Recurrent neural network-enhanced HMM speech recognition systems

by

Jan Willem Frederik Thirion

Submitted in partial fulfillment of the requirements for the degree

Master of Engineering (Electronics)

in the

Faculty of Engineering

UNIVERSITY OF PRETORIA

October 2001

To Carina, who makes every moment in life special.

Abstract

Keywords: speech segmentation, speech recognition, recurrent neural networks, extended recurrent neural networks, bi-directional recurrent neural networks, hidden Markov models

Speech recognition is fast becoming an attractive form of communication between humans and machines, due to the recent advances made in the field. The technology is, however, far from being acceptable when used in an unlimited form. Many attempts have been made to overcome some of the many difficulties faced in the automatic recognition of human speech by machine.

This dissertation attempts to accomplish two ambitious goals. The first is the reliable, automatic segmentation of speech signals, in a speaker independent manner, where no higher level lexical knowledge is used in the process. The system is limited to segmentation in an off-line manner. The second is to improve the phoneme recognition accuracy of a state-of-the-art hidden Markov model (HMM) based recognition system, using the segmentation information. A new technique of incorporating segmentation information into the Viterbi decoding process is presented, and it is shown that this technique outperforms other attempts to include the segmentation probabilities.

The segmentation system consists of a bi-directional recurrent neural network (BRNN), also called an extended recurrent neural network (ERNN). In contrast to conventional recurrent neural networks, that only use speech vectors from the past, present and possibly a fixed window of the future, BRNNs use all of the past, present and future

speech vectors. The problem of the choice of the fixed size window is thus eliminated, at the cost of difficult real-time implementation. The BRNN can be trained using a modified form of backpropagation through time (BPTT). The input to the BRNN is the entire speech sentence and the two BRNN outputs are the probability that a boundary between phonemes occurs in that frame of speech, as well as the probability that no boundary occurs. The segmentation system segments the speech signal into phonemes, using only the speech signal, and no higher level lexical knowledge such as the sequence of phonemes.

The recognition system can incorporate the segmentation probabilities in one of two ways. The first is to modify the HMM transition probabilities by combining the HMM transition probabilities and the BRNN outputs. The second method, developed in this dissertation, involves the use of an adaptive word (phoneme) transition penalty. Previously, only a fixed transition penalty was used between words (phonemes). By making the transition penalty adaptive (based on segmentation information), the phoneme recognition performance can be significantly improved. It is also shown that the adaptive word transition penalty outperforms the HMM transition probability modification technique, used by others.

All of the experiments used in this dissertation are conducted on the TIMIT database, in order to provide a convenient way to compare the results to that of others in the field. The hidden Markov toolkit (HTK) from Cambridge University is used for all phoneme recognition experiments.

Uittreksel

Sleutelwoorde: spraaksegmentering, spraakherkenning, terugvoer neurale netwerke, uitgebreide terugvoer neurale netwerke, bidireksionele terugvoer neurale netwerke, verskuilde Markov modelle

Spraakherkenning is tans een van die mees populêre intervlakke tussen mens en masjien. Talle probleme word egter nog steeds met hierdie tegnologie ondervind, veral wanneer dit in 'n onbeperkte vorm gebruik word.

Hierdie verhandeling pak twee ambisieuse doelwitte aan. Die eerste is die betroubare, outomatiese segmentering van spraakseine. Die stelsel word beperk tot 'n aflyn taak. Die tweede doelwit is om die foneemherkenning akkuraatheid van 'n hoë verrigting verskuilde Markov model (VMM) gebaseerde herkenning stelsel te verbeter, deur van die segmentering inligting gebruik te maak. Hier word 'n nuwe tegniek, om die segmentering inligting in die Viterbi dekodering proses in te sluit, voorgestel en daar word gewys dat hierdie tegniek beter resultate lewer as ander metodes wat tans gebruik word.

