

Representing Time in Automated Speech Recognition

David Richard Llewellyn Davies

March 2002

A thesis submitted for the degree of Doctor of Philosophy
of The Australian National University

Unless otherwise indicated the work presented in this thesis is entirely my own.

D. Davies

Acknowledgements

I would like to thank all those friends and colleagues who have helped in so many ways to make the completion of this thesis possible. To the one who made the greatest sacrifice, my daughter Annwn, I can only offer my love and wish her well in her own endeavours. To my supervisor Bruce Millar my thanks for providing the initial opportunity for me to develop my interests in speech, for developing the resources that made attempting this work possible, for his insights into the complex interdisciplinary nature of speech research and for his consistent guidance and encouragement during the development of this thesis and in helping to develop coherence from chaos in its presentation. I am indebted to the Computer Sciences Laboratory for access to its excellent facilities and in particular to Arthur McGuffin, Dennis Andriolo, Joe Elso for their technical support, Michelle Moravec for her friendship and patient assistance with administrative matters and to Markus Hegland for his valuable discussions on Fourier Transforms. Finally, I would like to gratefully acknowledge the financial assistance of an Australian Postgraduate Award.

Contents

1	Introduction	1
2	Time in Speech and ASR	4
2.1	Introduction	4
2.1.1	The Potential of ASR	4
2.1.2	Motivations	5
2.1.3	Temporal perspectives	6
2.2	Time in speech production and perception	8
2.2.1	General considerations	8
2.2.2	Source Synchronous Timing	9
2.2.3	Phoneme Duration	10
2.2.4	Dynamical Cues in the Frequency Domain	17
2.3	Time and Hidden Markov Models	18
2.4	Time and Artificial Neural Networks	21
2.5	Coding temporal information	24
2.6	Serial/parallel processing	27
2.7	Summary of time in speech and ASR	28
3	Objectives and Rationales	30
3.1	Introduction	30
3.2	Spectral Analysis	33
3.3	Source Synchronous Framing	34
3.4	The parameter similarity length	35
3.5	Acoustic-Phonetic Association Tables	36
3.5.1	The choice	36
3.5.2	Table input bandwidth	37
3.5.3	Acoustic-phonetic association rankings	38
3.6	Experimental Objectives	39
3.6.1	Smoothing and outlier constraint	40
3.6.2	Similarity lengths and differentials	40
3.6.3	Similarity lengths and duration	41
3.6.4	Voice Onset Transitions	42
4	Systems and Methods	43
4.1	Introduction	43
4.2	Development Platform	43
4.3	SLab Structure and Operational Overview	44
4.4	The Speech Data	46
4.5	Source Synchronous Framing	47
4.6	The Fourier Transforms	50
4.7	The Formant Picker Algorithm	55
4.8	Base Acoustic Parameters	61
4.9	Acoustic Parameter Scaling	65
4.10	Temporal Parameter Modifiers	67

4.10.1	Time Differentials	67
4.10.2	The Parameter Similarity Length	68
4.10.3	Smoothing and outlier constraint	70
4.10.4	Other Modifiers	70
4.11	The Association Table	71
4.12	Parameter Histograms and Temporal Slices	73
4.13	Discrimination Potentials	76
5	Results and Analysis	78
5.1	Introduction	78
5.2	Impact of smoothing and outlier constraint	79
5.3	Similarity lengths and differentials	81
5.3.1	The temporal range of PSLs and differentials	81
5.3.2	Comparison of PSLs and Differentials	85
5.4	Similarity lengths and duration	87
5.4.1	Duration-PSL comparison for all phonemes	87
5.4.2	Duration in phonetic contexts	91
5.5	CV contexts and F2 transitions	97
6	Conclusions and Projections	102
6.1	Summary of results	102
6.1.1	Temporal resolution and smoothing	102
6.1.2	PSLs and time differentials	103
6.1.3	PSLs and duration	104
6.1.4	PSLs and duration in context	105
6.1.5	PSLs and F2 transitions	105
6.2	Methodological objectives	105
6.3	Future exploration of the acoustic space	107
6.4	Conclusions	108

Appendices

A1	SLab system	A1
A2	Data specifications	A5
A2.1	Phonetic labels	A5
A2.2	Speaker information	A7
A3	Algorithms	A8
A4	Frequency domain smoothing	A11
A5	Formant histograms	A12
A6	Tx and Tx.Sim time streams	A21
A7	Parameter smoothing and constraint	A24
A8	Diferential-PSL comparisons	A30

Bibliography B1

The leading idea which is present in all our researches, and which accompanies every fresh observation, the sound which to the ear of the student of Nature seems continually echoed from every part of her works, is - Time! Time! Time!

George P. Scrope, 1827

Abstract

This thesis explores the treatment of temporal information in Automated Speech Recognition. It reviews the study of time in speech perception and concludes that while some temporal information in the speech signal is of crucial value in the speech decoding process not all temporal information is relevant to decoding. We then review the representation of temporal information in the main automated recognition techniques: Hidden Markov Models and Artificial Neural Networks. We find that both techniques have difficulty representing the type of temporal information that is phonetically or phonologically significant in the speech signal.

In an attempt to improve this situation we explore the problem of representation of temporal information in the acoustic vectors commonly used to encode the speech acoustic signal in the front-ends of speech recognition systems. We attempt, where possible, to let the signal provide the temporal structure rather than imposing a fixed, clock-based timing framework. We develop a novel acoustic temporal parameter (the Parameter Similarity Length), a measure of temporal stability, that is tested against the time derivatives of acoustic parameters conventionally used in acoustic vectors.

While the Similarity Length is directly applicable to conventional recognition techniques our analysis of these techniques points to the development of an approach to recognition that might prove more flexible and efficient in its ability to deal with temporal issues. We outline the requirements of such a system and in conformity with these requirements, develop an approach to evaluating the Similarity Lengths. We find that the Similarity Lengths generally provide a stronger association between the acoustic vector space and the phonetic space than do conventional parameter derivatives.