ĐẠI HỌC THÁI NGUYÊN
TRƯỜNG ĐẠI HỌC KỸ THUẬT CÔNG NGHIỆP

TRẦN VĂN NGHĨA

NGHIÊN CỨU ÁP DỤNG MÔ HÌNH MẠNG NƠ-RON END-TO-END CHO NHẬN DẠNG TIẾNG NÓI TIẾNG VIỆT

LUẬN VĂN THẠC SĨ KỸ THUẬT VIỄN THÔNG

THÁI NGUYÊN 2019

ĐẠI HỌC THÁI NGUYÊN
TRƯỜNG ĐẠI HỌC KỸ THUẬT CÔNG NGHIỆP

TRẦN VĂN NGHĨA

NGHIÊN CỨU ÁP DỤNG MÔ HÌNH MẠNG NƠ-RON END-TO-END CHO NHẬN DẠNG TIẾNG NÓI TIẾNG VIỆT

Chuyên ngành: Kỹ thuật viễn thông
Mã số: 8520208

LUẬN VĂN THẠC SĨ KỸ THUẬT VIỄN THÔNG

KHOA CHUYÊN MÔN
TRƯỞNG KHOA

NGƯỜI HƯỚNG DẪN KHOA HỌC

TS. NGUYỄN VĂN HUY

PHÒNG ĐÀO TẠO

THÁI NGUYÊN 2019

Lời nói đầu

Nhận dạng tiếng nói là mong ước của khoa học và con người. Những người máy có thể hiểu được tiếng người nói và thực thi nhiệm vụ theo mệnh lệnh người nói.

Các kỹ thuật nhận dạng tiếng nói đã và đang rất phát triển, đặc biệt với một số ngôn ngữ phổ dụng như Anh, Pháp, Trung Quốc,… Những yếu tố chính ảnh hưởng đến chất lượng của một hệ thống nhận dạng tiếng nói như: Người nói, tốc độ nói, hoàn cảnh nói, nhiễu, kích thước từ điển, cách thức phát âm,… tuy nhiên hiện nay vẫn chưa có một giải pháp nào hoàn thiện giải quyết tất cả các yếu tố đó. Các phương pháp cơ bản thường được sử dụng cho nhận dạng tiếng nói là: Kỹ thuật so khớp mẫu, mạng nơ-ron, phương pháp dựa trên tri thức và mô hình Markov ẩn. Trong đó phương pháp sử dụng mô hình Markov ẩn (Hidden Markov Model HMM) được sử dụng phổ biến nhất.

Đối với tiếng Việt hiện nay vẫn chưa thực sự được nghiên cứu rộng rãi về nhận dạng. Các công việc nghiên cứu mới đang ở những bài toán cơ bản. Tiếng Việt là một ngôn ngữ có thanh điệu, vì thế ngoài những khó khăn gặp phải tương tự như việc nhận dạng các ngôn ngữ không có thanh điệu khác (Anh, pháp,…), nhận dạng tiếng Việt còn phải nghiên cứu vấn đề nhận dạng thanh điệu. Tiếng Việt có sáu thanh điệu, một cách tổng quát có thể coi như mỗi âm tiết sẽ có thể có sáu ý nghĩa khác nhau khi ghép tương ứng với sáu thanh điệu đó. Việc nhận dạng thanh điệu là một công việc khó do thanh điệu chỉ tồn tại ở vùng âm hữu thanh. Vì thế đường đặc tính của nó không liên tục khi chuyển tiếp giữa hai vùng hữu thanh và vô thanh. Các đặc trưng được sử dụng phổ biến trong nhận dạng tiếng nói như MFCC (Mel Frequency Cepstral Coefficient) và PLP (Perceptual Linear Prediction) lại không mô tả được các đặc tính của thanh điệu, do vậy trước khi nhận dạng được thanh điệu ta phải áp dụng các kỹ thuật tính toán đặc trưng thanh điệu trong tín hiệu tiếng nói.

Khi áp dụng mô hình mạng nơ-ron (Deep Neural Network – DNN) cho nhận dạng tiếng Việt, cụ thể là trong quá trình trích chọn đặc trưng BottleNeck, đã giúp cải thiện chất lượng hệ thống nhận dạng. Tuy nhiên, nếu sử dụng mô

hình mạng nơ-ron truyền thống, các mô hình DNN này được huấn luyện trên tập dữ liệu đã được gán nhãn, sẽ cần tốn nhiều thời gian cho việc huấn luyện, và chất lượng mô hình huấn luyện phụ thuộc vào thủ tục liên kết các mô hình trong nó. Do mô hình truyền thống gồm ba phần chính: là mô hình phát âm (pronunciation model – PM), mô hình ngữ âm (acoustic model – AM) và mô hình ngôn ngữ (language model – LM), chúng được huấn luyện độc lập nhau.

