Available online at www.sciencedirect.com

ScienceDirect

Computer Speech and Language 24 (2010) 77–93

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

A computational auditory scene analysis system for speech segregation and robust speech recognition

Yang Shao [a,*], Soundararajan Srinivasan [b,1], Zhaozhang Jin [a], DeLiang Wang [a,c]

[a] *Department of Computer Science and Engineering, The Ohio State University, Columbus, OH 43210, USA*
[b] *Biomedical Engineering Department, The Ohio State University, Columbus, OH 43210, USA*
[c] *Center for Cognitive Science, The Ohio State University, Columbus, OH 43210, USA*

Received 5 July 2007; received in revised form 17 January 2008; accepted 3 March 2008
Available online 28 March 2008

Abstract

A conventional automatic speech recognizer does not perform well in the presence of multiple sound sources, while human listeners are able to segregate and recognize a signal of interest through auditory scene analysis. We present a computational auditory scene analysis system for separating and recognizing target speech in the presence of competing speech or noise. We estimate, in two stages, the ideal binary time–frequency (T–F) mask which retains the mixture in a local T–F unit if and only if the target is stronger than the interference within the unit. In the first stage, we use harmonicity to segregate the voiced portions of individual sources in each time frame based on multipitch tracking. Additionally, unvoiced portions are segmented based on an onset/offset analysis. In the second stage, speaker characteristics are used to group the T–F units across time frames. The resulting masks are used in an uncertainty decoding framework for automatic speech recognition. We evaluate our system on a speech separation challenge and show that our system yields substantial improvement over the baseline performance.

Keywords: Speech segregation; Computational Auditory Scene Analysis; Binary time–frequency mask; Robust speech recognition; Uncertainty decoding

1. Introduction

In everyday listening conditions, the acoustic input reaching our ears is often a mixture of multiple concurrent sound sources. While human listeners are able to segregate and recognize a target signal under such conditions, robust automatic speech recognition remains a challenging problem (Huang et al., 2001). Automatic speech recognition (ASR) systems are typically trained on clean speech and face the mismatch problem when

[*] Corresponding author.
E-mail addresses: shaoy@cse.ohio-state.edu (Y. Shao), srinivasan.36@osu.edu (S. Srinivasan), jinzh@cse.ohio-state.edu (Z. Jin), dwang@cse.ohio-state.edu (D. Wang).
[1] Present address: Research and Technology Center, Robert Bosch LLC, Pittsburgh, PA 15212, USA.

0885-2308/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.csl.2008.03.004

tested in the presence of interference. In this paper, we address the problem of recognizing speech from a target speaker in the presence of either another speech source or noise.

To mitigate the effect of interference on recognition, speech mixtures can be preprocessed by speech separation algorithms. Under monaural conditions, systems typically depend on modeling the various sources in the mixture to achieve separation (Ephraim, 1992; Jang and Lee, 2003; Kristjansson et al., 2004; Roweis, 2005; Raj et al., 2005). An alternate approach to employing speech separation prior to recognition involves the joint decoding of the speech mixture based on knowledge of all the sources present in the mixture (Varga and Moore, 1990; Gales and Young, 1996; Deoras and Hasegawa-Johnson, 2004). These model-based systems rely heavily on the use of *a priori* information of sound sources. Such approaches are fundamentally limited in their ability to handle novel interference (Allen, 2005). For example, systems that assume and model the presence of multiple speech sources only, do not lend themselves easily to handling speech in (non-speech) noise conditions.

In contrast to the above model-based systems, we present a primarily feature-based computational auditory scene analysis (CASA) system that makes weak assumptions about the various sound sources in the mixture. It is believed that the human ability to function well in everyday acoustic environments is due to a process termed auditory scene analysis (ASA), which produces a perceptual representation of different sources in an acoustic mixture (Bregman, 1990). In other words, listeners organize the mixture into *streams* that correspond to different sound sources in the mixture. According to Bregman (1990), organization in ASA takes place in two main steps: segmentation and grouping. Segmentation decomposes the auditory scene into groups of contiguous time–frequency (T–F) units or segments, each of which mainly originates from a single sound source (Wang and Brown, 2006). A T–F unit denotes the signal at a particular time and frequency. Grouping involves combining the segments that are likely to arise from the same source together into a single stream (Bregman, 1990). Grouping itself is comprised of simultaneous and sequential organizations. Simultaneous organization involves grouping of segments across frequency, and sequential organization refers to grouping across time.

From an information processing perspective, the notion of an ideal binary T–F mask has been proposed as a major computational goal of CASA by Wang (2005). Such a mask can be constructed from the *a priori* knowledge of target and interference; specifically a value of 1 in the mask indicates that the target is stronger than the interference within the corresponding T–F unit and 0 indicates otherwise. The use of ideal binary masks is motivated by the auditory masking phenomenon in which a weaker signal is masked by a stronger one within a critical band (Moore, 2003). Additionally, previous studies have shown that such masks can provide robust recognition results (Cooke et al., 2001; Roman et al., 2003; Srinivasan et al., 2006; Srinivasan and Wang, 2007). Hence, we propose a CASA system that estimates this mask to facilitate the recognition of target speech in the presence of interference. When multiple sources are of interest, the system can produce ideal binary masks for each source by treating one source as target and the rest as interference.

In this paper, we present a two-stage monaural CASA system that follows the ASA account of auditory organization as shown in Fig. 1. The input to the system is a mixture of target and interference. The input mixture is analyzed by an auditory filterbank in successive time frames. The system then generates segments based on periodicity and a multi-scale onset and offset analysis, producing voiced and unvoiced segments respectively (Hu and Wang, 2006). In the simultaneous grouping stage, the system estimates pitch tracks of individual sources in the mixture and employs periodicity similarity to group voiced segments into simultaneous streams. A simultaneous stream comprises multiple segments that overlap in time. Subsequently, a sequential grouping stage employs speaker characteristics to organize simultaneous streams and unvoiced segments across time into whole streams corresponding to individual speaker utterances. Within this stage, we first sequentially group simultaneous streams corresponding to voiced speech and then unvoiced segments. Finally, the CASA system outputs an estimate of the ideal binary mask corresponding to an underlying speaker in the input mixture. Such a mask is then used in an uncertainty decoding approach to robust speech recognition (Srinivasan and Wang, 2007). This approach reconstructs missing feature components as indicated by the binary masks, and also incorporates reconstruction errors as uncertainties in the speech recognizer. Finally, in the case of multiple speech sources, a target selection process is employed for identifying the target speech.

The rest of the paper is organized as follows. Sections 2–5 provide a detailed presentation of the various components of our proposed system. The system is systematically evaluated on a speech separation challenge