Available online at www.sciencedirect.com

SciVerse ScienceDirect

Speech Communication 54 (2012) 801–813

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Multi-accent acoustic modelling of South African English

Herman Kamper, Félicien Jeje Muamba Mukanya, Thomas Niesler *

Department of Electrical and Electronic Engineering, Stellenbosch University, South Africa

Received 9 February 2011; received in revised form 25 January 2012; accepted 31 January 2012
Available online 9 February 2012

Abstract

Although English is spoken throughout South Africa it is most often used as a second or third language, resulting in several prevalent accents within the same population. When dealing with multiple accents in this under-resourced environment, automatic speech recognition (ASR) is complicated by the need to compile multiple, accent-specific speech corpora. We investigate how best to combine speech data from five South African accents of English in order to improve overall speech recognition performance. Three acoustic modelling approaches are considered: separate accent-specific models, accent-independent models obtained by pooling training data across accents, and multi-accent models. The latter approach extends the decision-tree clustering process normally used to construct tied-state hidden Markov models (HMMs) by allowing questions relating to accent. We find that multi-accent modelling outperforms accent-specific and accent-independent modelling in both phone and word recognition experiments, and that these improvements are statistically significant. Furthermore, we find that the relative merits of the accent-independent and accent-specific approaches depend on the particular accents involved. Multi-accent modelling therefore offers a mechanism by which speech recognition performance can be optimised automatically, and for hard decisions regarding which data to pool and which to separate to be avoided.

Keywords: Multi-accent acoustic modelling; Multi-accent speech recognition; Under-resourced languages; South African English accents

1. Introduction

Despite steady improvement in the performance of automatic speech recognition (ASR) systems in controlled environments, the accuracy of these systems still deteriorates markedly when confronted with highly accented speech. In countries with non-homogeneous populations, non-mother-tongue speech is highly prevalent. When the language in question is also under-resourced, it is important to know how best to make use of the limited speech resources to provide the best possible recognition performance in the prevalent accents.

The South African constitution gives official status to 11 different languages, as summarised in Fig. 1. Although English is the lingua franca as well as the language of government, commerce and science, only 8.2% of the population use it as a first language. Hence, English is used predominantly by non-mother-tongue speakers, resulting in a large number of accents. In general, these accents are not bound to geographic regions as is often the case for other world accents. South African English (SAE) therefore provides a challenging and relevant scenario for the modelling of accents in ASR. It also can be classified as an under-resourced variety of English since the annotated speech available for the development of ASR systems is exceedingly limited.

The research presented in this paper considers the question of how best to optimise HMM-based acoustic models when presented with a very limited corpus of different SAE accents.[1] Although the speech databases used in this

* Corresponding author. Tel.: +27 21 808 4118.
E-mail addresses: kamperh@sun.ac.za (H. Kamper), trn@sun.ac.za (T. Niesler).

[1] According to Crystal (1991), the term 'accent' refers only to pronunciation differences, while 'dialect' refers also to differences in grammar and vocabulary. It is not always obvious whether we are dealing with accents or with dialects when considering varieties of SAE. We will therefore use the term 'accent' exclusively to avoid confusion.

0167-6393/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2012.01.008

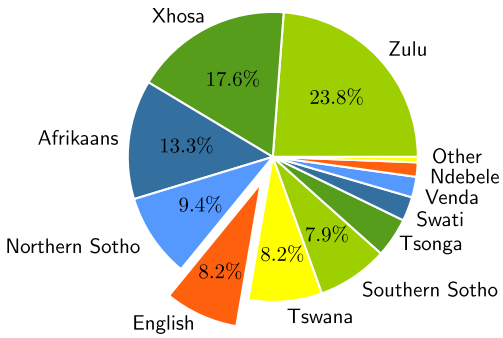

Fig. 1. Mother-tongue speakers of the 11 official languages in South Africa, as a percentage of the population (Statistics South Africa, 2004).

research are small compared to those used in state-of-the-art systems, the scenario considered here is representative of an under-resourced environment. In this environment the presence of multiple accents further aggravates the development of ASR technology.

2. Related research

Two main approaches are encountered when considering the literature dealing with multi-accent or multidialectal speech recognition. Some authors consider modelling accents as pronunciation variants which are added to the pronunciation dictionary employed by a speech recogniser (Humphries and Woodland, 1997). Other authors focus on multi-accent acoustic modelling. We will take the latter approach and begin by presenting a brief review.

2.1. Multi-accent acoustic modelling

A popular approach to multi-accent acoustic modelling is to pool data from all accents under consideration, resulting in a single accent-independent acoustic model set. An alternative is to train separate accent-specific systems that allow no sharing between accents. These two contrasting approaches have been considered and compared by many authors, including those whose work is summarised in Table 1. In most cases, accent-specific models lead to superior speech recognition performance when compared with accent-independent models. However, this is not always the case, as is demonstrated by Chengalvarayan (2001). The comparative merits of the two approaches appear to depend on factors such as the abundance of training data, the type of task and the degree of similarity between the accents involved.

In cases where the quantity of data is insufficient for the training of accent-specific models, adaptation techniques such as maximum likelihood linear regression (MLLR) and maximum a posteriori (MAP) adaptation can be employed. For example, MAP and MLLR have been successfully employed in the adaptation of Modern Standard Arabic acoustic models for improved recognition of Egyptian Conversational Arabic (Kirchhoff and Vergyri, 2005). The results obtained by Diakoloukas et al. (1997) suggest that training acoustic models on target accented data alone is superior to adaptation when larger amounts of accented data are available. However, Despres et al. (2009) found that accent-independent models which have been adapted with accented data outperformed both accent-specific and

Table 1
Literature comparing accent-specific and accent-independent modelling approaches, as well as various forms of adaptation.

Authors	Accents	Task	Training corpus	Best approach
Van Compernolle et al. (1991)	Dutch and Flemish	Isolated digit recognition	3993 Dutch and 4804 Flemish utterances	Accent-specific modelling
Beattie et al. (1995)	Three dialects of American English	Command and control (200 words)	Not indicated	Gender- and dialect-specific modelling
Fischer et al. (1998)	German and Austrian dialects	Large vocabulary continuous speech recognition	90 h German; 15 h Austrian speech	Accent-specific modelling
Chengalvarayan (2001)	American, Australian and British dialects of English	Connected digit recognition	7461 American, 5298 Australian and 2561 British digit strings	Accent-independent modelling
Caballero et al. (2009)	Five Spanish dialects (Spain, Argentina, Venezuela, Columbia, Mexico)	Isolated word recognition	50,000 Spanish utterances and 10,000 from each remaining dialect	Multidialect, followed by accent-independent modelling
Diakoloukas et al. (1997)	Stockholm and Scanian dialects of Swedish	Travel information task	21,000 Stockholm sentences; different amounts of Scanian adaptation data	Less data: adaptation; more data: accent-specific modelling
Wang et al. (2003)	Non-native English from German speakers	Spontaneous face-to-face dialogues	34 h native English; 52 min non-native adaptation data	Decision-tree-based adaptation, followed by MAP
Kirchhoff and Vergyri (2005)	Modern Standard Arabic and Egyptian Conversational Arabic	Large vocabulary continuous speech recognition	40 h Modern Standard Arabic; 20 h Egyptian Conversational Arabic	An approach employing both MAP and MLLR
Despres et al. (2009)	Northern and Southern dialects of Dutch	Broadcast news	100 h Northern Dutch; 50 h Southern Dutch	MAP