. The WD, on the other hand, has the property of correct marginals. It has nevertheless, other undesirable properties such as cross-terms and negative regions. The requirements of positivity and correct marginals are often emphasized in the literature, when proposing time-frequency distributions for signal analysis. By sacrificing both these properties, it is possible to obtain a distribution that has — a) better time-frequency localization properties than short-time spectral estimation techniques, and b) cross-terms of small enough amplitude as to have a negligible influence on the extraction of useful information. The partially smoothed WD presented in this dissertation is such a compromise. Such approaches are not generally accepted in the literature (see Cohen’s comments in [18]). The above compromise has, nevertheless, proved to be very useful in understanding the nature of stop consonants in speech signals.

The salient points regarding the development of the partially smoothed WD, as a tool for analyzing transient signals are summarised below:

• It was shown that the LP smoothed spectra were not suitable for studying the nature of the burst in stop consonants. This method not only introduces a bias in the time direction between successive spectra, but also ignores the portion of the signal that does not correspond to a typical steady state vowel (chapter 2, page 12).

• The WD smoothed for positivity, and the spectrogram were shown to be equivalent, in terms of resolution in the time domain and the frequency domain (page 24). An interpretation of the partially smoothed WD for multi-component signals was presented (page 25).

• The implementation of the partially smoothed WD was discussed in detail. A FORTRAN callable subroutine for computing the partially smoothed WD
was also included (page 27 and Appendix B).

- In order to overcome the drawbacks of the cross-terms in a partially smoothed WD, it is essential to have an understanding of the nature of the cross-terms. The nature of the cross-terms in the WD of multicomponent signals was illustrated in Section 3.6, by expressing the signal in terms of simpler signals, for which the cross-terms could be readily written.

- The choice of time-domain and frequency-domain smoothing parameters was discussed in Section 3.7. Spectral wrap was demonstrated in the WD of analytic signals. The choice of the frequency-domain smoothing parameter $\Omega$ should take the spectral wrap into consideration. It was pointed out that a priori knowledge of the signal is useful in choosing these parameters and that the choice essentially depends on the temporal and/or spectral separation between regions of interest in the signal. This was demonstrated using two chirps of opposite slopes in Figure 3.12. As the separation between the self-terms decreases in the frequency domain, the effect of cross-terms is seen to become more prominent.

- In a recent paper, Cohen pointed out that the WD is not necessarily zero, when the signal itself is zero [18]. It was shown that these terms do not inhibit extraction of useful information from the WD. The time-domain window used to compute the WD on a digital computer, suppresses these contributions in the WD where the signal is zero, to an acceptable level (see Section 3.7.3).

### 6.1.2 Acoustic Invariance

According to the notion of "acoustic invariance," there are invariant acoustic patterns present in the speech signal, which are related to the phonetic description of the signal in an abstract way. Many researchers have studied the time-varying spectral properties of stop consonants in order to identify these patterns [20,22,68,2,49,51]. These investigations were essentially based on short-time spectral analyses methods — spectrographic analysis, or time-varying LP smoothed spectra. Although these investigations unequivocally established the feasibility of extracting invariant cues in certain contexts from speech using signal processing
methods, they failed to provide a set of necessary and sufficient invariant cues in any given phonetic context. This context dependency is often quoted to argue that it is not possible to extract invariant patterns in the acoustic domain, that would correlate with the phonetic description of a speech segment [53].

From the discussions in Chapter 2 (page 12), it is clear that short-time spectral estimation techniques are not suitable for investigating invariant acoustic patterns in stop consonants. These methods not only introduce a smear across the burst and the vowel regions (giving rise to context dependencies), but also fail to provide sufficient information about the signal which might explain the coarticulation effects. The partially smoothed WD was used in a cross-language study of 200 stop consonants, spoken by native speakers of English, Telugu and French.

