Imperial College London

Final Year Project

Robust Speech Detection in High Levels of Background Noise

Author:
Marcin Baginski

Supervisor:
Mike Brookes

This report is submitted in fulfilment of the requirements
for the degree of *MEng Information Systems Engineering*
in the
Department of Electrical and Electronic Engineering
Imperial College London

November 2013

Declaration of Authorship

I, Marcin Baginski, declare that this thesis titled, 'Robust Speech Detection in High Levels of Background Noise' and the work presented in it are my own. I confirm that:

- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated
- Where I have consulted the published work of others, this is always clearly attributed
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work
- I have acknowledged all main sources of help
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself

Signed: ____

Date: ____

IMPERIAL COLLEGE LONDON

Abstract

Department of Electrical and Electronic Engineering

MEng Information Systems Engineering

Robust Speech Detection in High Levels of Background Noise

by Marcin Baginski

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Maecenas pretium sem nec nisi facilisis, vel consectetur libero rutrum. Curabitur rhoncus commodo leo, nec lobortis ante venenatis vehicula. Duis vel posuere risus. Nulla blandit risus elit, quis eleifend leo lacinia ut. Mauris rutrum vitae orci eu commodo. Suspendisse egestas, ipsum quis interdum rutrum, tortor lectus facilisis lorem, eget mollis ligula arcu at mi. Quisque accumsan orci magna, sit amet interdum lacus commodo non. Mauris elit magna, venenatis in auctor sit amet, laoreet in tortor. Suspendisse et dolor mattis, tempus libero sit amet, tempus lectus. Nunc in dolor et lorem dignissim elementum. Curabitur suscipit lectus lorem. Pellentesque ultrices venenatis neque, vitae consectetur arcu mattis sed. Quisque porta nisl elementum lacus mollis commodo. Sed vestibulum dolor sed lectus interdum eleifend. Quisque in libero ut augue blandit malesuada.

Acknowledgements

I would like to thank . . .

Contents

List of Figures

Chapter 1

Introduction

1.1 Voice Activity Detection

Voice Activity Detection (VAD) is a problem of separating parts of an audio recording which contain the presence of human voice from those which are only comprised of silence or the background noise. VAD is a trivial task in recordings which have high signal-to-noise (SNR) ratios, in which voice can be distinguished from noise simply by computing the short-time energy of all frames and setting an appropriate threshold for their classification. However, in real applications, the signal is almost always contaminated by some level of background noise which makes the VAD's performance to drop. VAD decision is especially difficult for the unvoiced phonemes [1] whose spectrum contains no periodicity and is often similar to the one of white noise [2].

There has been an active research in the VAD area from as early as 1975, when Rabiner and Sambur [3] proposed a VAD algorithm (then referred to as "algorithm for determining the endpoints of isolated utterances") based on the aforementioned short-time energy and the zero-crossing rate. This approach worked reasonably well for signals with SNR ratio on the order of 30 dB, however since then there has been a need for much better performance, including applications where algorithm robustness has to be achieved even at negative SNRs.

1.2 Applications

VAD is often a first step in many signal processing applications including speech recognition [4], speech coding [5], speech enhancement [6] or noise estimation [4].

In Automatic Speech Recognition (ASR), it is important to first extract the voice-active parts of a signal and then pass them to the actual recognition module. This procedure increases both the accuracy of the ASR system as well as its speed, since the recognition task is not performed on the parts of the signal which do not contain speech. An example structure of a ASR system which uses VAD module is presented in Figure 1.1 [4]. For this application, it is most important for the VAD module to be able to identify all speech segments, even if some of the returned frames are false positives. Typically, there is a trade-off in VAD performance which can be characterised as maximising the precision of the VAD decisions while keeping the recall at a high rate.

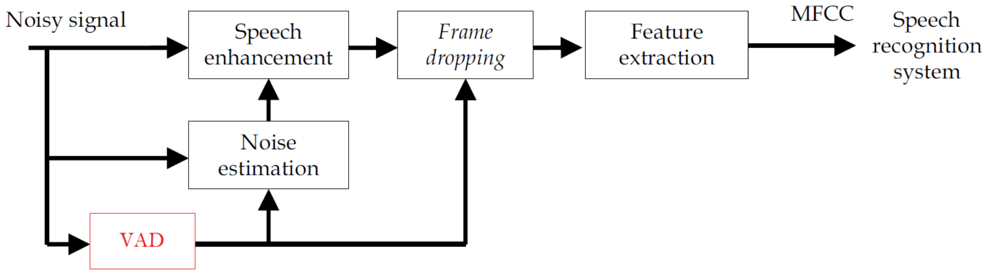

Figure 1.1: Block diagram of an ASR system which uses VAD as the first processing step [4]

Variable-rate speech coding techniques and discontinuous transmission (DTX) [7] systems also benefit from robust Voice Activity Detection. For DTX,

Bibliography

[1] A. M. Kondoz. *Digital Speech. Coding for Low Bit Rate Communication Systems*. John Wiley & Sons, 2004.

[2] P. R. Michaelis. Human Speech Digitization and Compression. In W. Karwowski, editor, *International Encyclopedia of Ergonomics and Human Factors*. CRC Press, 2006.

[3] L. R. Rabiner and M. R. Sambur. An Algorithm for Determining the Endpoints of Isolated Utterances. *The Bell System Technical Journal*, February 1975.

[4] J. Ramirez, J. M. Gorriz, and J. C. Segura. *Voice Activity Detection. Fundamentals and Speech Recognition System Robustness, Robust Speech Recognition and Understanding*. InTech, 2007.

[5] J. Sohn, N. S. Kim, and W. Sung. A Statistical Model-Based Voice Activity Detection. *IEEE Signal Processing Letters*, January 1999.

[6] Y. Park and S. Lee. Speech enhancement through voice activity detection using speech absence probability based on Teager energy, 2013.

[7] C. B. Southcott, D. Freeman, G. Cosier, D. Sereno, A. van der Krogt, A. Gilloire, and H. J. Braun. Voice Control of the Pan-European Digital Mobile Radio System. *Global Telecommunications Conference and Exhibition 'Communications Technology for the 1990s and Beyond' (GLOBECOM)*, 1989.