Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 68 (2015) 55–68

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Ensemble environment modeling using affine transform group

Yu Tsao [a,*], Payton Lin [a], Ting-yao Hu [a], Xugang Lu [b]

[a] Research Center for Information Technology Innovation, Academia Sinica, Taiwan
[b] Spoken Language Communication Laboratory, National Institute of Information and Communications Technology, Kyoto, Japan

Received 10 July 2014; received in revised form 22 November 2014; accepted 24 December 2014
Available online 8 January 2015

Abstract

The ensemble speaker and speaking environment modeling (ESSEM) framework was designed to provide online optimization for enhancing workable systems under real-world conditions. In the ESSEM framework, ensemble models are built in the offline phase to characterize specific environments based on local statistics prepared from those particular conditions. In the online phase, a mapping function is computed based on the incoming testing data to perform model adaptation. Previous studies utilized linear combination (LC) and linear combination with a correction bias (LCB) as simple mapping functions that only apply one weighting coefficient on each model. In order to better utilize the ensemble models, this study presents a generalized affine transform group (ATG) mapping function for the ESSEM framework. Although ATG characterizes unknown testing conditions more precisely using a larger amount of parameters, over-fitting issues occur when the available adaptation data is especially limited. This study handles over-fitting issues with three optimization processes: maximum a posteriori (MAP) criterion, model selection (MS), and cohort selection (CS). Experimental results showed that ATG along with the three optimization processes enabled the ESSEM framework to allow unsupervised model adaptation using only one utterance to provide consistent performance improvements.

Keywords: Ensemble modeling; Environment modeling; Prior knowledge; Maximum a posteriori; Model selection; Cohort selection

1. Introduction

Towards ubiquitous adoption of human–machine communication (Deng and Huang, 2004), robustness in automatic speech recognition (ASR) (Junqua et al., 1996) has been addressed by noise-robust techniques (Li et al., 2014), data reduction (O'Shaughnessy, 2008), and predictive classification (Huo and Lee, 2000). To address the technical challenges of performing according to the user's intention, selection, execution, and evaluation (Norman, 1984), environment modeling or model adaptation methods (Lee, 1998; Sankar and Lee, 1996,) extend workable systems to real-world situations by modeling specific speakers and acoustic environments with unlabeled and limited amounts of adaptation data. Fig. 1 presents the structure of environment modeling, where either one general model (Category-1) or multiple environment specific models (Category-2) are first prepared as a structure using the entire training data set. In the online phase, speech segments from incoming testing conditions are collected to derive a mapping function, $F_\varphi(\cdot)$, that performs model adaptation to obtain a target model, Λ^Y, minimizing the differences between training and testing conditions. Parameters in the mapping function can be estimated via criterion such as maximum likelihood (ML) and maximum a posteriori (MAP).

For Category-1, a single source model (Λ^X in Fig. 1) is built to reflect the average statistics of the whole training data set. The mapping function is then estimated to adapt the source model to the target model. Several estimation

* Corresponding author at: No 128, Academia Road, Section 2, Nankang, Taipei 11529, Taiwan. Tel.: +886 2 2787 2390; fax: +886 2 2787 2315.
E-mail address: yu.tsao@citi.sinica.edu.tw (Y. Tsao).

http://dx.doi.org/10.1016/j.specom.2014.12.007
0167-6393/

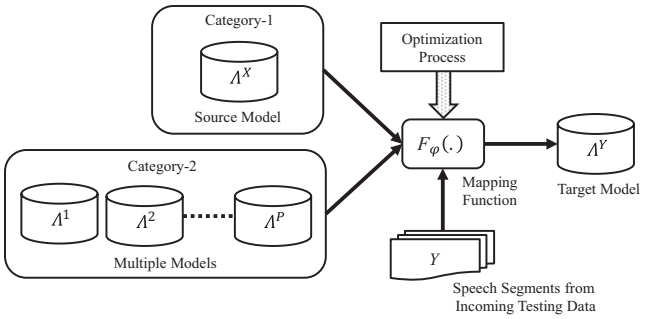

Fig. 1. Structure of environment modeling and model adaptation.

algorithms have been proposed such as linear and nonlinear stochastic matching approaches (Lee, 1998; Sankar and Lee, 1996; Suredran et al., 1999), signal bias removal (SBR) (Rahim and Juang, 1996), maximum likelihood linear regression (MLLR) (Leggetter and Woodland, 1995; Gales, 1997), maximum a posteriori linear regression (MAPLR) (Chesta et al., 1999; Siohan et al., 2001; Siohan et al., 2002), structural Bayesian linear regression (SBLR) (Watanabe et al., 2014), VTS-based model adaptation (Kim et al., 1998), joint compensation of additive and convolutive distortions (JAC) (Gong, 2005; Hu and Huo, 2007; Li et al., 2009), and JAC with unscented transform (JAC-UT) (Hu and Huo, 2006; Li et al., 2010).