The salient results of this cross-language study are as follows:

- The context dependency reported in the literature for bilabials (in the context of [i]), was partly due to the temporal smear between the burst and the following vowel, introduced by the analysis method. The context dependencies reported for alveolars and velars (in the context of [u] and [i] respectively), are attributed to inability of those methods to provide sufficient detail, with which to characterize the coarticulation effects.

- The temporal information in the WD was used to reduce the context dependency in specifying the invariant cues for place of articulation. The term “Formant Onset Duration” was defined to quantify the time taken for the steady state formants to set in after the release.

  - It was shown that the features “diffuse” and “compact” spectral shapes occur in both alveolars and velars, depending on the vowel context. The FOD and F2 of the following vowel were used to resolve ambiguities in the spectral shape during the burst.

  - About 78% of the bilabials, 81% of the alveolars and 98% of the velars were assigned correct place of articulation. Approximately 87% of the English tokens, 88% of the Telugu tokens, and 83% of the French tokens were identified correctly.
Unlike in other investigations, the errors in specifying the place of articulation were uniformly distributed over all vowel contexts.

- Both dentals and alveolars of Telugu were shown to have diffuse non-falling spectral shape during the burst, which supports the theoretical predictions of Blumstein and Stevens [2]. That Lahiri et al. [51] were unable to demonstrate the same finding for stop consonants from French and Malayalam is attributed to the inadequacy of the LP analysis.

Lahiri et al. [51] obtained correct identification for 89.3% of the bilabials, 94% of the dentals and 92.5% of the alveolars using time-varying LP smoothed spectra. It is however, not possible to compare their results with the results presented in this dissertation, for two reasons; (i) they did not consider velars, and (ii) the constraints on their experiment were quite different from those used in the present study. The corpus for the final experiment in [51] included 100 English tokens, 100 French tokens, and 95 of the 175 Malayalam tokens, that were analyzed in earlier pilot studies; and 198 additional tokens from French that were not analyzed in earlier pilot studies. Hence, the 493 tokens (on which their results were based) included 295 tokens that were used to determine the invariant patterns.

### 6.2 Suggestions for Future Research

The discussions in Chapter 3 suggest that the WD is a useful technique for analyzing transient signals, in spite of cross-terms and negative regions. By sacrificing the properties of correct marginals and positivity to some extent, the partially smoothed WD can be used to extract more information from the signal than is possible with short-time spectral analysis techniques. The partially smoothed WD can be used to analyze many other nonstationary signals in addition to speech signals. All the applications involving short-time spectra of transient signals can be considered as suitable candidates for the WD analysis techniques, e.g., EEG and EMG signal analysis for early detection of abnormal pathological conditions, harmonic analysis of transients in high voltage DC transmission systems etc. [32].

Boudreaux-Bartels [7,8] and Yu and Cheng [73] presented techniques for syn-
thesizing the signal from a modified WD. Boudreaux-Bartels pointed out the importance of including the cross-terms in the time-frequency description, in order to obtain a good approximation to the synthesised signal. This information is lost when the WD is partially smoothed. A signal synthesis technique from partially smoothed WD would be very useful in many applications. For example, speech signals with desired properties in the time and frequency domains could be synthesised and their perception studied. This would be similar to the Haskins Lab's pattern playback machine, which proved to be invaluable in many investigations.

Simple peak picking routines can be written to detect and measure the visual parameters described in Section 4.3.2. This would enable the experiments (described in Chapter 4) to be conducted on a large data base. The study could be extended to nasals, fricatives and other phonetic categories, and would allow the investigation of the properties of aspiration.

