Computers and Electrical Engineering 000 (2017) 1–14

Contents lists available at ScienceDirect

Computers and Electrical Engineering

journal homepage: www.elsevier.com/locate/compeleceng

Feature joint-state posterior estimation in factorial speech processing models using deep neural networks☆

Mahdi Khademian, Mohammad Mehdi Homayounpour*

Laboratory for Intelligent Multimedia Processing (LIMP), Amirkabir University of Technology, Tehran, Islamic Republic of Iran

ARTICLE INFO

Article history:
Received 17 November 2016
Revised 21 June 2017
Accepted 23 June 2017
Available online xxx

Keywords:
Factorial speech processing models
Deep neural networks
factorial hidden Markov models
State-conditional observation distribution
Model combination using vector Taylor series
Feature joint-state posterior

ABSTRACT

This paper proposes a new method for calculating joint-state posteriors of mixed-audio features using deep neural networks to be used in factorial speech processing models. The joint-state posterior information is required in factorial models to perform joint-decoding. The novelty of this work is its architecture which enables the network to infer joint-state posteriors from the pairs of state posteriors of stereo features. This paper defines an objective function to solve an underdetermined system of equations, which is used by the network for extracting joint-state posteriors. It develops the required expressions for fine-tuning the network in a unified way. The experiments compare the proposed network decoding results to those of the vector Taylor series method and show 2.3% absolute performance improvement in the monaural speech separation and recognition challenge. This achievement is substantial when we consider the simplicity of joint-state posterior extraction provided by deep neural networks.

1. Introduction

Factorial speech processing models are powerful generative tools for modeling dynamics of two or more audio sources and the way they combine to generate mixed-audio signals. Fig. 1 shows the graphical model of factorial speech processing models in which two audio sources create a mixed-audio signal. In fact, these models are constructed based on factorial hidden Markov models (FHMM) [1], which consist of multiple Markov chains for modeling dynamic systems with multiple underlying independent stochastic processes.

Factorial speech processing models have been applied in the past for two main purposes: robust automatic speech recognition (ASR) in environments with non-stationary noises [2] and monaural multi-talker speech separation and/or recognition [3]. In the first scenario, two stochastic processes are speech process and non-stationary noise process, which are combine to produce the corrupted speech signal. For this case, factorial speech processing models are suggested for handling non-stationary noise situations [4] in which they provide improvement over their single noise-state robust-ASR techniques like model compensation [2,5]. In the second scenario, two or more speaker voices are mixed together to produce a multi-talker speech utterance in which underlying source processes are the speakers' voices. In this case, previous achievements are surprising [6,7], even better than the results achieved manually by human listening (Fig. 8-left). The main reason for this is the ability of the factorial models for generative modeling of the mixing procedure in which the models are best fitted to

☆ Reviews processed and recommended for publication to the Editor-in-Chief by Associate Editor Dr. Z. Arnavut.

* Corresponding author.
E-mail address: homayoun@aut.ac.ir (M.M. Homayounpour).

http://dx.doi.org/10.1016/j.compeleceng.2017.06.028
0045-7906/

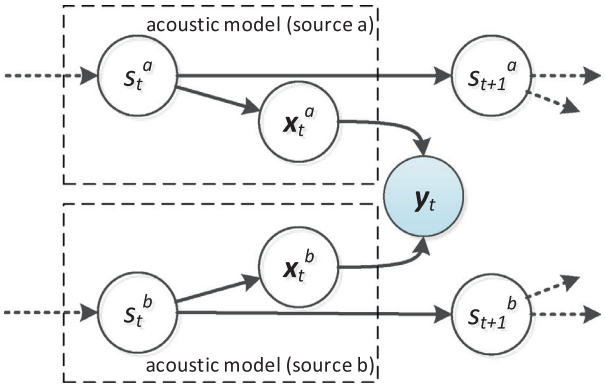

Fig. 1. Graphical model of factorial speech processing models. This figure shows all variables of frame t and only shows state variables of frame $t+1$. It shows two audio sources, which are related by $p(\boldsymbol{y}_t|\boldsymbol{x}_t^a, \boldsymbol{x}_t^b)$ to generate the mixed-audio feature of frame t. Source acoustic models are simply independent hidden Markov models for each audio source.

this application. Information flow and exchange in source speech Markov chains is very effective for resolving ambiguities during speech decoding since only one recording channel is available in this application.

