Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 58 (2014) 1–10

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Joint speaker and environment adaptation using TensorVoice for robust speech recognition

Yongwon Jeong *

School of Electrical Engineering, Pusan National University, Busan 609-735, Republic of Korea

Received 12 October 2012; received in revised form 5 October 2013; accepted 14 October 2013
Available online 25 October 2013

Abstract

We present an adaptation of a hidden Markov model (HMM)-based automatic speech recognition system to the target speaker and noise environment. Given HMMs built from various speakers and noise conditions, we build tensorvoices that capture the interaction between the speaker and noise by using a tensor decomposition. We express the updated model for the target speaker and noise environment as a product of the tensorvoices and two weight vectors, one each for the speaker and noise. An iterative algorithm is presented to determine the weight vectors in the maximum likelihood (ML) framework. With the use of separate weight vectors, the tensorvoice approach can adapt to the target speaker and noise environment differentially, whereas the eigenvoice approach, which is based on a matrix decomposition technique, cannot differentially adapt to those two factors. In supervised adaptation tests using the AURORA4 corpus, the relative improvement of performance obtained by the tensorvoice method over the eigenvoice method is approximately 10% on average for adaptation data of 6–24 s in length, and the relative improvement of performance obtained by the tensorvoice method over the maximum likelihood linear regression (MLLR) method is approximately 5.4% on average for adaptation data of 6–18 s in length. Therefore, the tensorvoice approach is an efficient method for speaker and noise adaptation.

Keywords: Acoustic model adaptation; Environment adaptation; Speaker adaptation; Speech recognition; Tensor analysis

1. Introduction

In hidden Markov model (HMM)-based automatic speech recognition (ASR) (Rabiner, 1989; O'Shaughnessy, 2008), speaker and environment variabilities are two major factors that affect the performance of ASR systems in real-world applications. Various techniques have been investigated to compensate for these two factors.

Linear models have been successfully used in the adaptation of acoustic models, e.g., eigenvoice adaptation (Kuhn et al., 2000). In the eigenvoice approach, basis vectors are constructed from the supervectors of training speakers by principal component analysis (PCA) (Jolliffe, 2002). A supervector for each training speaker is built by concatenating all the mean parameters of Gaussian mixture components. During adaptation, the model for the target speaker is assumed to be a linear combination of basis vectors and the weight vector is estimated in the maximum likelihood (ML) framework. Due to its low-dimensional speaker space, the eigenvoice method is suitable for rapid speaker adaptation. The eigenvoice approach can also be applied to speaker and environment adaptation by constructing a speaker and environment space using training models built from many speakers and various noise conditions, as described below. The eigenvoice approach is the application of *Eigenfaces* (Sirovich and Kirby, 1987; Turk and Pentland, 1991) to the adaptation of acoustic models, although the eigenvectors termed the *eigenvoices* obtained by the eigenvoice approach do not represent a voice signal.

Our approach is based on a tensor decomposition (or a multilinear decomposition) (Kolda and Bader, 2009) for

* Tel.: +82 51 510 1704.
E-mail address: jeongy@pusan.ac.kr

0167-6393/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.specom.2013.10.001

speaker and noise adaptation. We briefly review works on speaker and noise adaptation techniques in HMM-based speech recognition, and applications of tensor decompositions.

For robust speech recognition, speaker and environment variabilities can be compensated by using either different techniques or the same technique for both. First, speaker and environment variabilities can be separately compensated by combining a speaker adaptation technique and a noise compensation technique. In Nguyen et al. (1999), environment adaptation is performed in a speaker-independent (SI) manner using maximum likelihood linear regression (MLLR) adaptation (Leggetter and Woodland, 1995), thus compensating for the environmental effect, and speaker adaptation is performed in the eigenvoice framework. In Rigazio et al. (2001), the difference of mean vectors between training and target noise conditions is compensated by a Jacobian approach using a first-order approximation (Gales, 1998a; Sagayama et al., 1997), and the differences between the training and target speakers are compensated by the MLLR + maximum *a posteriori* (MAP) method (Gauvain and Lee, 1994). In Wang and Gales (2012), the training and target speaker differences are compensated by transforming HMM mean parameters by MLLR adaptation, and background noise is compensated by a model-based vector Taylor series (VTS) transform (Acero et al., 2000; Li et al., 2007); the transforms for the two factors are estimated using different adaptation data.