In speech research, narrowband and wideband spectrograms are used to study different aspects of the signal. Narrowband spectrograms suffer from poor temporal resolution and wideband spectrograms suffer from poor frequency resolution. The WD plots used in the experiment have a frequency resolution equivalent to that of a narrowband spectrogram and a temporal resolution equivalent to that of a wideband spectrogram. It is possible to alter the smoothing parameters dynamically, depending on some arbitrary parameter specifying the non-stationarity of the signal. It seems reasonable to assume that the auditory system analyzes transient signals with fine temporal resolution and stationary signals with fine spectral resolution. A similar strategy in computing and displaying the WD plots might enable the visualization of invariant acoustic patterns in speech with greater facility, and thus a closer match with peripheral auditory analysis.
Appendix A

Data Acquisition

In this appendix, the setup used for acquiring speech data in this thesis is presented.

A.1 Speech Signals

A.1.1 Data for Pilot Study

In the pilot study, three native speakers of English, French and Telugu (JG, AP and HG respectively), participated in recording the data. The informants spoke a list of CV syllables from a typed written page. The list contained each of the consonants shown in Table 4.1, except the dentals of Telugu, followed by the vowels [i, a, u]. There were 18, 18 and 36 stop consonants from English, French and Telugu. The subjects were asked to pause before each CV syllable in order to minimize the co-articulation effect. Two sets of the CV syllables were recorded and the second list was chosen for digitization.

A.1.2 Data for the Experiment

In the final experiment, six informants participated in recording the data; JG and DG for English, AP and JB for French and HG and CS for Telugu. The informants were provided with a list of CV syllables containing the nasal consonant /m/, all the consonants of their language shown in Table 4.1, and the nasal consonant /n/ followed by the vowels [i, e, a, o, u]. The list was recorded in triplicate for each informant. The last two sets were chosen for digitization. The unaspirated,
voiceless tokens from this set were chosen for final analysis. This resulted in 60 tokens each for English and French and 80 tokens for Telugu (including 20 dental stop consonants).

A.2 Instrumentation

A.2.1 Recording Setup

The analog recordings were made in a sound-proof room on a Revox tape recorder. An AKG microphone was used for recording the pilot data. A monotone signal at 120 Hz was played through the headphones and subjects (informants) were requested to try and maintain constant pitch throughout the recording.

A.2.2 Digitization

The data (i.e. speech segments) were low-pass filtered to 4.8 kHz using a Krohn-Hite filter and digitized using a LPA-11 I/O peripheral on a VAX-11/750 system. The data was digitized at 10 kHz sampling frequency. A rudimentary waveform editing program was developed to capture the burst and the beginning of the vowel. The program was capable of scanning the data for the burst, plotting different sections of the data and playing them through a D/A convertor. The digitized data was edited to extract 153.2 msec. of the signal covering the burst and the vowel. The edited data was transferred to the UBC General main frame computer on magnetic tape.

A.2.3 Computing and Plotting the Wigner distribution

On the main frame, the discrete time signal of each token was passed through a Hilbert transformer, as explained in chapter 3, to form a complex analytic signal. The names of the complex data files were formed with the initials of the speaker, the session number and the intended utterance. For instance, the data file JG1GA referred to speaker JG, session #1 and the intended utterance was syllable [ga].

An extensive menu-driven FORTRAN program called WIG4C.F was developed.
in order to conveniently compute and plot the WD. The program has features which modify the WD parameters and the plotting parameters interactively, and then plots the results on a terminal, or a laser printer, or stores the output for later use. The program can also be used in batch mode for computing the WD of large corpora economically. The usage of WIG4C.F and the parameters used in computing the WD of the speech data in this experiment are discussed in Appendix B.

A.2.4 Data Acquisition for the Cross-Language Study

The data was digitized on a PDP 11/23 and transferred to a HP9000 computer on a HPIB link for editing. A basic waveform editing facility was developed on the HP9000 to locate the burst. The digitized data was transferred back and forth between the PDP and HP9000 for audio feedback during the editing to locate the burst. The edited data of each utterance was stored on the tape before proceeding to the next utterance due to space constraints on the hard disk of the HP9000. The voiced and voiceless unaspirated consonants were digitized and edited. The data was transferred from the HP9000 to VAX and converted from UNIX format to ASCII. The ASCII data were then transferred to the main frame for computing and plotting the WD. The procedure for preprocessing the data using a Hilbert transformer and computing the WD was similar to that used in preparing the pilot data.

