Modelling Accents for Automatic Speech Recognition

Maryam Najafian, Martin Russell

School of EECE, University of Birmingham, Birmingham, UK

[mxn978,m.j.russell]@bham.ac.uk

Abstract

Accent is cited as an issue for speech recognition systems. If they are to be widely deployed, Automatic Speech Recognition (ASR) systems must deliver consistently high performance across user populations. Hence the development of accent-robust ASR is of significant importance. This research investigates techniques for compensating for the effects of accents on performance of Hidden Markov Model (HMM) based ASR systems. Recently, HMM systems based on Deep Neural Networks (DNNs) have achieved superior performance to more traditional systems based on Gaussian Mixture Models (GMMs), due to the discriminative nature of DNNs. Our research confirms, this by showing that a DNN system outperforms the GMM system even after an accent-dependent acoustic model was selected using Accent Identification (AID), followed by speaker adaptation. The average performance of the DNN system over all accent groups is maximized when either accent diversity is highest, or data from "difficult" accent-groups is included in the training set.

Index Terms: Multi-accent speech recognition, acoustic data selection, deep neural network

1. Research Summary

In the 'accent of English' book by Wells an accent is defined as a "a pattern of pronunciation used by a speaker for whom English is the native language or, more generally, by the community or social grouping to which he or she belongs". This differentiates accent from dialect, which includes the use of words or phrases that are characteristic of that community. It includes varieties of English spoken as a first language in different countries (for example, US vs Australian English), geographical variations within a country, and patterns of pronunciation associated with particular social or ethnic groups.

The recent growth in applications of ASR systems forces developers to consider approaches that deliver consistent high performance across different accent groups. It is of high importance for those systems to be able to deal with accented speakers.

Over the last decade, DNN based systems achieved superior accuracy compared to GMM based systems in many applications. This has been attributed to be due to their discriminative nature, and layer by layer invariant feature learning ability. This enhances their robustness against different sources of variation, for example accent. Various techniques, such as adding accent discriminative acoustic features, adding an accent-specific layer on top of a DNN acoustic model, accent-specific pronunciation adaptation, accent-specific polyphone decision trees, and acoustic model adaptation, were proposed in literature.

In this research we investigate, whether it is better to train a simple DNN system on a multi-accented data and rely on its layer-wise discriminative learning to learn different accent patterns in the training data, or it is better to do GMM based accent plus speaker adaptation. In the DNN system, we explored the effect of the amount of pre-training and training data and their accent diversity on the final performance. In our GMM system, we construct an accent-dependent acoustic model for 14 different British accents, and use Accent Identification (AID) to select the model that corresponds to a new speaker. For each new speaker we do accent-dependent acoustic model adaptation, followed by the speaker adaptation.

Results of our AID system shows that with 43s of speech an individual's accent can be determined with 81% accuracy using unsupervised AID (i-Vectors). Thus, a possible solution is to use AID for accent-dependent ASR model selection and then apply unsupervised speaker adaptation to the GMM-HMM system. In DNN-HMM systems it is possible to use this AID to analyse the accent diversity of the training data and investigates its effect on the final performance.

In our experiments we extracted the adaptation and test data from the Accents of British Isles (ABI-1) corpus containing data from 14 different regions of British Isles. These regional accents fall into four groups namely, Northern English, Southern English, Irish and Scottish.

We applied an AID system to explore the accent diversity of the WSJCAM0 speech corpus and realised that it consists of a range of Northern English, Southern English, Irish and Scottish accents. Using the WSJCAM0 corpus, we achieved a relative gain of 46.9% in recognizing the ABI-1 corpus by applying DNNs rather than GMMs for the acoustic modelling using the WSJCAM0 training data. This shows that DNN systems are better than GMMs in dealing with the multi-accented data.

A clear effect of accent is evident in the performance of our GMM-HMM speech recognition systems, even after applying multiple acoustic model adaptation. We observed that the accuracy of this GMM system, even after applying Maximum A Posterior adaptation to 40 minutes of accent-specific data, followed by unsupervised speaker adaptation using Maximum Likelihood Linear Regression with 43s (7.3%WER) of data, could not match the accuracy of our baseline DNN system (6.9%WER).

Despite, the major gain achieved by using the DNNs rather than GMMs in modelling the acoustic data, the effect of accent is still evident and the system doesn't perform uniformly well across all accent groups. Even in our best system (DNN-HMM) the percentage Word Error Rate (%WER) for the most challenging accent Glaswegian (13.34%) nearly 5 times bigger than that for standard southern English accent (2.84%).

Adding an accent-specific layer on top of a multi-accent neural network acoustic model is one potential solution and will be addressed in our future work. Although the work targets British English it is very likely that the techniques described are applicable to accented speech in other languages.

978-0-9928626-3-3/15/$31.00 ©2015 IEEE