Merging of Native and Non-native Speech for Low-resource Accented ASR

Sarah Samson Juan[1(✉)], Laurent Besacier[2], Benjamin Lecouteux[2], and Tien-Ping Tan[3]

[1] Faculty of Computer Science and Information Technology, Universiti Malaysia Sarawak, Kota Samarahan, Sarawak, Malaysia
sjsflora@unimas.my
[2] Grenoble Informatics Laboratory (LIG), University Grenoble-Alpes, Grenoble, France
{laurent.besacier,benjamin.lecouteux}@imag.fr
[3] School of Computer Science, Universiti Sains Malaysia, Gelugor, Penang, Malaysia
tienping@cs.usm.my

Abstract. This paper presents our recent study on low-resource automatic speech recognition (ASR) system with accented speech. We propose multi-accent Subspace Gaussian Mixture Models (SGMM) and accent-specific Deep Neural Networks (DNN) for improving non-native ASR performance. In the SGMM framework, we present an original language weighting strategy to merge the globally shared parameters of two models based on native and non-native speech respectively. In the DNN framework, a native deep neural net is fine-tuned to non-native speech. Over the non-native baseline, we achieved relative improvement of 15 % for multi-accent SGMM and 34 % for accent-specific DNN with speaker adaptation.

Keywords: Automatic speech recognition · Cross-lingual acoustic modelling · Non-native speech · Low-resource system · Multi-accent SGMM · Accent-specific DNN

1 Introduction

Performance of non-native automatic speech recognition (ASR) is poor when few (or no) non-native speech is available for training / adaptation. Many approaches have been suggested for handling accented-speech in ASR, such as acoustic model merging [2,16,22,23], applying maximum likelihood linear regression (MLLR) for adapting models to each non-native speaker [8], or adapting lexicon [1,4].

Lately, Subspace Gaussian Mixture Models (SGMMs) [17,18] have shown to be very promising for ASR in limited training conditions (see [11,13]). In SGMM modelling, the acoustic units are all derived from a common GMM called the Universal Background Model (UBM). This UBM, which in some way represents the acoustic space of the training data, can be estimated on large amount of

A.-H. Dediu et al. (Eds.): SLSP 2015, LNAI 9449, pp. 255–266, 2015.
DOI: 10.1007/978-3-319-25789-1_24