It should be that a lack of proper A/D and D/A facilities made consideration of a larger corpus impossible. The procedures described above, i.e., transferring the data between the PDP 11 and the HP9000 many times for each token, and then converting the data to VAX format for exporting the data to the UBCG main frame was extremely time consuming and costly.
Appendix B

The Program for computing the Wigner Distribution

An extensive menu-driven FORTRAN program WIG4C.F was developed to compute and plot the WD conveniently. The WIG4C.F program can be run either interactively to experiment with the various parameter values, or in batch mode to compute the WD of large corpora economically. The program was developed based on the principles presented in Chapter 3. This appendix concerns the usage of WIG4C.F to compute and plot the WD of a given signal. The choice of various parameter values used in computing the WD of the corpus is also explained. This appendix is intended as a manual for the version of WIG4C.F on UBC MTS mainframe computer.

The program has three major sections — the shell, the WD section, and the plotting section. The shell forms the interface between the user and the other sections of the program. The prompt CMD> is displayed in the shell. Figure B.1 depicts the help screen of the shell. The help command displays all the commands available to the user. The current values of the Wigner and the plot parameters can be displayed using the DW (display Wigner) and the DP (display plot) commands respectively. Figures B.2 and B.3 show the default values of various parameters when the program is started. The commands MW (modify Wigner) and MP (modify plot) can be used to modify the 11 Wigner parameters and 16 plot parameters. During the modification of the Wigner and plot parameters, the prompt is changed to MODWIG> and MODPLT> respectively. In order to change a parameter, the user needs to type in the serial number of that parameter, upon which the system
#run wig4c.o**displa
#Execution begins 14:47:53

Wigner Distribution.
Version 4.c

Type H for Help
CMD>help
To DISPLAY the Wigner parameters type .......... DW
To DISPLAY the Plot parameters type ............ DP
To MODIFY the Wigner parameters type .......... MW
To MODIFY the Plot parameters type ............ MP
To COMPUTE the Wigner Distribution type ....... C
To PLOT 3-D Perspective view type .............. PP
To PLOT the Contours type .................... PC
To PLOT the Input signal type .................. PI
To PLOT All the three plots type .............. PA
To STOP the execution of the program type ....... S
To get this HELP facility type ................ H
To go to MTS type ......................... # or E or 0.

CMD>

Figure B.1: Help Screen of WIG4C

CMD>dw
1) Input file ...................... = -1#
2) Length of Segment .......... = 384 points
3) Starting point .............. = 1
4) WD starts at ............... 66 (window center)
5) Computing WD till ......... 320 (window center)
6) WD computed every ......... 10 points
7) Frequency range spec ...... = 128 points
8) High-shaping (Y/N) ........ = Y
9) Time smoothing (Y/N) ...... = Y
10) Output file ................. = -WIG1
11) Sampling Frequency ...... = 10000

Smear in the frequency direction = 117.19 Hz.
Smear in the time direction ...... 2.1 ms.
OMEGA * T ....................... 0.25
Smoothing occurs over 21 data points

The identification code for this computation is
-1# 384/1/65-320:10# 128X128/256# 117/21/26# 200(A)#

CMD>

Figure B.2: Default Wigner Parameters

111
Figure B.3: Default Plot Parameters

prompts for the desired value. The display command D can be used to check the current status of the parameters within the modify routines. A brief description of these parameters along with the typical values used in computing the WD of the corpus is given here.

