Speech Communication 98 (2018) 1–16

Contents lists available at ScienceDirect

Speech Communication

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

# Monaural multi-talker speech recognition using factorial speech processing models

Mahdi Khademian, Mohammad Mehdi Homayounpour*

*Laboratory for Intelligent Multimedia Processing (LIMP), Computer Engineering and IT Department, Amirkabir University of Technology, Tehran, Islamic Republic of Iran*

ARTICLE INFO

*Keywords:*
Factorial hidden Markov model
Vector Taylor series
Monaural mixed-speech recognition
Joint-decoding
Two-dimensional Viterbi
Joint-speaker identification

ABSTRACT

A Pascal challenge entitled monaural speech separation and recognition challenge was developed, targeting the problem of robust automatic speech recognition against speech-like noises which significantly degrade the performance of automatic speech recognition systems. In this challenge, two competing speakers say a simple command simultaneously and the objective is to recognize speech of the target speaker. Surprisingly, a team from IBM research could achieve performance better than human listeners on this task during the challenge. The IBM system consists of an intermediate speech separation and two single-talker speech recognition modules. This paper reconsiders the recognition task of this challenge based on gain adapted factorial speech processing models. It develops a joint-token passing algorithm for direct joint-decoding of target and masker speakers' mixed-signals, simultaneously. It uses maximum uncertainty during the joint-decoding, which cannot be used in the two-phased IBM system. This paper provides a detailed derivation of inference on these models based on the general inference procedures of probabilistic graphical models. Additionally, it uses deep neural networks for joint-speaker identification and their gain estimation, which makes these two steps easier than before while producing competitive results for these steps. The proposed method of this work outperforms past super-human results and even the results recently achieved using deep neural networks by Microsoft research. It achieved 5.3% absolute task performance improvement compared to the first super-human system and 2.5% absolute task performance improvement compared to its recent competitor.

## 1. Introduction

Robustness of automatic speech recognition systems (ASR) against diverse speech processing environments and adverse disturbing noises remains an important research area in speech recognition systems (Baker et al., 2009; Li et al., 2014). Among all diversities and conditions in everyday environments that ASR systems must manage, dealing with the Babble noise and presence of competing speakers is a challenging problem for these systems. This problem is known as the cocktail party problem (Haykin and Chen, 2005) in which a person (or a system) wants to focus and follow a speaker's conversation in a place where people talk simultaneously. Roughly speaking, two groups of approaches are developed to address this problem. Approaches in the first group incorporate signals captured from several microphones and perform low level signal processing techniques such as beam forming and blind source separation, which reduces the footprint of competing audio sources. These approaches accomplish their speech processing tasks using improved signals. Approaches in the second group use only one recording channel and perform high level speech processing and machine learning techniques and accomplish their tasks in the presence of a competing audio source, which seems to be more challenging.

An interesting competition called the Pascal 2006 monaural speech separation and recognition challenge addresses problems related to the second group of approaches (Cooke et al., 2010). In this challenge, two competing speakers simultaneously issue two commands. The challenge objective is to recognize the command of the target speaker. The commands of the task follow a set of simple grammar rules and consist of a small vocabulary.

The problem of monaural multi-talker speech recognition becomes more difficult when the masker speaker's speech has higher energy than that of the target. The worst case is when the masker speaker voice is like that of the target, i.e. when the two speakers have the same gender or the two speakers are the same. Several teams attended this competition with variety of techniques (Cooke et al., 2010). Among the competitors, a team from IBM research presented a technique that outperformed others, even human listeners (Hershey et al., 2010). They performed the task in three main steps. First, they estimated the speakers' identity and their gains using a set of high resolution Gaussian

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

https://doi.org/10.1016/j.specom.2018.01.007
Received 17 August 2016; Received in revised form 10 September 2017; Accepted 30 January 2018
Available online 02 February 2018

mixture models (GMM) as the speaker models. In the main step, they jointly separated both speakers' speech from the mixed-speech signal using factorial speech processing models. This step considers the expected value of the source features given the observed feature and the inferred joint acoustic states for source estimation, then the source separation is performed. In the third step, the two separated speeches are decoded using a single-talker recognition system (Hershey et al., 2010). The team further developed their system to support a mixture of speeches of more than two speakers and separate their voices only by one recording channel. Researchers continued to work on this dataset. To the best of our knowledge, only the method proposed by a team from Microsoft research outperforms IBM's super-human system. The Microsoft system incorporates a pair of deep neural networks (DNN) for acoustic inference over semi-joint hidden Markov models (HMM) (Weng et al., 2015), a network for generating senone posteriors of high energy utterances, and another network for low energy utterances. They then perform decoding over these posteriors to complete the task. However, Microsoft's proposed DNN-based system of cannot take advantage of the two HMM dynamics information since it uses two disjoint DNNs for evaluating joint-state posteriors.

The method presented in this paper is a model-based approach based on factorial speech processing models for recognizing monaural mixed-speech signals, which is applied for the Pascal challenge. It directly performs joint-decoding over the model to simultaneously decode utterances of both the target and masker speakers. Direct speech recognition of this work is accomplished by a joint-decoder, which is developed by extending the token passing algorithm to support inference over factorial speech processing models composed of sub-word acoustic units. Moreover, this paper derives inference expressions of the factorial models using the general inference procedures of probabilistic graphical models. Based on the derived expressions, it develops the joint-token passing algorithm for performing joint-decoding. Additionally, this work uses deep neural networks for speaker identification and gain estimation as an important step for determining and adapting the factorial model's audio sources.

