UNIVERSITEIT•STELLENBOSCH•UNIVERSITY
jou kennisvennoot • your knowledge partner

Using computer vision to process speech

Marthinus Bosman
18955185

Report submitted in partial fulfilment of the requirements of the module Project (E) 448 for the degree Baccalaureus in Engineering in the Department of Electrical and Electronic Engineering at Stellenbosch University.

Supervisor: Dr. H. Kamper

October 2018

Acknowledgements

I would like to thank my dog, Muffin. I also would like to thank the inventor of the incubator; without him/her, I would not be here. Finally, I would like to thank Dr. Kamper for this amazing report template.

UNIVERSITEIT•STELLENBOSCH•UNIVERSITY
jou kennisvennoot • your knowledge partner

Plagiaatverklaring / *Plagiarism Declaration*

1. Plagiaat is die oorneem en gebruik van die idees, materiaal en ander intellektuele eiendom van ander persone asof dit jou eie werk is.
   Plagiarism is the use of ideas, material and other intellectual property of anothers work and to present is as my own.

2. Ek erken dat die pleeg van plagiaat 'n strafbare oortreding is aangesien dit n vorm van diefstal is.
   I agree that plagiarism is a punishable offence because it constitutes theft.

3. Ek verstaan ook dat direkte vertalings plagiaat is.
   I also understand that direct translations are plagiarism.

4. Dienooreenkomstig is alle aanhalings en bydraes vanuit enige bron (ingesluit die internet) volledig verwys (erken). Ek erken dat die woordelikse aanhaal van teks sonder aanhalingstekens (selfs al word die bron volledig erken) plagiaat is.
   Accordingly all quotations and contributions from any source whatsoever (including the internet) have been cited fully. I understand that the reproduction of text without quotation marks (even when the source is cited) is plagiarism.

5. Ek verklaar dat die werk in hierdie skryfstuk vervat, behalwe waar anders aangedui, my eie oorspronklike werk is en dat ek dit nie vantevore in die geheel of gedeeltelik ingehandig het vir bepunting in hierdie module/werkstuk of n ander module/werkstuk nie.
   I declare that the work contained in this assignment, except where otherwise stated, is my original work and that I have not previously (in its entirety or in part) submitted it for grading in this module/assignment or another module/assignment.

Studentenommer / *Student number*	**Handtekening / *Signature***
Voorletters en van / *Initials and surname*	**Datum / *Date***

Abstract

English

One of the fundamental difficulties in speech recognition is the task of extracting useful features from the highly variable time domain signal due to different speakers, tones, channels and acoustic conditions. However, In most state-of-the-art computer vision systems, convolutional neural networks are used to automatically learn how to extract relevant features. In this study, we aim to evaluate how general these features are. Specifically, we evaluate the features extracted from a trained vision CNN on speech spectrograms against existing techniques such as filter banks and MFCCs. *Our feature extraction technique showed a X% relative improvement over existing techniques.* Furthermore, we present some insight into the features extracted by the model.

Afrikaans

Die Afrikaanse uittreksel.

Contents

List of Figures

List of Tables

Nomenclature

Variables and functions

$p(x)$	Probability density function with respect to variable x.
$P(A)$	Probability of event A occurring.
ε	The Bayes error.
ε_u	The Bhattacharyya bound.
B	The Bhattacharyya distance.
s	An HMM state. A subscript is used to refer to a particular state, e.g. s_i refers to the i^{th} state of an HMM.
$\mathbf{S}$	A set of HMM states.
$\mathbf{F}$	A set of frames.
$\mathbf{o}_f$	Observation (feature) vector associated with frame f.
$\gamma_s(\mathbf{o}_f)$	A posteriori probability of the observation vector $\mathbf{o}_f$ being generated by HMM state s.
μ	Statistical mean vector.
Σ	Statistical covariance matrix.
$L(\mathbf{S})$	Log likelihood of the set of HMM states $\mathbf{S}$ generating the training set observation vectors assigned to the states in that set.
$\mathcal{N}(\mathbf{x}\|\mu, \Sigma)$	Multivariate Gaussian PDF with mean μ and covariance matrix Σ.
a_{ij}	The probability of a transition from HMM state s_i to state s_j.
N	Total number of frames or number of tokens, depending on the context.
D	Number of deletion errors.
I	Number of insertion errors.
S	Number of substitution errors.

Acronyms and abbreviations

AE	Afrikaans English
AID	Accent Identification
ASR	Automatic Speech Recognition
AST	African Speech Technology
BE	Black South African English
CE	Cape Flats English
DCD	Dialect-Context-Dependent
EE	White South African English
G2P	Grapheme to Phoneme
GMM	Gaussian Mixture Model
GPS	Global Phone Set
HMM	Hidden Markov Model
HTK	Hidden Markov Model Toolkit
IE	Indian South African English
IPA	International Phonetic Alphabet
LM	Language Model
LMS	Language Model Scaling Factor
LVCSR	Large Vocabulary Continuous Speech Recognition
MAP	Maximum a Posteriori
MFCC	Mel-Frequency Cepstral Coefficient
MLLR	Maximum Likelihood Linear Regression
MR	Multiroot
OOV	Out-of-Vocabulary
OR	One-Root
PD	Pronunciation Dictionary
PDF	Probability Density Function
SAE	South African English
SAMPA	Speech Assessment Methods Phonetic Alphabet

1 Introduction

intro paragraph

1.1. Motivations

1.2. Goals

1.3. Contributions

2 Existing Techniques and Models

intro

2.1. Speech Recognition

2.1.1. Speech Features

Spectrograms

Filterbanks

MFCCs

2.1.2. ML Models

2.2. Image Classification using cNNs

2.3. Summary of Existing Techniques

3 Feature Evaluation

3.1. Dynamic Time Warping

3.2. Same-different Speech Task

4 Experimental Setup

4.1. Dataset

4.2. Models

4.3. Visualization

5 Experiments

maybe something like:

5.1. First Layer

5.2. Second Layer

5.3. Third Layer

etc? (going deeper and deeper in the convNet to see how useful features are)

Will need to be something better, this is too simple

6 Summary and Conclusion

Bibliography

[1] G. E. Dahl, D. Yu, L. Deng, and A. Acero, “Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition,” *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 1, pp. 30–42, 2012.

[2] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-R. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury, “Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups,” *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 82–97, 2012.

A Sampling the segmentation

This is some appendix.

B Sampling using another bigram model

This is some other appendix.