Wigner Parameters

1. The file which holds the complex analytic version of the signal for which the WD is to be computed.

2. The time duration of the signal (in number of samples).

3. The starting point of the data in the file.

4. The first point in the data file for which the WD is desired. This is normally chosen as \( \text{IF}/2 + 1 \) where \( \text{IF} \) is the window length used in computing the WD, since to compute the WD at a given point, \( \text{IF}/2 \) past samples and \( \text{IF}/2 \) future samples are needed. If the past and future samples are not available, the program pads the data with zeros.

5. The last point in the data file for which the WD is desired. This is normally chosen as the duration of the signal, less \( \text{IF}/2 \).
6. This parameter specifies the decimation of the WD in time. Generally, its value should not exceed the larger of
   
   • the duration of quasi-stationarity in the signal or
   • the amount of time-domain smoothing in computing the smoothed WD.

   In the present study, one WD was computed for every 10 samples (i.e., 1 msec. for a sampling frequency of 10 kHz).

7. The length of the time-domain window IF, in number of samples (usually chosen as a power of 2). This parameter relates to the extent of smear in the frequency domain. In the present study, IF was chosen as 128.

8. A Boolean variable to specify pre-emphasis on the data. The speech signal in the corpus was pre-emphasized prior to computing the WD.

9. A Boolean variable to switch on or off the time-domain smoothing. In order to change the current level of smoothing, the parameter has to be reset to 'Y'. The program then computes the time duration required for $T\Omega = 1$ smoothing and prompts the user for the desired amount of smear in the time domain. In computing the WD of the corpus, 2.1 msec. of smear in the time domain was chosen corresponding to $T\Omega = 0.25$.

10. The output file to store the WD. The program provides -WIG1, -WIG2, ... as default files for every new computation.

11. The sampling frequency. In the present study, the signal was sampled at 10 kHz.

**Plot Parameters**

1. The file containing the WD of the input signal as a 2-D array. The default is the file in which the results of the most recent WD computation were stored.

2. The dimensions of the array in the input file (updated every time a new WD computation is performed).
3. The time duration of the signal (in number of samples). This information is necessary to setup the correct axes in plotting the WD. It holds the value corresponding to the default input file after a new computation.

4. A Boolean variable to delete the negative regions in the WD.

5. A Boolean variable to switch on or off a pseudo-log operation on the data prior to plotting.

6. A constant for scaling the normalised WD data prior to computing the log. Values in the range of 100 to 500 were found useful in amplifying the higher formants, while keeping the level of cross-terms low.

7. The azimuth view angle of the camera for the 3-D plot. The default value is $-60^\circ$.

8. The elevation angle for the camera. The default value is $60^\circ$.

9. The number of contours to be used in the contour map. The default value is 13.

10. The output device for the plots. The plot output can be produced on one of the following devices:
   
   - T — Terminal (default)
   - C — Compressed plot file
   - Q — QMS - 1200 Lasergraphics laser printer
   - P — Post Script file for plotting on the QMS - ps800 laser printer

11. A Boolean variable to get the frequency axis from $-\frac{\pi}{2}$ to $\frac{\pi}{2}$ instead of from 0 to $\pi$.


13. Legend for the plot. When a new WD is computed, the Wigner section sets the legend with the values of some of the parameters. This legend can be overwritten by the user. There are six fields of entry in the computed legend,
each separated by a ‘#’ sign. The entries within a field are separated by a ‘/’. Each field of entry in the legend is explained below.

(a) Name of the file containing the time domain signal.
(b) Wigner parameters 2, 3, 4, 5 and 6.
(c) Plot parameters 2 and 3.
(d) Frequency smear in Hz, time smear in number of samples, and $T\Omega$ in percentage.
(e) The log scaling factor followed by plot parameter 16.
(f) The peak value in the WD, prior to scaling and the log operation.

14. A Boolean variable to invoke an editor. This editor was used to set given regions in the time-frequency plane to 0. This parameter was built in to analyze somatosensory evoked potentials using the WD. It was set to N (for ‘no’ suppression) in speech research.

