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

Appendices

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.