The rest of this paper is organized as follows: the next section briefly describes the challenge and presents a detailed description of steps for applying factorial speech processing models to this challenge. In this section the joint-token passing algorithm is developed and presented. Section 3 describes the methods for determining and adapting the source models. Section 4 presents experiments, scoring procedure, results, and discussion about the super-human systems, and Section 5 concludes the paper.

## 2. Factorial speech processing models for single channel speech recognition

The goal of the Pascal monaural speech separation and recognition challenge is to recognize keywords of a target speaker from a mixed-speech of the target and a masker speaker (Cooke et al., 2010). Mixed-speech signals are artificially created in this task from speech materials of the Grid corpus (Cooke et al., 2006). This corpus contains simple six-word slot commands from a set of 34 speakers. Each command is a sequence of a command word, a color, a preposition, a letter, a digit, and an adverb, as depicted in Box 1.

Mixed-speech signals are created by selecting two utterances from the Grid corpus, one as a target speaker and the other as a masker. The target speaker always uses "white" as the command color but the masker does not. This is the clue for discriminating the target and the masker speakers. The following time domain relation mixes two speech signals:

$$y = x^a + gx^b \tag{1}$$

in which $x^a$ is the speech signal of the target and $x^b$ is for the masker speaker. The challenge is designed for different signal energy ratios of the target and the masker which is called Target to Masker Ratio (TMR).

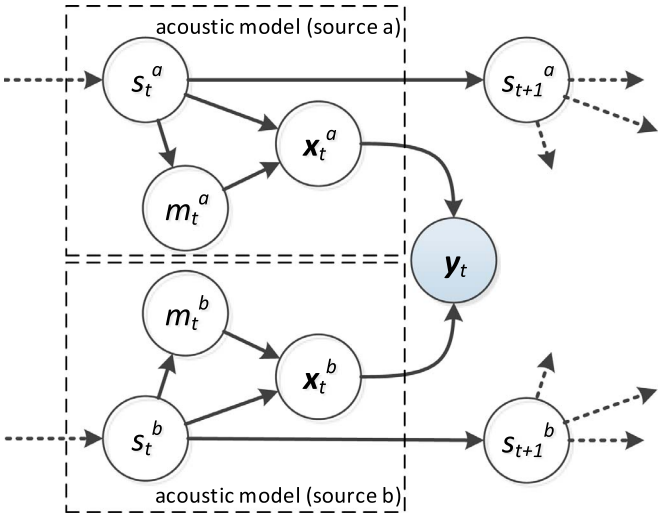


**Fig. 1.** Factorial speech processing model for recognition of two audio sources only by one recording channel. In this graphical model, $\boldsymbol{y_t}$ is the feature vector of captured signal at time frame $\boldsymbol{t}$ and is observable (other variables are hidden), $\boldsymbol{s_t^a}$ and $\boldsymbol{s_t^b}$ are hidden states of the two audio sources, $\boldsymbol{m_t^a}$ and $\boldsymbol{m_t^b}$ are mixture components of the audio sources when the Gaussian mixture model is used for acoustic modeling, and $\boldsymbol{x_t^a}$ and $\boldsymbol{x_t^b}$ are hidden feature vectors of the audio sources.

The gain coefficient, $g$, in (1) adjusts this.

The next sub-section provides an overview of factorial speech processing models that are applied for the Pascal challenge in this paper. The models' inference and decoding procedures are then explored in detail.

### 2.1. Factorial speech processing models

Factorial speech processing models are generative models for modeling the combination of multiple audio sources into one (or multiple) observable mixed-audio signals. These models are applicable for robust-ASR systems (Hershey et al., 2012) and are well suited for the Pascal challenge. The factorial speech processing models are developed based on factorial hidden Markov models (FHMM) (Ghahramani and Jordan, 1997), which are used for modeling processes with multiple independent underlying Markov chains (Koller and Friedman, 2009). The FHMMs were applied for a collection of Bach's chorales in (Ghahramani and Jordan, 1997) where these models could successfully capture statistical structure of these melody lines.

Fig. 1 shows the graphical model of a factorial speech processing model. In this figure, the two source Markov processes are the speech processes of speaker $a$ and $b$. Conditional probability distribution (CPD) of the Markov chains of the two audio sources are $p(s_{t+1}^a|s_t^a)$ and $p(s_{t+1}^b|s_t^b)$, which are modeled parametrically by individual stochastic matrices. Each factorial speech processing model chain contains an HMM for modeling its audio source, known as acoustic model (Young, 1996) in conventional speech recognition applications. Fig. 1 shows these source [acoustic] models by the dashed boxes around each audio source (sources $a$ and $b$).

Conventionally, in speech processing applications, Gaussian mixture models are used for HMM observation distribution, and left-to-right topology is used for modeling the Markov process. The observation probability distribution of HMMs is modeled by the following CPD:

$$p(\boldsymbol{x}_t^a|s_t^a) = \sum_{m_t^a} p(\boldsymbol{x}_t^a|m_t^a, s_t^a)p(m_t^a|s_t^a) \tag{2}$$

where $p(\boldsymbol{x}_t^a|s_t^a = i, m_t^a = j) = (\boldsymbol{x}_t^a; \boldsymbol{\mu}_{(i,j)}, \Sigma_{(i,j)})$ is the $j$th Gaussian component of the GMM of state $i$, and $p(m_t^a|s_t^a)$ models the component weights by a stochastic matrix. The observation model of the second chain is like the first one.