15. A constant to scale the WD data without prior normalization. This enables one to compare the plots with one another, since the absolute amplitude relationships within a corpus are preserved. A value of 200 was used in computing the WD of both pilot data and final data.

16. A character variable to choose between normalised log and absolute log. The Absolute log was used in this study.

After modifying the Wigner and plot parameters to the desired values, the WD can be computed using the command C. This command also sets the plot parameters for plotting the current WD. There are 4 commands available for plotting the 3-D perspective view, the contour map, the input signal or all the three plots on a single plot page. The command PA was used in plotting the WD of the corpus. The input signal is plotted in the top frame, the contour map is plotted in the middle frame and the perspective view is plotted in the bottom frame. The computed legend is printed at the bottom of the page. See Figure 4.3 for an example. The user can go to MTS without unloading the program from the memory by typing #, E or Q. The command S is used to stop the execution of the program.
Appendix C

Instructions for Identifying Stop Consonants from the Wigner Plots

C.1 Preamble

It is believed that there are “invariant acoustic features” present in the speech signal which correspond to a given phone. These features are expected to manifest themselves in the speech signal in the form of unique time-frequency patterns, across all languages and in all phonetic contexts. This research is directed towards establishing whether such features can be identified for stop consonants in the Wigner plots. The Wigner Distribution (WD) of bilabial, alveolar and velar stop consonants [b, p], [d, t] and [g, k] followed by the vowels [i], [e], [a], [o], and [u] were computed. The CV syllables were uttered by native speakers of English, French and Telugu. The Wigner plots you will be analyzing have three subplots referred to as Plot A, Plot B and Plot C. Plot A shows the speech wave in the top frame; Plot B and Plot C depict the WD of this signal as a contour map and a perspective plot, in the middle and bottom frames respectively. The signal duration in the plots is 51.2 ms, 78.6 ms, or 115.2 ms.

C.2 Overview of the Task

You will be requested to make rough estimates of the visual features of the WD plots. These features, comprising mainly of boundaries and shapes, are described below. Three templates are provided to facilitate the analysis. The program
prompts you for each of these features and stores your answers. Based on your estimates, the program then classifies the graph according to place of articulation of the consonant as bilabial, alveolar or velar. The experiment begins with your logging onto the PDP 11/73, and running the program FEATURE. The program keeps track of your earlier sessions so that you can start and stop the task whenever you want and work on the task at your convenience.

C.3 Feature Definitions

All the CV syllables presented here have a small segment of silence followed by a burst and the vowel region. In the first stage of the program, you will be identifying the start of the burst, start of the vowel and the frequency of the second formant of the vowel. In the second stage, you will be requested to identify the spectral shape within the burst region. The program will then determine whether this burst region corresponds to the release of a stop or some irrelevant acoustic artefact. If it corresponds to an artefact, you will be requested to ignore this region of the plot and identify another region in the burst. You will then be requested to estimate new values for the start of the burst, the start of the vowel and the spectral shape again. This process will be repeated until reliable burst cues are identified, or until you decide to move to the next plot. If a reliable spectral shape has been identified, you will be prompted for the burst duration. The program will then identify the place of articulation and proceed to the next plot. The features that you will be measuring are defined below. Do not hesitate to refer back to these definitions whenever you feel the necessity.

1) Start of the Burst The start of the burst is the time at which peaks start to appear in the contour and the 3-D plots. The beginning of the burst can also be seen from the speech plot (Plot A) where the signal departs from zero and becomes noise-like. Generally, it is easiest to measure the start of the burst on the contour map (Plot B). See Figures 1, 2, 3 and 4.¹

¹Figures 1, 2, 3 and 4 are included in chapter 4 as Figures 4.1, 4.3, 4.4 and 4.2.
2) **Start of the Vowel** The start of the vowel is characterized by the appearance of a well defined formant structure, in Plots B and C. Typically, the formant structure has the following features:

- At the beginning of each pitch period, a spectral line where the troughs are not as deep appears. This can usually be observed in both Plot B and Plot C.

