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 process of identifying parts of an audio recording which contain the presence of human voice as opposed to those which are only comprised of silence or the background noise. VAD is a relatively simple task in recordings which have high signal-to-noise ratios (SNR), 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 corrupted to some extent by background noise which makes the VAD performance to deteriorate. 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 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. Recently, numerous VAD approaches have been proposed, based on various features such as ....

1.2 Applications of VAD

VAD is often the first step in many signal processing applications including speech recognition [4–8], speech coding and transmission [4, 9–12], speech enhancement [4, 13, 14], noise estimation [4] or speaker recognition [15]. In most applications the noise-robust VAD decisions reduce the computational load required by the system and improve its accuracy. The reduced computational load is achieved since the voice-inactive frames are often not processed at all. At the same time, the clear boundaries of an utterance help to improve the accuracy of some systems (e.g. speech recognition).

1.2.1 Automatic Speech Recognition

In Automatic Speech Recognition (ASR), it is of importance to first extract the voice-active parts of a signal which can then be passed 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. A sample block diagram of an ASR system which uses a VAD module is presented in Figure 1.1 [4]. For ASR, and also most other applications, it is crucial for the VAD module to be able to identify all speech segments in order not to degrade the accuracy of the entire ASR system. Therefore, VAD systems often implement a fail-safe approach which means that if there is an uncertainty in classification of a frame, it is safer to label is as speech than otherwise. Typically, there is a trade-off in VAD performance which can be characterised as maximising the precision while keeping the recall at a steady, high rate.

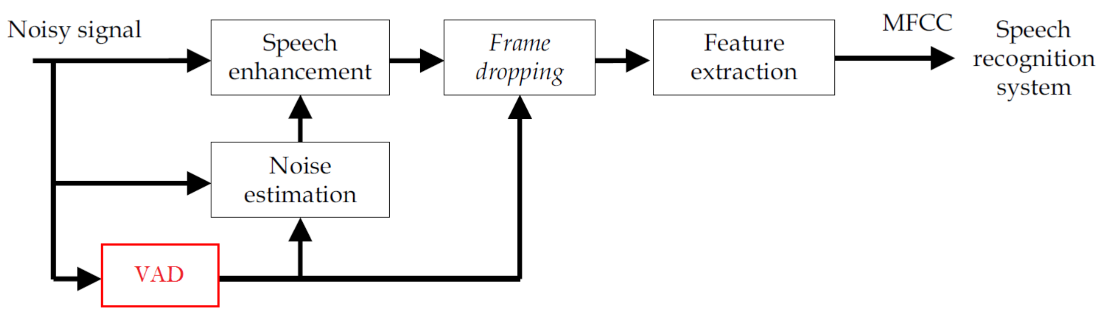

FIGURE 1.1: Block diagram of an Automatic Speech Recognition system with Voice Activity Detection module [4]

1.2.2 Speech Coding and Transmission

A typical phone conversation involves each person speaking on average no more than 50% of the time [12]. Using this fact it can be concluded that signal transmission would be greatly optimised if each transmitter was switched-off half of the time. Such approach could cause the overall system capacity to double. The technique of interrupted transmission during periods of silence is known as discontinuous transmission (DTX). In order to work properly, it requires a precise Voice Activity Detection to direct the operation of a transmitter between being switched on or off. As an alternate method to stopping the transmission, a dual-mode encoding technique could be employed, which uses a higher bit-rate for coding the voice-active frames and lower for silence/noise. The latter is precisely what the popular ITU-T G.729 Annex B [11] standard does, transmitting the voice-active parts at a fixed bit rate of 8 kb/s while the noisy ones at only 15 b/frame.

Figure 1.2 shows a structure of a dual-mode coding and transmission system, in which the VAD module is used to direct the incoming signal into either the active or inactive speech encoder. The noise can be either transmitted at a much lower bit-rate or the transmission might be switched off completely. In case of a stopped transmission, the receiving end often implements a *comfort noise* [4, 11, 12] generation module, which creates a synthetic signal similar to the background noise at the transmitter so that the listener does not notice the rapid, inconvenient switching during the conversation.

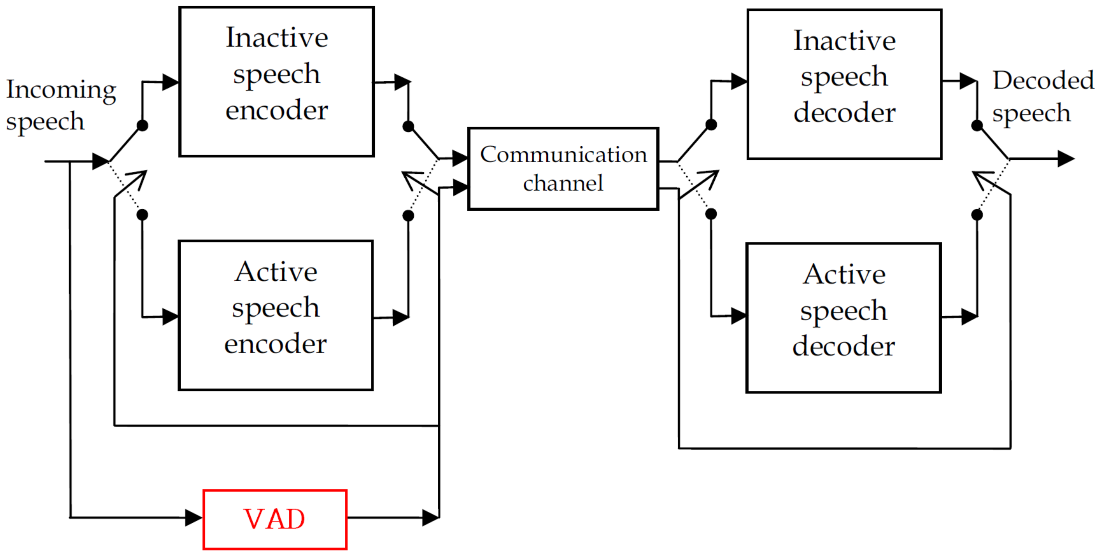

FIGURE 1.2: Block diagram of an dual-mode transmission system with Voice Activity Detection module [11]