Vì vậy, việc nghiên cứu loại mô hình mạng nơ-ron giúp tích hợp ba thành phần PM, AM và LM trong mô hình mạng nơ-ron truyền thống, vào một mô hình đơn nhất là cần thiết, và việc huấn luyện có thể thực hiện trực tiếp trên tập dữ liệu chưa được gán nhãn. Nghĩa là việc huấn luyện chỉ yêu cầu các file tiếng nói (audio file) và phiên âm của chúng – đây chính là mô hình End-to-End (E2E).

Xuất phát từ nhận thức trên, được sự gợi hướng của Thầy giáo, TS. Nguyễn Văn Huy, học viên xin trình bày luận văn tốt nghiệp Thạc sỹ chuyên ngành Kỹ thuật Viễn thông về **"Nghiên cứu áp dụng mô hình mạng nơ-ron End-to-End cho nhận dạng tiếng nói tiếng Việt".**

Nội dung chính của luận văn được trình bày thành 03 chương với bố cục như sau:

- **Chương 1**: Mở đầu. Giới thiệu tổng quan về nhận dạng tiếng nói và ứng dụng. Các vấn đề khó khăn cần giải quyết trong lĩnh vực nhận dạng tiếng nói. Giới thiệu tổng quan về tình hình nghiên cứu nhận dạng tiếng Việt trong và ngoài nước. Giới thiệu các nội dung nghiên cứu chính của luận văn.
- **Chương 2**: Mô hình mạng nơ-ron học sâu End-to-End cho nhận dạng tiếng nói. Giới thiệu về các thành phần cơ bản trong hệ thống nhận dạng tiếng nói từ vựng lớn. Mô hình dựa trên mạng nơ-ron học sâu (Deep Neural Network - DNN) cho nhận dạng tiếng nói. Phân loại mô hình mạng DNN truyền thống, mạng DNN End-to-End; và ứng dụng trong nhận dạng tiếng nói ngôn ngữ không phải tiếng Việt.

❖ **Chương 3**: Áp dụng mô hình mạng nơ-ron End-to-End cho nhận dạng tiếng Việt. Trình bày tổng quan về cấu trúc ngữ âm tiếng Việt, đề xuất cho việc nhận dạng tiếng nói tiếng Việt và thử nghiệm thực tế.

Tôi xin được gửi lời cảm ơn đặc biệt đến TS. Nguyễn Văn Huy, đã luôn chỉ bảo, định hướng, tạo điệu kiện thuận lợi nhất để tôi có thể hoàn thành luận văn này.

Thái Nguyên, ngày tháng năm 2019

Trần Văn Nghĩa

Lời cam đoan

Tôi xin cam đoan đây là công trình nghiên cứu của riêng tôi và được sự hướng dẫn khoa học của **TS. Nguyễn Văn Huy**. Các nội dung nghiên cứu, kết quả trong đề tài này là trung thực và có nguồn gốc rõ ràng. Những số liệu trong các bảng biểu phục vụ cho việc phân tích, nhận xét, đánh giá được thu thập từ các thử nghiệm thực tế.

Nếu phát hiện có bất kỳ sự gian lận nào tôi xin hoàn toàn chịu trách nhiệm về nội dung luận văn của mình.

Tác giả

Trần Văn Nghĩa

Mục lục

Danh mục các từ viết tắt

TT	Viết tắt	Nghĩa
1	AM	Acoustic Model
2	AMDF	Average Magnitude Difference Function
3	CNN	Convolution Neural Network
4	CP	Character-based Phoneset
5	CTC	Connectionist Temporal Classification
6	DCT	Discrete cosine transform
7	DFT	Discrete Fourier transform
8	DNN	Deep Neural Network
9	E2E	End-to-End
10	F0	Fundamental Frequency
11	FST	Finite-State Transducer
12	G2P	Grapheme to Phoneme
13	GMM	Gaussian Mixture Model
14	GPU	Graphical processing unit
15	HMM	Hidden Markov Model
16	IDFT	Invert Discrete Fourier transform
17	LDA	Linear Discriminant Analysis
18	LM	Language Model
19	LPC	Linear Prediction Coding
20	LSTM	Long Short-Term Memory
21	MFCC	Mel Frequency Cepstral Coefficients
22	MLLT	Maximum Likelihood Linear Transform
23	MLP	Multilayer Perceptron
24	MSD	Multispace Distribution
25	NCC	Normalized Cross-Correlation
26	NN	Neural Network
27	PLP	Perceptual Linear Prediction
28	PM	Pronunciation Model
29	PP	Phoneme-based Phoneset
30	RNN	Recurrent Neural Networks
31	T1	Tone 1
32	T2	Tone 2
33	T3	Tone 3

34	T4	Tone 4
35	T5	Tone 5
36	T6	Tone 6
37	TDNN	Time Delay Deep Neural Network
38	VN-G2P	Vietnamese Grapheme to Phoneme
39	WER	Word Error Rate
40	WT	phoneset Without Tone informations