- 2 to 5 peaks distributed in frequency and extending in time (formants).

Sometimes, the burst can excite the vocal tract, resulting in a formant-like structure in the WD plots. This formant-like structure looks a little different from the formant structure excited by phonation alone. *If you notice this in the WD plots, take the beginning of the formants due to phonation as the onset of formant structure.* At this time instant, a large downward deflection can normally be observed in Plot A. See Figures 1, 2, 3 and 4.

3) **Second formant** The second formant frequency can be measured either in Plot B or in Plot C using the appropriate template. In Plot B, you should align the grid on the successive peaks corresponding to the second formant and read the corresponding frequency. If you use Plot C, superimpose the vertical axis of the template (on Plot C) to its homologue on the graph until the last spectral line on the template is at a tangent to the top of the second formant on the last spectral line of the graph. The second formant frequency can then be read off the frequency axis of Plot C. This will become clearer once it has been demonstrated to you in the training session.

4) **Spectral Shape** The spectral shape of the burst region can be classified into three classes. When you are prompted as to whether the spectral shape of the burst falls into one of these three categories, please answer “Y” for yes and “N” for no. For the three possible spectral shapes, the typical visual cues are listed below.

---

2The second formant measurement could sometimes be tricky both in B and C. In Plot B, it is not always easy to differentiate the valleys from the peaks. In Plot C, the second formant peak in the last spectral line could be hidden by another peak in front of it. If the second formant measurement is ambiguous in both B and C, your best approximation adequate.
a) Compact Pattern

A compact pattern can be recognized by the presence of one or more of these visual features.

- A single major peak, above 700 Hz in the frequency domain. Usually, but not always, occupying a region of less than 1.5 kHz. Ignore the peak below 700 Hz, if there is one.

- More than one peak in the frequency domain, but one of these peaks is much larger in amplitude than the others, making the spectral shape appear more compact than diffuse. In this case, the number of contours in the tallest peak is at least twice that of the other peaks in Plot B.

- Several successive peaks in the time domain, near the frequency of second formant of the vowel. These peaks need not have less than 1.5 kHz spread in the frequency domain. This feature alone is sufficient for you to classify the spectral shape as compact.

- Sometimes, there is some energy in the low frequency region (0 to 700 Hz) due to pre-voicing. This is usually referred to as the pre-voicing bar. Unlike the peaks of the burst, the amplitude of the pre-voicing bar does not decrease prior to the start of the vowel. Sometimes, the pre-voicing bar can make the compact spectral shape look like a diffuse pattern. The pre-voicing bar should not be considered in determining the compact spectral shape.

See Figures 1 and 4.

b) Diffuse Non-Falling Pattern

A diffuse non-falling pattern can be recognized by the presence of one or more of these visual features.

- Two or more significant peaks in the burst spectrum, covering a range of 1.5 kHz or more.

- A decrease in the amplitude of the burst peaks in time, prior to the start of the vowel.
• Sometimes, there is some energy in the low frequency region (0 to 700 Hz) due to pre-voicing. This is usually referred to as the pre-voicing bar. Unlike the peaks of the burst, the amplitude of the pre-voicing bar does not decrease prior to the start of the vowel. In general, the amplitude of the pre-voicing bar is not large enough to make a diffuse pattern look like a falling pattern. The pre-voicing bar should not be considered in determining the diffuse non-falling spectral shape.

See Figure 2.

c) Diffuse Falling Pattern
A diffuse falling pattern can be recognized by the presence of one or more of these visual features.

• Rapid increase in energy in the low frequency region (0 to 700 Hz) over time.
• Often, no other peaks are present at higher frequencies.
• If there are peaks at higher frequencies, their amplitude is smaller than those of the low frequency peaks. Thus, the line joining the peaks in frequency domain is always falling (negative slope).
• There is no noticeable burst region. The signal appears to go without transition from silence into the vowel.