1.2.3 Noise Estimation and Speech Enhancement

Speech enhancement aims to improve the intelligibility and quality of speech signals corrupted by additive noise of some kind. Many speech enhancement systems use a technique called *spectral subtraction* [1, 4]. It assumes, that the clean speech can be represented in the frequency-domain in the form:

$$|S(f)| = |Y(f)| - |N(f)| \tag{1.1}$$

where $|Y(f)|$ is the amplitude spectrum of the corrupted speech, $|S(f)|$ of the clean speech and $|N(f)|$ of the noise. In order for this technique to work, the noise needs to be additive, stationary and uncorrelated with the clean speech signal. Additionally, one needs to estimate the spectrum of the noise, which in real-world applications where a variety of different, often nonstationary, noise types are encountered, is a nontrivial problem. A robust Voice Activity Detector can become very useful in this task by identifying the voice-inactive frames of a signal from which the noise statistics could be estimated. A precise VAD can also become useful in applications dealing with slowly varying piecewise stationary noises, where the noise statistics can be adaptively estimated based on the most recent VAD decisions.

1.3 Structure of a typical VAD system

Figure 1.3 shows a high-level structure of a Voice Activity Detector, however only the two middle blocks are considered a core of a typical VAD system.

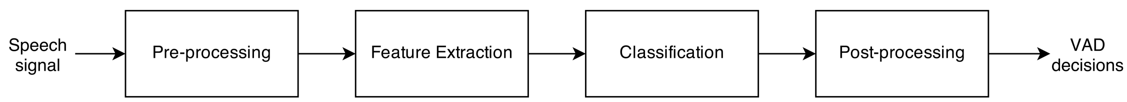

FIGURE 1.3: Block diagram of a typical Voice Activity Detection system

1.3.1 Pre-processing

Describe pre-processing

1.3.2 Feature Extraction

Describe feature extraction

1.3.3 Classification

Describe classification

1.3.4 Post-processing

Describe post-processing

The noisy speech signal is passed to a pre-processing module which might perform a variety of tasks, including noise suppression in order to improve the performance of the actual VAD algorithm. Additionally, during pre-processing the signal is often split into frames which are typically 20-100 ms long. In the next step, a number of features is calculated for each frame. These can be time-domain, frequency-domain, cepstral-domain features or other, depending on the specific VAD algorithm. The extracted information is passed to a classification module, which makes a decision whether to label each frame as speech or non-speech. The classification can be based on a variety of decision rules, starting from simple thresholding to more advanced methods such as machine learning or statistics. The last block, post-processing, often tries to *smooth* the VAD decisions in order to reduce the number of false positives and false negatives. For example, if among 50 consecutive frames, each of 20 ms duration, only one is classified as speech, the post-processing module might change the decision for this particular frame, since it is highly unlikely for speech to be active during such short-time window.

1.4 Thesis organisation

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] S. Kuroiwa, M. Naito, S. Yamamoto, and N. Higuchi. Robust Speech Detection Method for Telephone Speech Recognition System. *Speech Communication*, 1999.

[6] A. Martin and L. Mauuary. Robust Speech/Non-Speech Detection Based on LDA-Derived Parameter and Voicing Parameter for Speech Recognition in Noisy Environments. *Speech Communication*, 2006.

[7] I. Shafran and R. Rose. Robust Speech Detection and Segmentation for Real-Time ASR Applications. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2003.

[8] J. M. Gorriz, J. Ramirez, E. W. Lang, G. C. Puntonet, and I. Turias. Improved Likelihood Ratio Test Based Voice Activity Detector Applied to Speech Recognition. *Speech Communication*, 2010.

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

[10] R. Venkatesha Prasad, A. Sangwan, H. S. Jamadagni, M. C. Chiranth, R. Sah, and V. Gaurav. Comparison of Voice Activity Detection Algorithms for VoIP. *Seventh International Symposium on Computers and Communications*, 2002.

[11] A. Benyassine, E. Shlomot, H. Y. Su, D. Massaloux, C. Lamblin, and J. P. Petit. ITU-T Recommendation G.729 Annex B: A Silence Compression Scheme for Use with G.729 Optimized for V.70 Digital Simultaneous Voice and Data Applications. *IEEE Communications Magazine*, 1997.

[12] 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.

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

[14] K. R. Borisagar, D. G. Kamdar, B. S. Sedani, and G. R. Kulkarni. Speech Enhancement in Noisy Environment Using Voice Activity Detection and Wavelet Thresholding. *IEEE International Conference on Computational Intelligence and Computing Research*, 2010.

[15] M. Sahidullah and G. Saha. Comparison of Speech Activity Detection Techniques for Speaker Recognition, 2012.