For Category-2, multiple models ($\{\Lambda^1, \Lambda^2, \ldots, \Lambda^P\}$ in Fig. 1) that are trained using subsets of the entire training data allow more effective local statistics of environment conditions. In these cases, the mapping function for adaptation needs to transform multiple models to the target model. For efficient estimation of mapping functions, several techniques have been proposed such as reference speaker weighting (RSW) (Hazen, 2000), eigenvoice (Kuhn et al., 2000), cluster adaptive training (CAT) (Gales, 2000; Yu and Gales, 2006), speaker clustering (Kosaka et al., 1996; Padmanabhan et al., 1998), probabilistic 2DPCA/GLRAM (Jeong, 2012), tensor voices (Jeong, 2014), and ensemble speaker and speaking environment modeling (ESSEM) (Tsao and Lee, 2009). Generally, a simple mapping function such as best first (BF) (Tsao et al., 2012), linear combination (LC) (Kuhn et al., 2000; Gales, 2000), or linear combination with correction bias (LCB) (Tsao et al., 2014) is used to perform adaptation. However, a mapping function that utilized more free parameters could enable more accurate model estimation when larger amounts of adaptation data become available. Therefore, the present study proposes an affine transform group (ATG) mapping function that applies an affine transform for each model in $\{\Lambda^1, \Lambda^2, \ldots, \Lambda^P\}$ to compute the target model. The ATG mapping function expands upon the previous ESSEM framework (Tsao et al., 2014; Tsao et al., 2012) and is denoted as ATG-ESSEM in the following discussion. While the usage of more free parameters can provide better environment modeling capabilities, over-fitting issues must be considered when the amount of adaptation data is insufficient. A previous study proposed to adopt the MAP criterion to handle over-fitting (Tsao et al., 2012). The present study proposes two additional approaches to enhance optimization processes: model selection (MS) and cohort selection (CS). This study also compares four different types of affine transform matrix: full, diagonal, scalar, and identity matrices, in order to evaluate the benefits of added complexity.

To verify effective model adaptation using ATG-ESSEM with the MAP criterion, MS, and CS, experiments were conducted on Aurora-4, a large vocabulary continuous speech recognition (LVCSR) task (Parihar and Picone, 2002; Parihar et al., 2004; Hirsch, 2001; Au Yeung and Siu, 2004). Unsupervised ESSEM adaptation could also enhance the parameter estimation of deep neural networks (DNNs) (Seltzer et al., 2013). Some adaptation methods have been proposed in DNN–HMM systems by using linear transformations (Neto et al., 1995; Li and Sim, 2010; Yao et al., 2012; Gemello et al., 2007; Ochiai et al., 2014). Due to the enormous amount of parameters, DNN has limited adaptation capability when only limited adaptation data is available. Since DNN parameter estimation is based on discriminative criterion, adaptation performance is sensitive to label errors. A combination of GMM and DNN has also effectively enhanced ASR performance (Liu and Sim, 2014) since the GMM–HMM framework is based on generative training paradigms for more robust unsupervised adaptation. This study evaluates the ATG-ESSEM framework using a difficult task designed to simulate "real-world" conditions: per-utterance unsupervised adaptation with lots of fluctuating SNRs. Experimental results confirmed the effective adaptation capability of ATG-ESSEM with only one adaptation utterance. Discussion related to adaptation under future DNN–HMM systems will be included following results of our GMM–HMM findings.

The rest of this paper is organized as follows: Section 2 reviews the ESSEM framework and the ATG mapping function, Section 3 derives three optimization processes to enhance ATG-ESSEM performance, Section 4 reports results and discusses extensions of ESSEM to DNN model parameter adaptation, and Section 5 offers concluding remarks.

2. Ensemble environment modeling and affine transform group (ATG)

In this section, the ESSEM framework is first described, followed by the presentation of the proposed ATG mapping function.

2.1. Ensemble speaker and speaking environment modeling (ESSEM)

Fig. 2 illustrates the ESSEM framework, which consists of offline and online phases. In the offline phase, a single source model Λ^X is first estimated based on the entire set of training data. This source model is trained on speech data collected from a variety of environment conditions