See Figure 3.

Once the spectral shape has been determined, the program asks a few questions to establish that the region being analyzed is the relevant portion of the burst and not an acoustic artefact.

5) Burst Duration (BD) BD is defined as the duration from the start to the end of the burst. Measurement of BD for each of the spectral shapes is outlined below:

• For compact spectra with a single peak in the time-domain, determine the tallest peak (choose the one with the largest number of peaks in Plot B). The end of the burst is the time at which the number of contours
is down from the maximum by half. When there are multiple peaks in the time-domain, the duration of the first peak will be taken as the BD. See Figures 1 and 4.

- For diffuse non-falling spectra, there is a decrease in the energy of the burst prior to the start of the vowel. Some of the peaks could be connected with the formants of the vowel. Choose the tallest peak in Plot B. The end of the burst is the time instant where the number of contours is down from the maximum by half. In situations where it is difficult to identify the end of the burst, it can be taken to be the time at which more than two peaks in the diffuse spectra disappear completely. See Figure 2.

- For diffuse falling spectra, the burst blends with the vowel. There is no noticeable decrease in the energy of the burst prior to the start of the vowel. For this case, the end of the burst coincides with the start of the vowel. The BD is taken to be the duration between the start of the burst and the start of the vowel. See Figure 3.

After reading the value of BD, the program proceeds to identify the place of articulation. At this point, you are given an option of entering a comment to be stored in the data-base along with your analysis. The program then starts to analyze another plot, or quits processing, depending on your choice.

In the training session, you will be familiarized with some typical time-frequency patterns in the WD of stop consonants. Running the program FEATURE and measuring various features of the WD plots using the templates will also be demonstrated. There are approximately 200 WD plots for which the above analysis is to be carried out.

C.4 Addendum

Some additional instructions and tips to use the program FEATURE are given here. These were compiled based on the analysis of the training session data. Modifications to FEATURE which affect the user interaction are also described.
• The maximum time reading of each Wigner plot and the second formant of all the utterances are stored in a file that can be accessed by FEATURE. This eliminates the need for the user to input this information to the program.

• In order to decide whether a peak aligns the second formant, a tolerance of about ±200 Hz is permissible. A larger tolerance could be considered, if the peak under consideration migrates into the second formant over this frequency range.

• Compact and Diffuse non-falling spectral shapes generally have a frequency spread of less than 1.5 kHz and greater than 1.5 kHz, respectively. But it is not necessary that they strictly confirm to the above range. You are encouraged to eye-ball the first few milliseconds of the burst and decide on the spectral shape, unless the burst happens to belong to a borderline situation.

• There were many errors in the training session, due to wrong identification of the Start of the Vowel (VST). In addition to the cues described in the original instructions, VST can also be identified by

  - An abrupt increase of energy in the low-frequency region (< 500 Hz).
  - Appearance of the second formant.

(See Plot # 22 of the training session).

• Another source of errors in the training session was wrong classification of the spectral shape. Some hand-drawn pictures are provided on the next page to illustrate various spectral shapes.\(^3\)

In general, you are requested to refer to the instructions, whenever in doubt. It is a good idea to review the instructions before starting a session, especially if it is more than a couple of days since your earlier session.

\(^3\)This figure has been reproduced as Figure 4.5 in chapter 4.
Bibliography


[32] The software developed in this thesis for computing and plotting the WD is being used by Dr. Eisen of the Faculty of Medicine at UBC to analyze EMG signals and Dr. Begleiter of State University of New York Medical College to analyze EEG signals. The author is also collaborating with Dr. Bhattacharya of ASEA, Milwaukee to analyze the transients in high voltage DC transmission lines. Dr. Clapton of the University of Michigan, Ann Arbor and Dr. Rhode of the University of Wisconsin have expressed interests in obtaining a portable version of this software.