Second, speaker and environment variabilities can be compensated using the same adaptation technique. In Seltzer and Acero (2011a), a series of constrained MLLR (CMLLR) transforms (Gales, 1998b) are estimated to transform a speaker-adapted model in one environment to a different environment. In the training phase, separate transforms for speaker and environment are obtained by speaker-adaptive training (SAT) (Anastasakos et al., 1996). In SAT, a transformation-based adaptation technique such as the MLLR method is applied during both training and adaptation such that inter- and intra-speaker variations are decoupled and the model for intra-speaker variation is transformed to the target speaker. However, this approach needs knowledge about the training and testing conditions. This problem is addressed by using an unsupervised environment clustering technique described in Seltzer and Acero (2011b). Environment clustering is performed by building a Gaussian mixture model (GMM) using the silence portion of training utterances, and the silence portion from utterances of a test environment is used to decide the cluster that the test environment belongs to. In ensemble speaker and speaking environment modeling (ESSEM) (Tsao and Lee, 2009), a joint speaker and noise adaptation approach is proposed using the eigenvoice approach. In ESSEM, HMMs are constructed for different speakers, noise types, and channel distortions. A supervector for each training condition is built in the same way as in the eigenvoice approach. Supervectors built in different training conditions form an ensemble speaker and speaking environment space. The supervector for a new testing environment is estimated by a stochastic matching criterion via either the environment clustering algorithm (where supervectors are clustered into groups) or the environment partitioning algorithm (where each supervector is partitioned into smaller vectors). ESSEM is closely related to the eigenvoice technique and cluster adaptive training (CAT) based speaker adaptation (Gales, 2000) in that supervectors of the training condition play the role of basis vectors. In CAT, cluster means, which play the role of basis vectors, are derived from the cluster analysis of training models, and the model for the target speaker is expressed as a linear interpolation of all the cluster means. The interpolation weight vector (which is equivalent to a weight vector in the eigenvoice technique) is estimated in the ML framework.

Tensor decompositions (Kolda and Bader, 2009) have been successfully used in image and computer vision applications. Tensor decompositions are higher-order generalizations of matrix decompositions such as PCA and singular value decomposition (SVD). Extending the eigenface method, Vasilescu and Terzopoulos (2002a) uses a tensor decomposition to model an ensemble of face images of various expressions, viewpoints, and illumination conditions using tensor analysis, introducing the *TensorFaces*. The tensorface approach is applied to facial recognition under varying expression, viewpoint, and illumination in Vasilescu and Terzopoulos (2002b). In Vlasic et al. (2005), the authors apply a tensor decomposition to face animation. Using the tensor decomposition of 3D scans of faces from many people and expressions, the authors build a model that can control the identity, expression, and viseme (the visual equivalent of phoneme). As a generalization of SVD to tensor objects, multilinear SVD is introduced in Lathauwer et al. (2000). Multilinear PCA (MPCA), the tensor equivalent of PCA, is applied to gait recognition in Lu et al. (2008). Vasilescu and Terzopoulos (2005) introduces a generalization of independent component analysis (ICA) to tensor objects called multilinear ICA (MICA), which provides better performance than the tensorface approach in facial recognition under varying viewpoint and illumination.

In this paper, we present a tensor decomposition based approach to adapt acoustic models jointly to the target speaker and noise environment. From the decomposition of a fourth-order tensor (4-D array) consisting of acoustic models trained from many speakers and various noise conditions, we obtain an extended core tensor, which we call the *TensorVoices*, that is common across training speakers and noise conditions. In Jeong (2010), a tensor decomposition is applied to the clean acoustic models of training speakers where the HMM mean parameters of training speakers are collectively represented in a third-order tensor. In Jeong (2011), the approach is extended to a fourth-order tensor in which an additional dimension is added for a noise space. However, in their approach,