Die segmenteringstelsel bestaan uit 'n bidireksionele terugvoer neurale netwerk (BTNN), ook 'n uitgebreide terugvoer neurale netwerk (UTNN) genoem. Konvensionele terugvoer neurale netwerke gebruik slegs spraakvektore van die verlede, hede en moontlik vaste venster van die toekoms. In kontras hiermee, gebruik BTNN'e al die inligting van die verlede, hede en die toekoms. Die probleem van die keuse van 'n vaste grootte venster word dus vermy ten koste van 'n moeilike intydse implementasie. Die BTNN kan afgerig word met 'n aangepaste weergawe van terugvoering deur tyd (TVDT). Die inset na die

BTNN is die volledige sin met spraak en die twee BTNN uitsette is die waarskynlikheid dat 'n grens tussen foneme in daardie raam van spraak voorkom, as ook die waarskynlikheid dat geen grens voorkom nie. Die segmenteringstelsel segmenteer die spraak in foneme deur slegs van die spraaksein gebruik te maak. Geen hoër vlak leksiese kennis, soos bv. die volgorde van foneme, word gebruik nie.

Die herkenningstelsel kan die segmentering waarskynlikhede in een van twee maniere insluit. Die eerste is om die VMM oorgang waarskynlikhede aan te pas deur die VMM oorgang waarskynlikhede en BTNN uitsette te kombineer. Die tweede metode, ontwikkel in hierdie verhandeling, maak van aanpasbare woord (foneem) oorgang penalisasie gebruik. Voorheen was slegs van vaste oorgang penalisasie tussen woorde (foneme) gebruik gemaak. Deur die oorgang penalisasie term aanpasbaar te maak (gebaseer op segmentering inligting), kan die foneemherkenning akkuraatheid aansienlik verhoog word. Daar word ook gewys dat aanpasbare woord oorgang penalisasie die VMM oorgang waarskynlikheid aanpassing tegniek, soos deur ander gebruik, oortref.

Al die eksperimente wat in hierdie verhandeling gedoen word, word op die TIMIT databasis gedoen, om sodoende 'n gerieflike manier daar te stel vir die vergelyking van resultate met dié van ander in die veld. Die stel van verskuilde Markov model programme (HTK) van Cambridge Universiteit word vir alle foneemherkenning eksperimente gebruik.

Acknowledgements

I would like to thank Professor E.C. Botha for allowing me the opportunity to study under her. She provided me with the academic guidance that made the work given here possible. It was an honour to learn from such a great intellectual.

Special thanks goes to Darryl Purnell for all the advice and discussions we had throughout the completion of my studies. His invaluable comments guided me towards achieving something that I am proud of. I also thank Willie Smit, with whom I had many discussions about concepts and ideas.

I am grateful to my family and friends, for their love, encouragement, and patience. I have great appreciation for my parents who coped with me, and always encouraged and supported me in my studies.

Finally, I would like to give my deepest thanks to my Saviour, Jesus Christ, for giving me the talents and gifts to be able to think.

Contents

List of Abbreviations

ADC	Analogue to digital converter
AM	Auditory modeling
AR	Autoregressive
ASP	Auditory speech processing
ASR	Automatic speech recognition
BCD	Bayesian changepoint detector
BRNN	Bi-directional recurrent neural network
CRNN	Chaotic recurrent neural network
DC	Direct current
DP	Dynamic programming
DTW	Dynamic time warping
ERNN	Extended recurrent neural network
FBDYN	Forward-backward auditory dynamic cepstra
FFT	Fast Fourier transform
HMM	Hidden Markov model
HTK	Hidden Markov toolkit
KL	Kullback Leibler
LP	Linear prediction
LPC	Linear predictive coefficient
MAP	Maximum a posteriori
MFC	Mel frequency cepstrum
MFCC	Mel frequency cepstrum coefficient
MIT	Massachusetts Institute of Technology
MLP	Multilayer perceptron
NIST	National Institute of Standards and Technology
NN	Neural network
OGI	Oregon Graduate Institute
PDF	Probability density function

PSM	Polynomial segment model
PSS	Pseudosyllables
RASTA	Relative spectral
RMLP	Recurrent multilayer perceptron
RMS	Root mean square
RNN	Recurrent neural network
S/M-R	Synchrony/mean-rate
SCVQ	Sequence-constrained vector quantisation
SDU	Semantic dialogue unit
SLAM	Segmentation and labeling automatic module
SRN	Simple recurrent network
SVF	Spectral variation function
TI	Texas Instruments
VCV	Vowel-consonant-vowel, or vocal-consonant-vocal
VQ	Vector quantisation