The information flow requires having joint-state likelihoods (or posteriors) for the mixed-audio features, i.e., $p(\boldsymbol{y}_t|s_t^a, s_t^b)$ or $p(s_t^a, s_t^b|\boldsymbol{y}_t)$. There exist at least two problems with providing joint-state likelihoods for the decoder. First, the calculation of joint-state likelihoods requires a potentially squared number of states for which the likelihood should be calculated. This significantly increases the required calculations during the decoding. Second, there are some approximations involved in most of the methods for calculating joint-state likelihood. For example, in the max model, the max approximation is valid for high resolution spectral features [6], while for recognition purposes, we need more elaborated feature spaces [4]. As another example, the data-driven parallel model combination (PMC) and vector Taylor series (VTS) methods rely on mismatch functions where there is at least one approximation involved in developing the mismatch functions [8,9]. The VTS method also uses a linear approximation of the nonlinear mismatch function. Considering the third problem, the data-driven PMC and the method of weighted stereo samples (WSS) [5] need an additional parametric modeling step along with saving joint-state parametric models, which require quadratic space compared to the separate hidden Markov model (HMM) acoustic models. Moreover, while the WSS method provides the most accurate results since it does not rely on the mismatch function, it suffers from a need for stereo data.

Recent efforts make the deep neural networks (DNN) a powerful acoustic model in ASR applications, which is referred to as DNN/HMM architecture [10,11]. In these works, DNNs are used to extract senone posteriors. The senone posteriors are then passed to the decoder that performs the decoding. In this architecture, hidden Markov models (HMM) are only used for decoding and not extracting the acoustic likelihoods. A recent work by Microsoft research [12] applies DNNs to a task that is closely related to factorial models: the monaural speech separation and recognition challenge [13]. In this challenge, we consider two chains in the factorial model in which each models a speaker, while both speakers simultaneously issue a simple command. In [12], two separate DNNs are used for extracting marginal posteriors that are used in a decoder. In this case, there cannot be any information flow between the two chains of the factorial model. The state posteriors must be extracted in the joint form to make the information flow possible between chains of factorial models.

This paper proposes a new method for extracting true joint-state posteriors to be used in factorial speech processing models. This way of applying DNNs along with using factorial speech processing models is comparable to applying DNNs in the DNN/HMMs [11]. Therefore, by proposing the mentioned method, this work brings the DNNs to be used in factorial models as a DNN/FHMM architecture. This architecture can be used in robust-ASR applications that require further development of the proposed method.

The rest of this paper is organized as follows. The next section describes the exact method for the calculation of joint-state likelihoods and a variety of approximated estimation and modeling techniques for this purpose. It simplifies presenting the current methods in a unified view to bring a clear insight into the problem. The third section describes the proposed method for calculating joint-state posteriors using deep neural networks. It defines the objective function, the way to optimize it, and some practical considerations for training the network. Section 4 describes the experiment framework, dataset, and achieved results, and the last section concludes the paper.

2. Methods for calculating features' joint-state likelihoods

The role of joint-state likelihoods in factorial models of speech processing is like the observation distribution of HMMs, which is required for decoding. The following expression exactly calculates the joint-state likelihood of $\boldsymbol{y}_t$ (the time index,

Please cite this article as: M. Khademian, M.M. Homayounpour, Feature joint-state posterior estimation in factorial speech processing models using deep neural networks, Computers and Electrical Engineering (2017), http://dx.doi.org/10.1016/j.compeleceng.2017.06.028