Available online at www.sciencedirect.com

ScienceDirect

Computer Speech and Language 41 (2017) 180–194

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

Acoustic adaptation to dynamic background conditions with asynchronous transformations

Oscar Saz *, Thomas Hain

Speech and Hearing Group, University of Sheffield, 211 Portobello St., Sheffield S1 4DP, UK

Received 11 January 2016; received in revised form 23 May 2016; accepted 25 June 2016
Available online 4 July 2016

Abstract

This paper proposes a framework for performing adaptation to complex and non-stationary background conditions in Automatic Speech Recognition (ASR) by means of asynchronous Constrained Maximum Likelihood Linear Regression (aCMLLR) transforms and asynchronous Noise Adaptive Training (aNAT). The proposed method aims to apply the feature transform that best compensates the background for every input frame. The implementation is done with a new Hidden Markov Model (HMM) topology that expands the usual left-to-right HMM into parallel branches adapted to different background conditions and permits transitions among them. Using this, the proposed adaptation does not require ground truth or previous knowledge about the background in each frame as it aims to maximise the overall log-likelihood of the decoded utterance. The proposed aCMLLR transforms can be further improved by retraining models in an aNAT fashion and by using speaker-based MLLR transforms in cascade for an efficient modelling of background effects and speaker. An initial evaluation in a modified version of the WSJCAM0 corpus incorporating 7 different background conditions provides a benchmark in which to evaluate the use of aCMLLR transforms. A relative reduction of 40.5% in Word Error Rate (WER) was achieved by the combined use of aCMLLR and MLLR in cascade. Finally, this selection of techniques was applied in the transcription of multi-genre media broadcasts, where the use of aNAT training, aCMLLR transforms and MLLR transforms provided a relative improvement of 2–3%.

Keywords: Speech recognition; Acoustic adaptation; Factorisation; Dynamic background; Media transcription

1. Introduction

Complex and dynamic acoustic backgrounds usually cause significant loss of performance on Large Vocabulary Continuous Speech Recognition (LVCSR) systems in many scenarios. Research has focused mostly on situations where the background is stationary or, at least, synchronous with the speech, following the assumption that the characteristics of the background noise remain unchanged through each utterance to decode. Multiple techniques, designed for ASR systems based on Hidden Markov Models (HMMs) and Gaussian Mixture Models (GMMs), have been

* Corresponding author at: Speech and Hearing Group, University of Sheffield, 211 Portobello St., Sheffield S1 4DP, UK. Fax: +44 (0) 114 222 1810.
E-mail addresses: o.saztorralba@sheffield.ac.uk (O. Saz).

http://dx.doi.org/10.1016/j.csl.2016.06.008

reported to provide solid improvements in ASR tasks (Li et al., 2014). These techniques can be categorised depending on whether they operate in the acoustic space, the feature space or in the model space.

Acoustic-based techniques aim to remove the background noise in the audio via some speech enhancement techniques like Wiener filtering or Minimum Mean-Square Error (MMSE) (Ephraim and Malah, 1984, 1985). Several works have reported significant improvement in recognition rates on benchmark tasks using such techniques (Astudillo et al., 2009; Paliwal et al., 2010). In a similar approach are techniques based on missing features that aim to reconstruct the clean speech signal from the input noisy signal, also with successful results (Cooke et al., 2001). More recently, techniques based on exemplars and Non-negative Matrix Factorisation (NMF) have provided also substantial gains in several tasks (Raj et al., 2010; Schuller et al., 2010).

Techniques in the feature space aim to enhance or transform the input features in order to reduce the mismatch with the GMM–HMM model used for decoding. These include Stereo-based Piecewise Linear Compensation for Environment (SPLICE) (Droppo et al., 2001) or Multi-Environment Model-based Linear Normalization (MEMLIN) (Buera et al., 2007), which have been successfully employed in conventional benchmarks for robust ASR. Another well-known technique is Constrained Maximum Likelihood Linear Regression (CMLLR) (Gales, 1998), which has been widely used to reduce variability caused by multiple sources, like speaker or background.

Model space techniques aim to re-estimate and adapt the parameters of the GMM–HMM model used for recognition. Methods like Parallel Model Combination (PMC) (Gales and Young, 1996) or Vector Taylor Series (VTS) (Moreno et al., 1996) have been especially targeted to speech recognition in noisy environments, while adaptation techniques like Maximum a Posteriori (Gauvain and Lee, 1994) and Maximum Likelihood Linear Regression (MLLR) (Gales and Woodland, 1996) have been used for adaptation to different speakers or different background conditions. Other types of model-based methods are adaptive training regimes, where the parameters of the GMM–HMM are re-estimated jointly with the parameters of some of the previously mentioned techniques. In Speaker Adaptive Training (SAT), for instance, MLLR transforms trained from a set of target speakers are used to update the model parameters (Anastasakos et al., 1996). Extending this, other types of adaptive training regimes have been used for adaptation to the background effects (Kalinli et al., 2010; Liao and Gales, 2007).

The assumption of stationarity and synchrony of the background noise is true for corpora such as NOISEX (Varga and Steeneken, 1993) or Aurora (Hirsch and Pearce, 2000), traditional benchmarks for noise adaptation and compensation techniques. These corpora were generated by adding noise to clean speech signals. This process guaranteed that a single type of noise was added to each utterance. However, in naturally occurring audio, the assumption of stationarity is often not valid. Non-stationary background effects, such as music or overlapping speech, can be common in many tasks, including the multimedia domain and meeting recognition. Furthermore, acoustic background conditions can, by nature, be independent, and hence asynchronous to the target speaker. Typical examples of asynchronous acoustic events can be applause, laughter or door slamming. The common feature of both non-stationary and asynchronous events is that their acoustic properties are not tied to the beginning and end of a speaker utterance; hence, modelling them as a single static environment does not have to be optimal.

The work in this paper aims to deal with asynchrony and non-stationarity of the background in ASR tasks, performing a thorough evaluation of the benefits that an explicit modelling of non-stationary backgrounds can provide. The initial technique will be asynchronous CMLLR (aCMLLR) transforms, which will provide adaptation to dynamic acoustic backgrounds in the feature space. This work will be then expanded with two further techniques: an asynchronous Noise Adaptive Training (aNAT) regime will be defined to provide asynchronous adaptation in the model space; and factorisation using cascading aCMLLR and MLLR transforms following work in Seltzer and Acero (2011, 2012). The proposed techniques will be evaluated in state-of-the-art GMM–HMM systems, with acoustic front-ends like Perceptual Linear Predictive (PLP) features (Hermansky, 1990), and Deep Neural Network (DNN)-front-ends like bottlenecks features (Grezl and Fousek, 2008; Liu et al., 2014). This paper expands and describes a common framework for the techniques briefly introduced in Saz and Hain (2013) and Saz et al. (2015).

This paper is organised as follows: Section 2 will introduce a novel technique to perform asynchronous background adaptation with feature transforms. Section 3 will describe the two extensions to this method for adaptive training and factorisation. Section 4 will evaluate the proposed techniques with asynchronous transforms in a controlled scenario with WSJCAM0, and Section 5 will provide the results on the automatic transcription of Multi-Genre Broadcasts (MGB). Finally, Section 6 will present the conclusions to this work.